MEMORY IS A DERIVATION: THE DISTRIBUTED-EVIDENCE PARADOX IN LONG-TERM AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-running LLM agents compress past interactions into persistent memories that may be reused as premises for later tasks. This creates a distinct *derivation problem*: whether the memory actually follows from what the interaction history supports. Relevant evidence may be scattered across earlier interactions, while compression can introduce relations or event status that the history never established. A valid memory may therefore appear unsupported because its citations omit relevant evidence, while individually supported facts may be composed into a stronger statement the history never established. We characterize this problem through three coupled requirements: *(1) Evidence scope*; *(2) Compositional validity*; *(3) Admission reliability*. We therefore ask whether the interaction history available at write time supports what enters persistent memory. We introduce **DERIVAUDIT**, a framework for auditing whether a memory is actually supported by the history available when it was written. The audit separates three questions: whether supporting evidence lies beyond writer-provided citations, whether the composed memory introduces unsupported meaning, and how write-time admission decisions affect later memory use. Across two natural memory corpora, audits using broader pre-write history recover support for nearly 60% of memories that appear unsupported from citations alone, while 17–21% remain unsupported after expansion. Yet broader evidence does not by itself make admission reliable: unsupported memories are still frequently admitted (59–89%) across verification models, and evidence expansion alone worsens it on two backbones. Controlled experiments reveal a *distributed-evidence paradox*: valid memories often require combining evidence across interactions, yet several plausible facts can make unsupported compositions harder to detect. Moreover, requiring a verifier to check more independently true details can itself increase rejection of valid memories. On 391 labelled real memory writes, composition-aware verification with expanded history improves retention of valid memories rescued by broader-history evidence but does not consistently reduce unsupported admission across backbones. Downstream probes further show that both distorted and missing memories can harm later answers.

1 INTRODUCTION

Long-running LLM agents use persistent memory to carry information across turns, sessions, and tasks without retaining every past interaction in context. Prior systems store conversational experience, reflections, and reusable knowledge in persistent state, and increasingly organize or evolve that state over time (Park et al., 2023; Zhong et al., 2023; Shinn et al., 2023; Zhao et al., 2024; Xu et al., 2025). Recent surveys accordingly treat memory formation, evolution, and retrieval as core components of long-running agents (Zhang et al., 2024). This makes reliable memory writing important: once information is stored, later tasks may treat it as an established fact rather than return to the original interaction history, and may further combine stored memories to draw new conclusions. Yet a memory may not faithfully reflect what the history actually supports. A writer can combine observations from different turns, omit an important qualification, or introduce a relation that was never established. When such an error enters persistent memory, it can continue to shape later agent behavior even after the original interaction is no longer in context.

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.

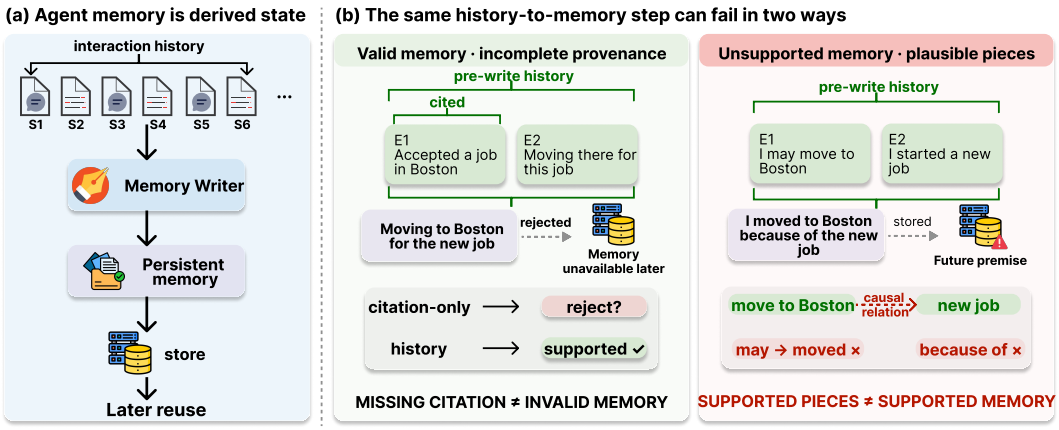

Figure 1: **Persistent memory as derived state. (a)** Interaction history is compressed into persistent memory that can later be reused as agent state. **(b)** This write step has two distinct failure modes: incomplete citations can make a valid memory appear unsupported, while supported historical pieces can be combined into a memory whose full meaning is not supported.

This raises question for reliable memory writing: how can we tell whether a new memory is actually supported by the interaction history? Figure 1 illustrates why this is not straightforward. In one case, a memory is supported by earlier interactions, but its attached citations capture only part of the relevant evidence, so a verifier that checks citations alone may reject a valid memory. In the other, individual facts are each supported somewhere in the history, but the memory combines them into a relation or changes event's status in a way the history never established. A reliable verifier must look beyond citation coverage and determine whether the memory as a whole is justified by the history (Rashkin et al., 2023; Gao et al., 2023b;a). We study this as a *derivation problem*: whether proposed memory can be derived from the interactions that came before it. This problem has three closely related aspects. **(1) Evidence scope:** the evidence needed to support a memory may extend beyond the sources explicitly attached to it (Joshi, 2026; Jin et al., 2026). **(2) Compositional validity:** even when individual facts are supported, their combination may introduce meaning that the history does not support. **(3) Admission reliability:** a write-time verifier should reject unsupported memories without discarding valid information that may be needed later (Cui et al., 2026; Zhang & Li, 2026).

We introduce **DERIVAUDIT**, a framework for auditing how interaction history is turned into persistent memory. We call the underlying correctness requirement *semantic derivation integrity*: what is written into persistent memory should be supported by the interaction history available at write time. To audit this requirement, DERIVAUDIT compares writer-provided citations with broader pre-write history, checks whether the full meaning of a memory is supported rather than only its individual facts, and evaluates which supported and unsupported memories survive write-time verification.

We evaluate DERIVAUDIT on 400 unedited memory writes from LoCoMo (Maharana et al., 2024) and HaluMem-Medium (Chen et al., 2026), drawn from two memory-writing pipelines and evaluated with three verification models. Across both corpora, broader pre-write history recovers support for about 60% of memories whose attached citations are insufficient, showing that citation-only verification can mistake incomplete provenance for unsupported memory; even after expansion, roughly one fifth of writes remain unsupported. Controlled experiments reveal a complementary difficulty: valid memories often require evidence distributed across multiple interactions, yet several individually plausible facts can make unsupported combinations harder to reject, a pattern we call the *distributed-evidence paradox*. No verification approach consistently resolves this tension across models, and requiring the verifier to check more independently supported details can itself increase rejection of valid memories. On 391 labelled natural memory writes, broader-history and composition-aware verification improve retention of valid memories rescued by additional historical evidence on some models, while rejection of unsupported memories remains inconsistent. Finally, downstream interventions show why these write-time errors matter: distorted memories can propagate incorrect information, while rejecting valid memories can remove information needed for later tasks.

Our contributions are threefold. **(1) A derivation view of memory writing.** We frame persistent memory as a derivation from interaction history and define *semantic derivation integrity*: what is

written should be supported by the history available at write time. **(2) An empirical framework for studying write-time memory reliability.** DERIVAUDIT provides a common audit setup for separating incomplete provenance from unsupported composition and examining admission decisions across verification settings. **(3) A systematic study of memory-writing failure modes.** Across natural and controlled settings, we characterize incomplete provenance, the *distributed-evidence paradox*, and verification-induced rejection of valid memories, and show that both distorted and missing memories can affect downstream behavior.

2 RELATED WORK

We situate our work along three lines. **(1) Long-term memory for LLM agents.** Persistent-memory systems enable agents to retain and reuse information across extended interactions through extraction, consolidation, updating, and retrieval (Packer et al., 2024; Chhikara et al., 2025; Latimer et al., 2026), while benchmarks evaluate recall, temporal reasoning, knowledge updating, and failures across memory operations (Latimer et al., 2026; Wu et al., 2025; Hu et al., 2026; Chen et al., 2026). Rather than asking primarily whether stored information can later be recalled, we study whether a memory was semantically justified by the interaction history when it entered persistent state. **(2) Reliable memory writing and admission.** Recent work increasingly treats memory state as a reliability boundary: Eywa links memories to provenance, MemIR distinguishes source evidence from truth-bearing memory content, and ConsistencyGate and MemTxn control whether candidate updates should be committed (Joshi, 2026; Jin et al., 2026; Zhang & Li, 2026; Cui et al., 2026). Our setting exposes two additional challenges: writer-supplied provenance may omit historical evidence that supports a valid memory, while individually supported facts may still fail to license the relations introduced when they are composed into a memory. DERIVAUDIT therefore evaluates natural memory writes under controlled evidence scopes and measures both unsupported memories retained and historically supported memories discarded. **(3) Fine-grained factuality and compositional verification.** Prior work uses dependency-level entailment, semantic graphs, predicate–argument structure, and turn-level verification to expose unsupported relations (Goyal & Durrett, 2020; Laban et al., 2022; Honovich et al., 2022; Ribeiro et al., 2022; Cattan et al., 2025; Lewis et al., 2026; Liu et al., 2025), while studies of abstractive generation document cases in which source-present content is recombined into unsupported statements (Maynez et al., 2020; Pagnoni et al., 2021). We extend these ideas to persistent agent memory, where a record is derived from longitudinal interaction history and may later be reused without its original sources; DERIVAUDIT jointly studies *where* its support resides, *what* meaning composition adds, and *which* memories survive write-time verification and affect later agent behavior. Table 1 summarizes these distinctions.

Table 1: Comparison with closely related work. DERIVAUDIT studies whether interaction history supports what is written into persistent memory, including evidence beyond attached provenance, compositional meaning, and the consequences of admission decisions.

Work	Agent Memory	Evidence Beyond Attached Provenance	Compositional Semantics	Write-time Admission	Downstream Memory Effects
Long-term memory and reliability					
LoCoMo / LongMemEval[a]	✓	–	–	–	–
HaluMem (Chen et al., 2026)	✓	–	–	–	✓
Eywa (Joshi, 2026)	✓	–	–	✓	–
ConsistencyGate (Zhang & Li, 2026)	✓	–	–	✓	–
MemTxn (Cui et al., 2026)	✓	–	–	✓	–
MemIR (Jin et al., 2026)	✓	–	–	–	–
Fine-grained factuality					
DAE / FactGraph / QASemConsistency[b]	–	–	✓	–	–
VISTA (Lewis et al., 2026)	–	–	–	–	–
DERIVAUDIT (Ours)	✓	✓	✓	✓	✓

[a]Maharana et al. (2024); Wu et al. (2025); [b]Goyal & Durrett (2020); Ribeiro et al. (2022); Cattan et al. (2025). **Evidence Beyond Attached Provenance** denotes explicit comparison or recovery of support outside the writer-provided evidence for the same memory write. **Compositional Semantics** denotes explicit verification of relations or qualifications beyond isolated factual ingredients. **Downstream Memory Effects** denotes analysis of how stored or missing memory affects later agent behavior.

3 DERIVAUDIT: AUDITING MEMORY AS DERIVATION

We use *agent memory* to mean a compact record distilled from earlier interactions and stored for use in later turns or tasks. A *candidate memory write* is a compact record proposed for persistent storage; admission determines whether that record is committed to memory. DERIVAUDIT audits this process by asking three questions: where the memory's support lies in the earlier interaction history, whether that history supports the full meaning of the memory, and whether write-time verification retains the right information for later use.

3.1 EVIDENCE SCOPE: RECOVERING THE HISTORY BEHIND A MEMORY

Writer-provided citations may capture only part of the history that supports a memory. Our goal is therefore to distinguish a genuinely unsupported memory from a valid memory whose provenance is incomplete. For each candidate memory, we compare two views of the same pre-write history. The *citation-only* view contains only the passages attached by the writer. The *expanded-history* view supplements these citations with up to 12 BM25-retrieved passages from interactions that occurred before the memory was written, subject to a total budget of 16 passages. Later interactions are excluded, and previously written memories are not treated as independent evidence for claims derived from the same history.

We evaluate the unchanged memory independently under both evidence views using the same reference-labeling procedure described in Section 3.2. Passages are presented chronologically and without indicating whether they were cited or retrieved, so the paired judgments isolate the effect of seeing more of the pre-write history. We call a memory *provenance-repaired* when it is insufficiently supported by its citations but supported after history expansion, and report the fraction of citation-insufficient memories repaired in this way. Because history expansion is retrieval-bounded, an insufficient judgment does not always establish that a memory was unsupported by the full history. We therefore treat search-limited cases as unresolved rather than as historically unsupported.

3.2 COMPOSITIONAL VALIDITY: CHECKING WHAT THE AGENT WILL REMEMBER

Recovering the relevant history does not guarantee that a memory is supported. A writer may combine individually supported facts into a stronger statement, for example by adding a causal relation, changing when an event occurred, or turning a plan into an accomplished fact. We therefore ask whether the evidence supports the *full meaning* that will be stored in memory.

(1) Semantic obligations. We decompose each memory into the factual claims, relations, and qualifications that must all be supported for the memory as a whole to be valid. We refer to these requirements as *semantic obligations*. For memory m_i, let

$$\mathcal{O}(m_i) = \{o_{i1}, \ldots, o_{iJ_i}\}.$$

For example, "The user moved to Boston because of the new job" requires evidence not only for the move and the job, but also that both occurred and that the job caused the move. We consider a memory supported only when all of its obligations are supported by the available evidence:

$$Y(m_i, E_i) = \mathbf{1}[\forall o \in \mathcal{O}(m_i),\ \ell(o, E_i) = \textsc{Supported}]\,.$$

(2) Reference labels. Each obligation receives a four-way reference label from two independent model assessments, with disagreements resolved by blinded adjudication. Support and contradiction are assessed separately so that missing evidence is not treated as evidence against a claim.

(3) Verification views. We next study how different representations of the same memory affect write-time verification. Every view receives the same candidate memory, evidence, and verification model; only the representation used for checking changes:

$$s_i^{(v)} = V_\theta(\phi_v(m_i), E_i)\,,$$

where ϕ_v denotes verification view v. We compare holistic factuality, atomic-claim verification, and predicate–argument QA (Luo et al., 2023; Min et al., 2023; Cattan et al., 2025) with views that explicitly represent relations and qualifications beyond isolated claims. This comparison tests

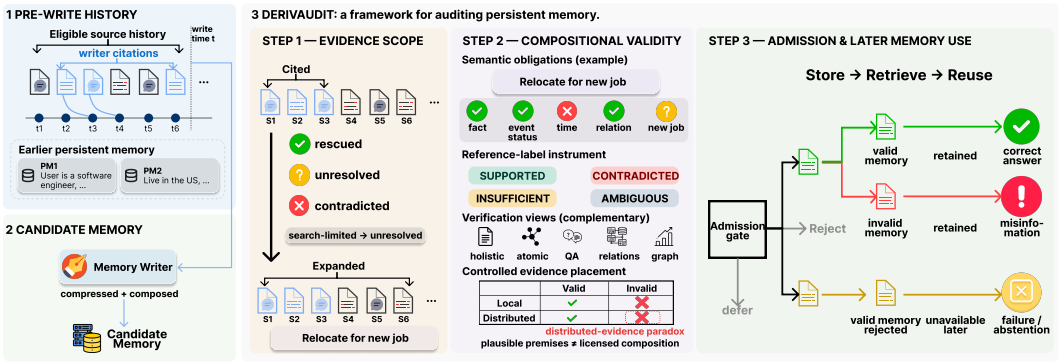

Figure 2: **DERIVAUDIT audits memory derivation from pre-write history to later use.** The framework organizes the analysis around three questions: where a memory's historical support resides, whether that support licenses the full meaning introduced during composition, and how alternative admission decisions affect the memory available to later tasks.

whether different verification representations preserve the compositional meaning needed to determine whether the memory is supported by its history.

(4) Controlled evidence distribution. Finally, we isolate the effect of where supporting information appears in the history. Across 70 matched families of examples, we construct valid and invalid memories under *local* and *distributed* evidence conditions while keeping the candidate memory and its validity fixed. Valid distributed cases require combining evidence across passages; invalid cases contain individually plausible premises but no support for the relation that combines them. Additional controls remove complete sets of evidence needed to support the target relation or replace plausible premises with matched irrelevant passages. These interventions distinguish difficulty integrating distributed evidence from difficulty rejecting unsupported compositions.

3.3 ADMISSION RELIABILITY: RETAINING KNOWLEDGE FOR LATER TASKS

A write-time verifier does more than classify a memory: its decision determines what information remains available to the agent later. Admitting an unsupported memory can introduce an incorrect premise, while rejecting a supported one can remove useful information. We therefore study whether verification makes reliable *admission decisions* and what happens when those decisions are wrong.

(1) Write-time admission. We replay the same natural candidate writes under three settings: citation-only holistic verification, expanded-history holistic verification, and expanded-history obligation-aware verification. For verification setting v, a memory is admitted when its score exceeds the corresponding threshold:

$$a_i^{(v)} = \mathbf{1}\left[s_i^{(v)} \geq \tau_v\right].$$

The two expanded-history settings receive the same evidence, allowing us to separate the effect of seeing more history from the effect of checking the memory differently. We measure both sides of the decision: retention of supported memories and admission of unsupported ones, with separate analysis of provenance-repaired memories.

(2) Verification granularity. We next ask whether checking a memory in greater detail can itself make valid memories harder to retain. For supported memories, we ask the verifier to check two or four additional obligations that are independently known to be true, while keeping the candidate memory, evidence, and reference validity unchanged. The additional obligations change what the verifier must check, not the content of the memory. Any systematic decrease in verification score or admission therefore arises from requiring more correct checks rather than from introducing an actual defect. We call this effect *accumulated verification noise*.

(3) Later memory use. We examine the consequences after a write-time decision is made. We track whether later rewriting preserves qualifications such as timing and uncertainty, and whether stored memories are retrieved into later contexts. In controlled downstream probes, we fix the task context and provide either the correct memory, a distorted version, or no memory. This isolates the cost of carrying an incorrect premise forward from the cost of removing information that a later task needs.

4 EXPERIMENTS

We use DERIVAUDIT to study three aspects of memory writing: evidence scope, compositional validity, and admission reliability. We combine natural memory writes with controlled interventions and then examine how write-time errors affect later memory use.

Evaluation Data. Our natural-data audit contains 400 unedited memory writes, split evenly between LoCoMo (Maharana et al., 2024) and HaluMem-Medium (Chen et al., 2026). Each write is evaluated under citation-only and expanded pre-write evidence; 391 receive resolved expanded-history reference labels and enter the admission evaluation. Our controlled evaluations include 70 matched families (280 examples) that vary whether evidence is local or distributed while holding the candidate memory and its validity fixed, and a 512-record suite for comparing verification approaches on single- and multi-passage support and unsupported relations. Additional experiments examine repeated memory rewriting, retrieval, and downstream use.

Models and Verification Settings. We evaluate Qwen3-30B-A3B, Gemma-4-31B, and Llama-3.3-70B-Instruct. Across experiments, these models serve as verification models and, where applicable, as memory writers or downstream readers. For write-time verification, we compare established approaches, holistic factuality judgment (Zheng et al., 2023), atomic-claim verification (Min et al., 2023), and predicate–argument QA (Cattan et al., 2025), with representations designed to make compositional relations explicit, including compact-relation, composition-graph, and obligation-level views. Within each comparison, all views receive the same candidate memories and evidence.

Metrics and Labels. Following Sections 3.1–3.3, we measure provenance repair for evidence scope; detection of unsupported relations and retention of valid multi-passage memories for compositional validity; and supported-memory retention and unsupported-memory admission for admission reliability. Natural reference labels are derived from two independent model assessments, with disagreements resolved by blinded adjudication. Threshold selection and model-specific decision rules appear in Appendix F.

4.1 WHAT DOES WRITE-TIME VERIFICATION RETAIN AND REJECT?

We first examine write-time admission on the 391 natural memories with resolved expanded-history labels: 313 are supported and 78 are unsupported or contradicted. Table 2 compares three settings: citation-only holistic verification, expanded-history holistic verification, and expanded-history obligation-aware verification.

Broader History Recovers Valid Memories with Incomplete Citations. On Qwen, retention of provenance-repaired memories rises from 82.1% with citation-only verification to 93.8% with expanded history, an 11.6-point paired improvement ($p = 0.007$). Adding obligation-aware checking yields similar retention at 92.9% ($p = 0.023$ versus citation-only). Overall valid-memory retention remains nearly unchanged across the three settings at 90.4%, 90.1%, and 90.7%. The same pattern appears on the other verification models: expanded history raises provenance-repaired retention from 58.0% to 95.5% on Gemma and from 85.7% to 98.2% on Llama.

Unsupported Writes Remain Frequently Admitted. On Qwen, unsupported-memory admission decreases from 84.6% with citation-only verification to 80.8% with expanded history and 76.9% with obligation-aware checking, but the latter difference from citation-only verification is not statistically resolved on the 78 negative records ($p = 0.24$). Across models, obligation-aware checking is not consistently associated with lower unsupported admission: the rate falls from 66.7% to 59.0% on Gemma but rises from 83.3% to 88.5% on Llama. Thus, broader history consistently improves retention of valid memories with incomplete provenance across the three models, whereas rejection of unsupported writes remains much less stable.

4.2 WHY IS MEMORY ADMISSION DIFFICULT?

Memory admission determines whether a candidate write is stored for later use. The preceding results reveal an asymmetry: broader history often helps retain supported memories with incomplete citations, while unsupported writes remain frequently admitted. We examine this gap through evidence scope, compositional validity, and admission reliability.

Table 2: Memory admission under three write-time verification settings on 391 decidable natural writes. Expanded history adds evidence from pre-write interactions beyond writer-provided citations; composition-aware checking additionally verifies meaning introduced when information is composed into memory. Lower unsupported admission and higher retention are better. A no-gate policy accepts every write and is omitted.

Model	Verification Setting	History Expansion	Composition Check	Unsupported Admission↓	Valid Retention↑	Repaired Retention↑
Qwen	Citation-only	–	–	0.846	0.904	0.821
	+ Expanded history	✓	–	0.808	0.901	0.938
	+ Obligation-aware check	✓	✓	0.769	0.907	0.929
Gemma†	Citation-only	–	–	0.667	0.837	0.580
	+ Expanded history	✓	–	0.769	0.971	0.955
	+ Obligation-aware check	✓	✓	0.590	0.939	0.920
Llama†	Citation-only	–	–	0.833	0.946	0.857
	+ Expanded history	✓	–	0.897	0.974	0.982
	+ Obligation-aware check	✓	✓	0.885	0.974	0.973

Notes. Repaired memories are valid writes supported only after expanding beyond writer-provided citations. Qwen uses cross-fitted thresholds; †Gemma and Llama use argmax operating points. On Qwen, repaired retention improves with expanded history ($+0.116$, $p = 0.007$) and composition-aware checking ($+0.107$, $p = 0.023$), while the reduction in unsupported admission is not statistically resolved ($p = 0.24$).

Table 3: Compositional verification on the 512-record controlled suite. Established and composition-aware verification views are evaluated on the same candidate memories and evidence in a shared scoring harness. U denotes unsupported-relation detection, and Multi denotes retention of valid memories requiring joint support from multiple passages.

Model	View Type	Verification View	Unsupported Detection U ↑	Multi-span Retention↑	Valid Rejection↓
Qwen	Established	Holistic judge (Luo et al., 2023)	**0.608**	0.846	0.088
		Atomic claims (Min et al., 2023)	0.165	0.744	0.121
		Predicate–argument QA (Cattan et al., 2025)	0.392	0.564	0.115
	Composition-aware	**Compact relations**	0.595	**0.949**	**0.082**
		Composition graph	0.582	0.872	0.093
		Per-obligation	0.291	0.897	0.093
Gemma†	Established	Holistic judge (Luo et al., 2023)	0.937	0.923	0.022
		Atomic claims (Min et al., 2023)	0.000	**1.000**	**0.005**
		Predicate–argument QA (Cattan et al., 2025)	0.975	0.205	0.176
	Composition-aware	**Compact relations**	*degenerate: rejects all 39 multi-span memories*		
		Composition graph	0.823	0.974	0.011
		Per-obligation	**0.987**	0.769	0.055

Notes. Candidate memory and evidence are fixed within each comparison. Established rows instantiate prior verification paradigms in the same shared harness. Single-span valid retention is at least 0.91 in every reported cell and is omitted. Qwen uses out-of-fold tuned thresholds; †Gemma uses its argmax operating point. Highlighted cells mark within-model extrema.

Evidence Scope: Valid Memories Often Need Evidence Beyond Their Citations. Figure 3(a) compares citation-only and expanded-history reference judgments on the same natural writes. Citations alone are insufficient for 42.5% of LoCoMo and 51.0% of HaluMem-Medium memories, but broader history recovers support for 60.0% and 59.8% of these cases, raising the overall supported fraction from 53.5% to 79.5% and from 47.0% to 77.0%, respectively. Moreover, 25 of 107 citation-supported LoCoMo memories and 14 of 94 HaluMem-Medium memories require multiple passages jointly for support. Thus, expanding the evidence recovers incomplete provenance, but also raises a harder question: whether evidence distributed across interactions supports the memory as a whole.

Compositional Validity: Supported Pieces Do Not Guarantee a Supported Memory. Table 3 shows a trade-off between detecting unsupported relations and retaining valid multi-passage memories. On Qwen, atomic verification detects only 16.5% of unsupported relations, while predicate–argument QA reaches 39.2% but retains only 56.4% of valid multi-passage memories. Our compact-relation view retains 94.9% while detecting 59.5%, close to the 60.8% detection rate of holistic verification. No single verification view consistently performs well on both criteria across models. Figure 3(b–c) further isolates the role of distributed evidence while holding the candidate memory and its validity fixed. Under distributed evidence, holistic and obligation-aware verification detect only 8.6% and 11.4% of invalid Qwen cases, respectively. We call this the *distributed-evidence paradox*: valid memories often require combining evidence across interactions, yet the same setting can make unsupported combinations harder to reject. Matched controls suggest that this pattern is not explained solely by longer context or greater evidence distance. Removing all sufficient evidence for the target relation flips 88–100% of affected decisions to rejection, while replacing plausible premises with matched irrelevant history reduces invalid-memory acceptance from 65–88% to 0–10%.

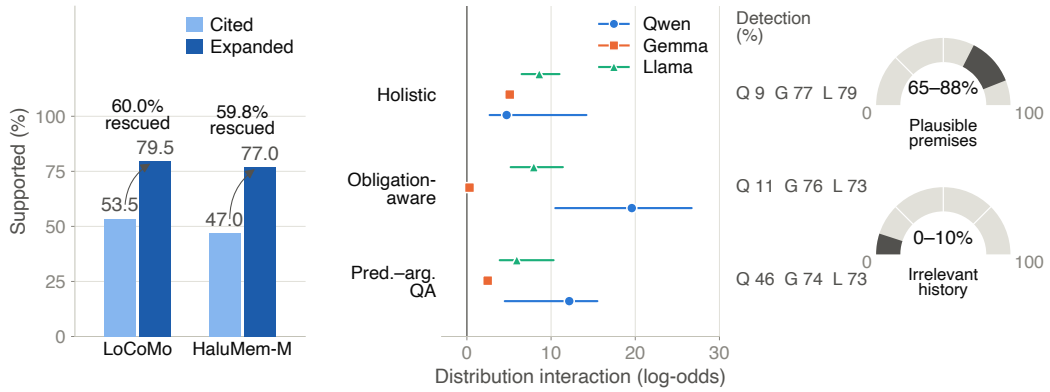

Figure 3: Evidence scope and compositional verification. **(a)** Historical evidence recovers support beyond writer-supplied citations. **(b–c)** Distributed plausible premises reduce valid–invalid separation and increase acceptance of unsupported memory compositions relative to matched irrelevant history.

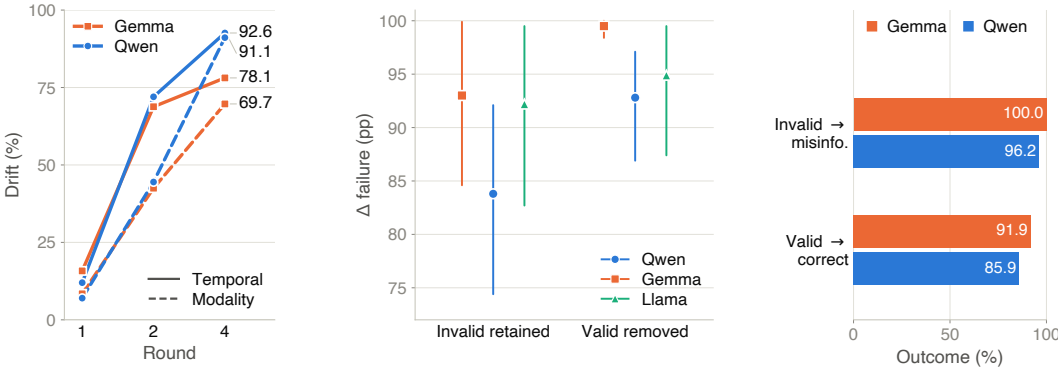

Figure 4: From persistent memory to downstream behavior. **(a)** Semantic qualifications drift during repeated consolidation. **(b)** Distorted and missing memories both increase downstream failure. **(c)** Natural retrieval links invalid memory to misinformation and valid memory to correct answers.

Admission Reliability: More Detailed Checking Can Reject Valid Memories. More detailed verification introduces a complementary failure mode. When the verifier is asked to check two or four additional obligations that are independently known to be true, while the memory itself remains unchanged, its confidence often declines. With four additional checks, Qwen's median support score drops by 7.7 log-odds units, with 98.7% of records decreasing. The resulting admission rate changes little on Qwen, from 99.4% to 98.1%, but falls from 93.3% to 2.6% on Gemma and from 97.4% to 35.5% on Llama. We refer to this pattern as *accumulated verification noise*. Together, these results expose a two-sided tension: insufficient checking can miss unsupported compositions, while increasingly detailed checking can make valid memories harder to retain. Alternative aggregation rules shift this trade-off but do not eliminate it (Appendix G).

4.3 What Happens After the Write-Time Decision?

We finally examine what happens after information enters persistent memory: whether timing, uncertainty, and other qualifications remain stable during later rewriting, and how distorted or missing memories affect downstream tasks.

Stored Memories Can Drift during Later Rewriting. Figure 4(a) follows naturally generated memories through repeated extraction and consolidation. Both temporal drift, which changes when an event is said to occur, and modality drift, which changes whether an event is planned, uncertain, or realized, increase substantially across rounds. For Gemma, the two rates rise from 8.4% and 15.8% at extraction to 69.7% and 78.1% by round 4; for Qwen, from 7.0% and 12.0% to 91.1% and 92.6%. These trends are descriptive rather than causal, since citation breadth, writer settings, and record populations also vary across rounds.

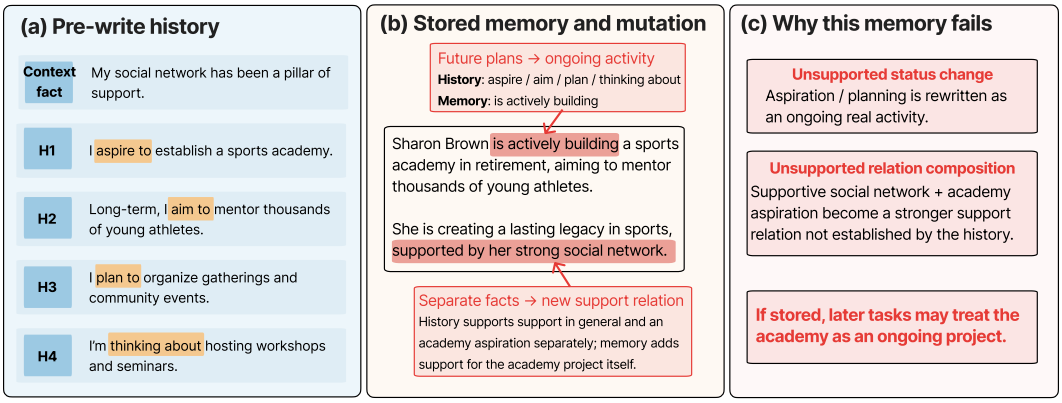

Figure 5: **A natural failure of memory derivation.** The writer compresses future-oriented plans into an ongoing activity and combines separately supported facts into a relation that the pre-write history does not establish.

Admission Errors Affect Downstream Answers. In Figure 4(b), we vary only the memory supplied to a downstream reader while holding its retrieval position and surrounding context fixed. Relative to a correct memory, supplying a distorted memory increases downstream failure by 93.0, 83.8, and 92.2 percentage points for Gemma, Qwen, and Llama, while removing a valid memory increases failure by 99.5, 92.8, and 94.9 points on queries that require that information. Natural retrieval is consistent with the same pattern (Figure 4(c)): invalid retrieved memories produce misinformation in 100% and 96.2% of Gemma and Qwen cases, while valid retrieved memories support correct answers in 91.9% and 85.9%.

A Natural Write Illustrates a Derivation Failure. Figure 5 shows an unedited HaluMem example. The history states that Sharon Brown *aspires* to establish a sports academy and separately describes her social network as supportive. The written memory introduces two unsupported changes: a future plan becomes the present-tense assertion that she *is actively building* the academy, while a generally supportive social network is recast as supporting that specific effort. Neither stronger statement is established by the pre-write history, and the memory remains unsupported after evidence expansion.

5 Conclusion and Limitations

We study persistent agent memory as a derivation from interaction history into persistent records that later tasks may reuse. Across natural memory writes and controlled interventions, we identify three challenges to reliable memory writing. First, *evidence scope*: broader pre-write history recovers support for nearly 60% of memories whose citations alone are insufficient. Second, *compositional validity*: individually plausible facts can still be combined into unsupported meaning, and distributed evidence can make such compositions harder to reject, a pattern we call the *distributed-evidence paradox*. Third, *admission reliability*: recovering missing provenance substantially improves retention of provenance-repaired memories, but rejection of unsupported writes remains inconsistent across verification models, while more detailed checking can itself reject valid memories. Downstream interventions further show that both distorted and missing memories can substantially affect later answers. Together, these findings show that reliable memory writing requires more than finding relevant evidence: verification must also determine whether that evidence supports the full meaning that enters persistent memory.

Limitations. Reference labels rely primarily on two independent model assessments, with disagreements resolved by blinded adjudication, although stratified human review shows 96.0% agreement. Evidence expansion is retrieval-bounded, so failure to recover support does not establish historical invalidity. Natural data contain relatively few contradicted cases and cases involving explicit relations, while our controlled suites cover only a subset of the ways meaning can be altered during memory writing. Results also vary across verification models, verification views, prompting, and aggregation rules. Finally, consolidation results are observational, and downstream interventions do not capture a fully closed-loop memory system. We therefore treat these findings as evidence about the evaluated settings rather than universal claims about memory verification.

Reproducibility Statement

Sections 3 and 4 describe the memory-writing setting, evidence scopes, reference-label procedure, verification views, admission decision rules, evaluation metrics, and controlled interventions used in our analyses. The appendix provides annotation attrition and label denominators, robustness and protocol controls, threshold-selection details, and additional implementation information. The accompanying artifact records model configurations, prompts and chat templates, retrieval indexes and evidence budgets, random seeds and sampled record IDs, annotation and adjudication outputs, decision thresholds, experiment manifests, and analysis code. Results reused across analyses are linked to their original recorded runs rather than reconstructed from prose.

AI Use Statement

Generative AI tools were used to assist with code and LaTeX polishing. Separate model instances also served as annotators and blinded adjudicators for the natural memory audit; agreement from this procedure is reported as model-instance agreement rather than human validation. Human review was conducted separately on a stratified sample, as described in the appendix. The authors remain responsible for citations, mathematical claims, data processing, and reported results.

Ethics Statement

Our experiments use public long-term conversation benchmarks and do not collect new private user conversations. Human annotators were used only to review sampled benchmark records for validation of the model-produced reference labels and were compensated hourly at standard rates. No new personal profiles or inferred real-world identities were created as part of the study. Released artifacts should preserve the licenses and redistribution requirements of the underlying datasets, minimize unnecessary reproduction of conversation text, and avoid adding identities or personal attributes not present in the source data.

References

Arie Cattan, Paul Roit, Shiyue Zhang, David Wan, Roee Aharoni, Idan Szpektor, Mohit Bansal, and Ido Dagan. Localizing factual inconsistencies in attributable text generation, 2025. URL https://arxiv.org/abs/2410.07473.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. Halumem: Evaluating hallucinations in memory systems of agents, 2026. URL https://arxiv.org/abs/2511.03506.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Hanshuai Cui, Zhiqing Tang, Zhi Yao, Fanshuai Meng, Qianli Ma, and Weijia Jia. Memtxn: A transaction boundary for source-supported updates and complete-state recovery in agent memory, 2026. URL https://arxiv.org/abs/2607.27834.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Tanya Goyal and Greg Durrett. Evaluating factuality in generation with dependency-level entailment, 2020. URL https://arxiv.org/abs/2010.05478.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. TRUE: Re-evaluating factual consistency evaluation. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3905–3920, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.287. URL https://aclanthology.org/2022.naacl-main.287/.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions, 2026. URL https://arxiv.org/abs/2507.05257.

Zhengda Jin, Bingbing Wang, Jing Li, Ruifeng Xu, and Min Zhang. Mitigating provenance-role collapse in long-term agents via typed memory representation, 2026. URL https://arxiv.org/abs/2605.25869.

Resham Joshi. Eywa: Provenance-grounded long-term memory for ai agents, 2026. URL https://arxiv.org/abs/2605.30771.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. SummaC: Re-visiting NLI-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177, 2022. doi: 10.1162/tacl_a_00453. URL https://aclanthology.org/2022.tacl-1.10/.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In Greg Durrett and Ping Jian (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 275–285, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-392-0. doi: 10.18653/v1/2026.acl-demo.27. URL https://aclanthology.org/2026.acl-demo.27/.

Ashley Lewis, Andrew Perrault, Eric Fosler-Lussier, and Michael White. Vista: Verification in sequential turn-based assessment, 2026. URL https://arxiv.org/abs/2510.27052.

Hongjun Liu, Yilun Zhao, Arman Cohan, and Chen Zhao. Sucea: Reasoning-intensive retrieval for adversarial fact-checking through claim decomposition and editing, 2025. URL https://arxiv.org/abs/2506.04583.

Zheheng Luo, Qianqian Xie, and Sophia Ananiadou. Chatgpt as a factual inconsistency evaluator for text summarization, 2023. URL https://arxiv.org/abs/2303.15621.

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

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization, 2020. URL https://arxiv.org/abs/2005.00661.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation, 2023. URL https://arxiv.org/abs/2305.14251.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics, 2021. URL https://arxiv.org/abs/2104.13346.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior, 2023. URL https://arxiv.org/abs/2304.03442.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, December 2023. doi: 10.1162/coli_a_00486. URL https://aclanthology.org/2023.cl-4.2/.

Leonardo F. R. Ribeiro, Mengwen Liu, Iryna Gurevych, Markus Dreyer, and Mohit Bansal. Factgraph: Evaluating factuality in summarization with semantic graph representations, 2022. URL https://arxiv.org/abs/2204.06508.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning, 2023. URL https://arxiv.org/abs/2303.11366.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory, 2025. URL https://arxiv.org/abs/2410.10813.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents, 2025. URL https://arxiv.org/abs/2502.12110.

Yan Zhang and Shibo Li. Consistencygate: Preventing memory contamination in llm agents via self-consistency admission control, 2026. URL https://arxiv.org/abs/2607.22962.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents, 2024. URL https://arxiv.org/abs/2404.13501.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners, 2024. URL https://arxiv.org/abs/2308.10144.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. URL https://arxiv.org/abs/2306.05685.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. Memorybank: Enhancing large language models with long-term memory, 2023. URL https://arxiv.org/abs/2305.10250.

A Formal Boundaries and Proofs

A.1 Evidence and information boundaries

The first boundary concerns evidence scope. A verifier that checks each evidence unit independently and aggregates by a maximum cannot accept every valid record whose support is genuinely joint.

Proposition 1 (Single-span evidence boundary)**.** *Suppose m is licensed by $S = \{e_i, e_j\}$ under Γ, while neither e_i nor e_j licenses m alone. Any verifier whose accept decision requires $\max_k V(m, e_k)$ to exceed a threshold rejects m whenever every individual score remains below that threshold.*

The second boundary concerns information, not architecture. If an interface maps two candidates to identical inputs after deleting their only differing truth-conditional relation, no downstream scorer can recover that relation.

Lemma 1 (Representation-deletion boundary)**.** *Let m^+ and m^- share the same mention-only representation and evidence but differ in a relation that is licensed in m^+ and unlicensed in m^-. Every verifier that factors exclusively through that shared representation assigns the pair the same score.*

These statements do not rule out scalar or holistic verification. For example,

$$V^*(m, H_t) = I_\Gamma(m; \mathcal{E}(H_t)) \tag{1}$$

is scalar-valued yet semantically complete by construction. Verification can fail because its evidence scope is insufficient or because its checks omit truth-conditional components; a single output score is not itself the cause.

Repair requires a separate semantics. Logical weakening is allowed only when the original assertion entails the replacement under the declared logic. Evidence-grounded correction is a different operation: it replaces an unsupported obligation with one separately licensed by the history. For example, $\textsc{completed}(p)$ does not generally entail $\textsc{planned}(p)$, so that edit cannot be presented as logical weakening without an additional premise.

Proof of Proposition 1. By assumption, every individual evidence unit yields a score below the acceptance threshold, although the joint set licenses the record. The maximum of those individual scores remains below the threshold, so the verifier rejects the record. The proposition makes no claim about a verifier that scores evidence subsets jointly. □

Proof of Lemma 1. The verifier receives identical representation and evidence inputs for m^+ and m^-. Substituting identical inputs into its factorization yields identical scores. The result does not apply when the interface preserves the differing relation or exposes the original record. □

B Annotation Instrument and Adjudication

Primitives, never labels. Reference labels are produced by routing, not by typing: annotators never write a support-status label. For each obligation, an annotator answers three primitive questions in a fixed order. **D** (*determinate*): does the obligation have determinate truth conditions under Γ, answered from the record alone before the evidence spans are consulted, with a reason code required for a negative answer; every reason code names a property of the record, never of the evidence, which is what keeps D answerable before the spans are read. **E** (*entailed*): is there a subset of the spans that entails the obligation under Γ; if yes, the minimal such subset is listed. **R** (*refuted*): is there a subset that entails the negation of the obligation under Γ; if yes, that subset is listed. E and R have identical scope—one obligation, all spans—and differ only in direction, so neither question can collapse into the negation of the other.

Routing. The four-way label is a pure function of the primitives:

D	E	R	label
no	–	–	ambiguous under Γ (D's reason code)
yes	yes	no	supported
yes	no	yes	contradicted
yes	no	no	insufficient evidence
yes	yes	yes	ambiguous under Γ (Γ-inconsistent)

Routing exists for stability: directly typed multi-way labels measured 0.69–0.78 agreement stability under rubric-wording revision alone, against 0.81–0.91 for the primitives. Under routing, rubric wording moves the primitives, and the label moves only as far as the primitives do.

Refutation is time-indexed. A span refutes an obligation only when both concern the same subject, event, and time frame under Γ; a later state change is an update, not a contradiction. This rule is part of the frozen instrument text and is applied identically in both evidence scopes.

Adjudication. Every record is annotated by two independent model instances. Disagreements—and only disagreements—go to a blinded adjudicator that sees both primitive sets without identities or provenance and itself writes primitives, never labels. A record-level verdict is then routed from its obligation labels by the pre-registered precedence contradicted $>$ unsupported $>$ ambiguous $>$ supported. Two clerical defects found during review (one omitted obligation, one transposed obligation id) were repaired as transcription by full re-annotation from a fresh instance, never by editing labels.

C Natural-Audit Accounting and Pilot Gate

Attrition. Every unit between the issued scaffold and the agreement coefficients reported in Section 4 is accounted for. An obligation marked spurious by either annotator leaves the support-status family, since there is no pair left to compare.

records annotated by both	160
scaffold obligations issued	643
judgements if none were spurious ($\times 2$)	1286
spurious marks	-20
obligation judgements	1266
obligations dropped, both annotators marked spurious	8
obligations dropped, one annotator marked spurious	4
support-status family	631

The two drops close exactly: $8 \times 2 + 4 = 20$ spurious marks, and $643 - 8 - 4 = 631$ units. Annotator additions do not enter this family at all; they appear only in the inventory row, which is why that row's n exceeds 643.

Four denominators for *contradicted*. The label is rare enough that the choice of denominator changes the number reported, so all four are given. Pre-adjudication: 9 of 1266 judgements (0.71%); 6 of 631 units where either annotator used it (0.95%); 3 of 631 where both did (0.48%). Post-adjudication: 4 of 631 final labels (0.63%). The reliability coefficients reported in Section 4 are pre-adjudication by construction; any prevalence statement uses the post-adjudication count.

The pilot gate. The pre-registered gate required all four labels to occur in a 12-record pilot carrying 102 obligation judgements. At the base rate the main round subsequently measured ($p = 0.0071$), the expected count of *contradicted* in the pilot was $102p = 0.72$ and $P(N = 0) \approx e^{-0.72} = 0.487$. A gate demanding the occurrence of a label below 1% in a sample that size fails roughly half the time on a correctly functioning instrument. Nothing about the instrument was altered in response: prompts, routing thresholds and the instrument hash are identical before and after. The evidence that refutation is reachable comes from an independent 12-item probe with stipulated gold, whose three refutation items all routed to *contradicted*; that probe is logically prior to the main round and does not use it. We record the gate as a protocol deviation, keep every rare-label result descriptive, and

note the lesson that generalises: a pre-registered occupancy check must be powered for the rate of the rarest label it demands, or it manufactures failures that then need explaining away.

D Evidence Scopes and Retrieval Configuration

Two audits of the same records. Every natural record is audited twice under the same instrument. The citation-only audit presents the record with its writer-supplied cited spans. The expanded-history audit presents the same record under a new, unlinkable item id with evidence equal to the cited spans united with the BM25 top-12 passages retrieved from the *pre-write* history—everything up to the last span position the writer could see—capped at 16 spans total. Spans are shown in chronological order with no marking of which were cited and which retrieved; provenance and all construction flags live only in a key file the annotator never sees.

Unresolved is its own outcome. Retrieval is an operational approximation to the full pre-write evidence, so its failures must not be converted into semantic verdicts. Records whose retrieval hit the search cap are flagged `search_limited` in the key, and an expanded-scope *insufficient* verdict on such a record is reported as unresolved—reference-limited, not unsupported-by-history. This rule is applied mechanically from the flag, never by judgement.

Prevalence estimands. The headline citation-insufficiency rates are *conversation-balanced*: each conversation contributes equally, so a single prolific conversation cannot dominate the estimate. The *write-weighted* alternative, a Horvitz–Thompson estimate weighting conversations by their number of writes, moves every reported rate by at most 0.017 and is reported alongside. Proportions carry Wilson intervals; contrasts across scopes are record-paired, with conversation-resampled intervals where the cluster structure matters.

E Controlled Suite Construction

Matched families. Each of the 70 validated seed families instantiates the four cells $\{\text{local}, \text{distributed}\} \times \{\text{valid}, \text{invalid}\}$ over one candidate memory pair. A cell's evidence block contains its target span or spans, the family's two fixed distractors, and deterministic neutral padding drawn from the same conversation's real history up to a common evidence budget, so the only cross-cell differences are where the licensing premises sit and whether they license. Families are matched on semantic type, obligation count, candidate length, and distractors. Truth is by construction; nothing visible in the prompt names a cell or a label, construction labels live only in the key file, and item files are leak-checked mechanically with word-boundary matching.

Licensing-path interventions. The remote-evidence conditions operate on support sets rather than individual spans: the label-flipping intervention removes *every* sufficient support set for the target obligation, the label-preserving intervention removes only redundant evidence and retains at least one sufficient set, and the correction condition adds a temporally applicable explicit refutation. The mechanism-control variant presents the same invalid record with its plausible distributed premises replaced by irrelevant conversation content matched in token length and position (one family was dropped for residual content overlap, leaving 69).

Equal-budget prompt search. Each compared interface receives the identical budget: three fixed paraphrase templates crossed with $\{0, 2, 4\}$ few-shot examples, nine configurations per interface. Shot examples come only from four families excluded from scoring. The winning configuration per interface is selected on development conversations by the declared metric and frozen, then scored once on untouched test conversations. Within-interface development ranges span 2–$3.5\times$ (holistic $+21$ to $+53$; QA $+15$ to $+54$), larger than most between-interface gaps under default prompts; the faithful-atomic interface's best of nine configurations remains below the other interfaces' two-shot means, so the decomposition deficit survives its entire budget.

F Threshold Selection and Decision Rules

Verdict extraction. Every verifier decision is read from the true token log-probabilities of the verdict words rather than from sampled text. The serving window is widened to the top 100 log-probabilities, and the request is rejected rather than silently truncated if the server cannot honour it. Two quantities are recorded for each judgement: the normalised probability $p(\textit{yes})/(p(\textit{yes}) + p(\textit{no}))$, which saturates against 1.0, and the log-odds $\log p(\textit{yes}) - \log p(\textit{no})$, which ranks identically but keeps near-ties apart. All thresholding and paired analyses operate on log-odds. A record for which neither verdict token appears in the window is excluded and counted as a scoring error, never defaulted to mid-scale; a record for which only one token appears yields a real score with an unmeasurable margin and is counted separately as a rail hit.

The rail audit. On Gemma, 84–99% of scores sit on the window floor. A full-vocabulary echo audit on 120 Gemma rows (stratified over rail direction) and 33 Qwen rows recovered the losing verdict's exact probability: median 10^{-10}, maximum 4.2×10^{-9}. Replacing the floored values with exact probabilities flips 0 of 120 Gemma decisions and 1 of 33 Qwen decisions (a $+0.031$-margin record inside the known rerun variation). Near-binary verdicts are therefore a property of the model, not of the scoring pipeline.

Backbone-specific decision rules. The graded backbone (Qwen) is thresholded on log-odds with leave-one-conversation-out cross-fitting: thresholds are selected on the held-in conversations at a selection false-positive rate of at most 0.10 and every reported number is evaluated out-of-fold. The dense backbones produce near-binary score distributions that cannot support a tuned threshold, so Gemma and Llama are read at their own argmax operating point, $p(\textit{yes}) > p(\textit{no})$, with nothing fitted. The same asymmetry appears in deployment: at the untuned argmax point the graded backbone gates almost nothing on natural records (96–99% false admission, with only 2 of 400 records inside $|\text{log-odds}| < 1$), while the dense backbones' argmax points already gate. Comparisons are therefore made within a backbone, never across operating points.

Paired uncertainty. Policy contrasts on the natural records are paired at the record level. We report exact McNemar tests on the discordant pairs, together with $10{,}000$-replicate record-paired and conversation-clustered bootstrap intervals; the clustered intervals are the ones quoted in the main text. For the admission contrast the discordant split is 12:6 on 78 negative records, which is why that reduction is reported as directionally consistent but unresolved.

Determinism. Every scoring job runs two identical passes and records the gap rather than asserting bit-equality: median absolute log-odds difference per rerun is ≈ 0.2–0.5 on the mixture-of-experts backbone and $\approx 10^{-8}$ on the dense backbones. The one Qwen rail-audit flip above sits inside this recorded variation.

G Robustness and Protocol Controls

This appendix collects the controls referenced in Section 4.2. Table 4 varies how the same candidate memory and evidence are rendered: correct, isomorphic, and shuffled composition graphs, length-matched filler, and compact relations. Structure helps or hurts depending on backbone and rendering, and a shuffled graph can dominate the judgement entirely, so no representation is a universal repair. Table 5 reports the protocol-level controls: oracle upstream inputs, the equal-budget prompt search under which the top three interfaces converge to within four points while the winning interface flips across backbones, alternative aggregation rules over separately scored obligations (minimum 0.266, mean 0.101, noisy-AND 0.304 unsupported detection on Qwen—the operating point moves, the tension remains), and verification compute.

Table 6 gives the full accumulated-verification-noise results behind Section 4.2: adding obligations independently verified as true lowers support scores monotonically on every backbone, with backbone-scaled magnitude. Table 7 shows the two controls that locate the distributed-evidence paradox's mechanism: the verifier notices in-input support deletion and honours applicable corrections, and the paradox disappears when plausible premises are replaced by token-matched irrelevant content—the failure is plausible-premise composition bias, not context length or neglect of distant evidence.

Table 4: Representation robustness on fixed candidate memories and evidence. Values are unsupported-relation detection / multi-span valid retention.

Model	Representation	Unsupported Detection↑	Multi-span Retention↑	Observation
Qwen	Raw memory	0.608	0.846	holistic baseline
	Correct composition graph	0.468	0.923	no stable gain
	Isomorphic graph	0.354	0.974	rendering-sensitive
	Shuffled graph	0.430	0.846	graph form alone does not help
	Length-matched filler	0.544	0.846	length alone does not explain effect
	Compact relations	0.595	0.949	strong retention / detection balance
Gemma†	Raw memory	0.937	0.923	holistic baseline
	Correct composition graph	0.924	0.333	retention drops
	Isomorphic graph	0.924	0.154	rendering-sensitive
	Shuffled graph	505 / 512 records rejected		structure can dominate judgment
	Length-matched filler	0.734	0.923	length alone does not explain effect
	Compact relations	all 39 multi-span memories rejected		degenerate operating point
Llama baseline-only transfer				
Llama†	Holistic	0.595	1.000	valid rejection = 0.016
Llama†	Atomic claims	0.000	1.000	valid rejection = 0.000

†Gemma and Llama use their argmax operating points. Llama derivation-aware representation variants were not run in this suite.

Table 5: Protocol controls. Correct upstream inputs, prompt search, alternative aggregation, and additional verification compute change the operating point but do not provide a universal solution.

Oracle upstream inputs

Setting	Model	Full	Atom	Graph	Metric
Gold decomposition + gold evidence	Qwen	0.292	0.555	0.394	unsupported detection

Equal-budget prompt search

Interface	Qwen		Llama		Observation
	Default	Tuned	Default	Tuned	
Holistic	12.8	46.2	55.4	61.7	interface gap narrows
Obligation-aware	28.8	47.2	57.5	66.7	ranking changes
Predicate–argument QA	37.6	48.1	66.6	58.8	ranking changes

Aggregation over separately scored obligations

Model	Minimum	Mean	Noisy-AND	Metric
Qwen	0.266	0.101	**0.304**	unsupported detection

Verification compute

Interface	Model	Calls	Input tokens	Detection	Relative cost
Holistic	Qwen	512	201,766	0.608	1.0×
Per-obligation	Qwen	1,106	436,340	0.291	2.2×

Notes. Oracle-input results use a separate 235-member minimal-pair suite and should not be numerically compared with the 512-record main verification matrix. Equal-budget prompt search uses the same nine-configuration search budget for each interface.

H Downstream Causal Protocol

Design. The downstream probe changes exactly one memory slot available to a reader—correct, distorted, or absent—while holding the slot's retrieval position and all surrounding context fixed. Each example is evaluated on two queries: the *harm* query, whose answer the distortion changes, and the *content* query, which the record uniquely answers. Two causal quantities are computed, paired within example and never averaged: $\mathrm{CACE}_{\mathrm{harm}} = P(\mathrm{fail} \mid do(\mathrm{distorted})) - P(\mathrm{fail} \mid do(\mathrm{correct}))$ on the harm query (the cost of false admission) and $\mathrm{CACE}_{\mathrm{reject}} = P(\mathrm{fail} \mid do(\mathrm{absent})) - P(\mathrm{fail} \mid do(\mathrm{correct}))$ on the content query (the cost of false rejection). All six answer-option permutations are run and averaged so option position cannot carry the effect, with three replicates at temperature zero capturing server nondeterminism only.

Control floors gate every causal number. No causal quantity is read unless four control floors pass at 0.80, all-or-nothing: the reader must answer the content query correctly with the correct record present, must say the record is absent when it is absent, must recall correctly, and must not manufacture reasons. These floors can genuinely fail—an always-name policy fails the no-recall floor and an always-reason policy fails the no-reason floor—and every backbone passed all gating floors. Malformed responses are excluded as errors, never counted as data. The off-diagonal cells (distortion evaluated on the content query, removal on the harm query) are reported as specificity checks, not

Table 6: Accumulated verification noise. Supported memories receive k additional obligations independently verified as true; the candidate memory is unchanged, so any score decrease measures noise from checking more correct content, not detection of a defect.

Model	Median Δlog-odds $k{=}2$	Median Δlog-odds $k{=}4$	Records moving down ($k{=}4$)	Acceptance $k{=}0$	Acceptance $k{=}2$	Acceptance $k{=}4$
Qwen	-5.7	-7.7	98.7%	0.994	0.987	0.981
Gemma†	-43.3	-45.9	100.0%	0.933	0.121	0.026
Llama†	–	-20.0	99.0%	0.974	–	0.355

Notes. Fillers are drawn from other records of the same conversation whose expanded-scope verdict is supported and whose every resolved obligation was entailed; the $k{=}2$ list is a prefix of the $k{=}4$ list, so k is the only difference between conditions. The direction is universal and the magnitude is backbone-scaled; on the graded backbone the binary acceptance moves little while the underlying log-odds fall, which is why threshold-based deployments of the same verifier can behave very differently. †Argmax operating point. "–": condition not run on this backbone.

Table 7: Remote-history and mechanism controls (family-paired, Qwen verifier). Top: when the manipulation is inside the verifier's input, the verifier is competent—deleted support is noticed, redundant-path deletion never flips a decision, and a temporally applicable correction is honoured. Bottom: the same invalid record is accepted when its plausible premises are present but rejected when they are replaced by token-matched irrelevant history, isolating plausible-premise composition bias as the mechanism of the distributed-evidence paradox.

Remote-evidence interventions (69 families, five conditions)

Interface	Accept intact↑	Still accept after support deletion↓	Wrong flip after redundant deletion↓	False admit after applicable correction↓
Holistic	0.986	0.116	0.000	0.014
Obligation-aware	1.000	0.000	0.000	0.000
Predicate–argument QA	1.000	0.000	0.000	0.000

Mechanism control: plausible premises versus token-matched irrelevant content

Interface	Accept with plausible premises	Accept with irrelevant content	Paired Δlog-odds [95% CI]
Holistic	0.884	0.101	$+21.9$ [$+19.0$, $+25.5$]
Obligation-aware	0.884	0.000	$+29.9$ [$+27.7$, $+31.4$]
Predicate–argument QA	0.652	0.000	$+27.0$ [$+24.6$, $+30.7$]

Notes. Median paired Δlog-odds: full support deletion 28–37; redundant-path deletion 0.1–1.4; explicit correction 32–37. On Gemma the applicable-correction false-admit rate is 0.043–0.145. In the mechanism control both conditions present the same invalid record with distributed spans matched in length and position; only whether the spans carry premise content differs.

pooled into either estimate: a swapped holder makes the gold statement unrecorded and is expected to move on the content query, while a spurious connective leaves it recorded and is not.

Results across backbones. $\mathrm{CACE}_{\mathrm{harm}}$ is $+0.838$ (Qwen), $+0.930$ (Gemma), $+0.922$ (Llama); $\mathrm{CACE}_{\mathrm{reject}}$ is $+0.928$, $+0.995$, $+0.949$. Persistent two-sided consequences are a property of the setting, not of a model class. One profile difference is worth recording: with a distorted record on the content query, Llama abstains ("records neither") at 0.505 rather than answering.

I Additional Natural Cases

Table 8 complements the aspiration-to-activity case in Section 4.3 with four further unedited writes from the truth-labelled natural set, covering event-status mutation, cross-entity re-binding of durations, attributes, and causes, and a time-indexed refutation.

J Verification Interface Prompts

All interfaces share one scoring harness, one evidence rendering, and one verdict extraction; only the prompt below changes. Multi-unit interfaces score each unit with the per-unit template and aggregate by the minimum, since a record is licensed only if every obligation is. Placeholders in braces are filled per record.

Holistic.

Table 8: Additional unedited natural writes exhibiting derivation failures. Each row quotes the pre-write history and the resulting memory; verdicts are the adjudicated labels under the citation-only and expanded-history audits. The second case is the study's canonical illustration that citation matching and history-level validity are different targets: its cited spans pass the provenance check while the expanded audit finds the meaning unsupported.

Pre-write history says	**Memory says**	**Mutation**	**Cited verdict**	**Expanded verdict**
"I've been working on this car, doing engine swaps and suspension modifications. Now I'm *learning about* body modifications."	"…transforming it with engine swaps, suspension modifications, *and body modifications*."	prospective → performed	unsup.	unsup.
"What instrument are you playing?" — "I'm learning how to play the *violin* now…" / "How long have you been playing the *piano* again?" — "I've been playing for about *four months*."	"Tim has been learning to play the *violin* for about *four months*…"	duration re-bound across entities	supp.	unsup.
"I'm still just learning how to draw, but I love expressing myself *through writing*." / "I've been a bit *frustrated* lately with my *new phone*."	"Sam has recently taken up drawing as a new form of *self-expression* …despite occasional *frustration with his progress*."	attribute and cause re-bound	unsup.	unsup.
"I scored a deal to *continue* collaboration with Frank Ocean!"	"Calvin expressed excitement about his *new* collaboration with Frank Ocean…"	ongoing → new (refuted)	contra.	contra.

Notes. All four records are unedited writer outputs from the truth-labelled natural set; none was constructed for this table. Emphasis added. Together with the aspiration-to-activity case in the main text, these span the mutation families the obligation decomposition targets: event status (prospective versus performed), binding (attributes, durations, and causes attached to the wrong entity), and time-indexed refutation.

```
You are checking whether a memory record is licensed by
conversation evidence.
RECORD: {record}
EVIDENCE ({n} spans):  {evidence}
Is the record as a whole licensed by this evidence?  A
record is licensed only if every claim it makes, and every
relation it asserts between claims, is supported.  If it
asserts a relation, the evidence must license the relation
itself, not merely each side separately.
Answer with one word, yes or no.
```

Atomic claims (per unit).

```
You are checking whether ONE claim from a memory record is
licensed by conversation evidence.
CLAIM: {unit}
EVIDENCE ({n} spans):  {evidence}
Is this claim licensed by the evidence?  Judge only this
claim, not the rest of the record.
Answer with one word, yes or no.
```

Residual (composition beyond listed claims).

```
You are checking a memory record for anything it asserts
BEYOND the individual claims listed below.
RECORD: {record}
CLAIMS ALREADY CHECKED SEPARATELY: {parts}
EVIDENCE ({n} spans):  {evidence}
Assume each listed claim is licensed.  The question is only
about what the record asserts in addition:  a relation
it draws between claims, a modifier or circumstance it
attaches, or a binding of an attribute to a particular
person.  Is everything the record asserts beyond the listed
claims licensed by the evidence?
Answer with one word, yes or no.
```

Composition graph.

```
You are checking whether a memory record is licensed by
conversation evidence.
RECORD: {record}
ITS STRUCTURE: claims it makes: {claims} plus whatever the
record asserts to connect, modify or bind them
EVIDENCE ({n} spans): {evidence}
The record is licensed only if every claim above is licensed
AND everything the record asserts to connect, modify or
bind those claims is licensed. Evidence that supports the
claims separately does not by itself license a relation
drawn between them, a circumstance attached to them, or
which person an attribute belongs to.
Answer with one word, yes or no.
```

Compact relations.

```
You are checking whether a memory record is licensed by
conversation evidence.
The record, written out as the claims it makes and the links
it draws between them: {relations}
EVIDENCE ({n} spans): {evidence}
Every line above must be licensed for the record to be
licensed, including the lines that link claims together.
Evidence supporting the linked claims separately does not
license a link drawn between them.
Answer with one word, yes or no.
```

Obligation-aware joint (admission gate).

```
You are checking whether a memory record is licensed by
conversation evidence.
RECORD: {record}
The record asserts, at minimum, each of these obligations:
{obligations}
EVIDENCE ({n} spans): {evidence}
Considering every obligation against the evidence together:
is the record licensed by the evidence? Answer with one
word, yes or no.
```

Predicate–argument QA. The QA interface derives predicate–argument questions from the record and scores each question with the full evidence block; both decision flows are evaluated: one joint verdict over all questions, and the faithful per-question variant with minimum aggregation. Every call in every interface sees the identical evidence block—decomposing interfaces restrict the claim, never the evidence.

K Implementation and Reporting

Experiment settings. Table 9 records the writer, verifier, and reader configurations, serving parameters, and evidence budgets used by each experiment.

Reporting checklist. Every result table must report sample size, sampling frame, denominator, history-level split, threshold-selection split, actual test FPR, model and prompt version, retrieval scope and budget, number of development trials, cluster unit, interval construction, tie rate, latency, calls, and tokens where applicable. Within-pair ranking and standard ROC-AUC are named and reported separately. Unclipped label log-probability differences are used to audit score saturation; the number of unique scores and pairwise ties accompanies each ranking result.

Table 9: Evaluation suites used to study evidence scope, compositional validity, admission reliability, and later memory use.

Evaluation	Agent-memory question	Scale	Primary comparison
Natural paired audit	Is a memory unsupported, or is writer provenance incomplete?	400 writes	Cited evidence $\rightarrow$ expanded pre-write evidence
Natural admission replay	Which supported and unsupported memories survive write-time verification?	391 writes	Citation-only / expanded history / composition-aware
Matched distribution	Can a verifier distinguish licensed memory from plausible composition when support is distributed?	70 families $\times$ 4	Local/distributed $\times$ valid/invalid
Semantic verification	Which representation best preserves relations asserted by a memory?	512 records	Prior paradigms vs. derivation-aware interfaces
Remote-history controls	Are failures caused by ignored history or by plausible but non-licensing premises?	Matched interventions	Necessary-path removal / redundant-path removal / correction
Oracle diagnostics	Do errors remain with correct evidence and decomposition?	235 records	Gold evidence + gold decomposition
Supported-filler intervention	Can additional correct checks cause valid memory to be rejected?	391 labelled records	$k = 0, 2, 4$ verified-true added obligations
Natural gate replay	Does verification exert selection pressure on naturally written memory?	663 writes	Single-span licensor present / absent
Memory lifecycle probes	How do distorted or missing memories affect later agent behavior?	39 pairs + 35 retrieval items	Correct / distorted / absent memory