EVIDENCE IS NOT AUTHORITY: A CAUSAL STUDY OF STATE ARBITRATION IN LONG-HORIZON AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-horizon agents can retrieve both historical and current user-state evidence yet still act on the historical state. Because existing memory benchmarks primarily score final answers, they often cannot distinguish a lost update from one that was visible but ignored. We call this choice among jointly visible states *state arbitration* and study it through an evidence-to-action audit and paired interventions. Across 240 paired histories and three model families, retaining an early explicit declaration while holding later behavior fixed lowers current-state accuracy by 33–39 points. Our primary intervention changes only the lifecycle relation between two records already in context. Marking the current record active and the old record superseded raises current-state actions from 24% to 53–55% on two 70B-class models. Natural validity/provenance fields also help, whereas timestamps alone do not. The contrast also transfers to structured action versions of STALE, LongMemEval, and MemoryAgentBench, but incorrect labels can reverse the result: marking a valid record superseded reduces one model's correct actions from 241 to 8 of 400. In native paths, managers often fail either to expose new evidence or to link it to the state it replaces. These findings identify a read-time failure distinct from retrieval and suggest preserving history while making the currently applicable record explicit and inspectable.

1 INTRODUCTION

When a user's state changes, which part of an agent's memory determines what it does next? Consider a user who once asks for Go, later works repeatedly in Python, and never announces a preference change. An agent may retrieve both observations and still write Go. The question is not only whether the new evidence is present, but whether the memory representation lets it guide the action.

We call this problem *state arbitration*. A memory pipeline preserves evidence, retrieves it, and may materialize a normalized state. At the action boundary, however, an old state and a current state can be visible at the same time. Retrieval answers "what is present?"; arbitration answers "which state governs the output?" Existing evaluations often combine these stages into a single score, making it difficult to distinguish missing evidence from a current state that loses control beside an older one.

We examine this distinction with an evidence-to-action audit and paired interventions. The audit records four checkpoints—retention, retrieval, materialization, and arbitration—as well as the exact context sent to the host model. After the relevant state has been materialized and made visible, the primary intervention changes only how the lifecycle relation between the two records is expressed.

The declaration intervention first establishes the phenomenon. Across 240 paired histories and three model families, retaining an early declaration while keeping the later behavioral trace unchanged lowers current-state accuracy by 32.9–38.8 points. In structured code, naming, format, and tool tasks, the same declaration increases stale actions by 20.0–25.0 points. We then change the labels attached to two jointly visible records. Across 192 prospectively selected drift contexts per host, accuracy on the current value rises by 28.6 points for Qwen-72B and 31.3 points for Llama-70B. This result is not confined to literal status words: natural validity/provenance fields yield a pooled gain of 12.0 points, whereas timestamp-only metadata yields -0.5 points.

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.

Native audits show that current memory managers often fail before arbitration. Under fixed lexical diagnostics, new evidence is detectable in only 5/400 audited LangMem drift histories and in 387/400 for Mem0, but neither manager reliably links the newer record to the older one under the audited schema. Supplying that relation at read time adds 24.7 points on the native panel. Retrieval alone is therefore insufficient: the final context must make clear which retrieved record should be followed.

Our contributions are:

- We define an evidence-to-action audit that separates retention, retrieval, materialization, and arbitration, making the source of stale action observable.
- Our primary causal result is a conditional read-time effect: with old and current records held fixed and jointly visible, lifecycle roles change which state controls action across prospectively selected contexts and two hosts. Task-stratified and six-schema analyses characterize its heterogeneity and robustness.
- We test the same interface in native memory paths and three external task families, and measure its safety boundary with stable-state and reversed-label controls.

The practical implication is to retain historical records for provenance while making current validity or supersession explicit before generation. Evaluation should then follow the resulting action: retrieved evidence may still be ineffective when the relation between current and historical records remains unspecified.

2 Related Work

Evaluating memory under updates. LoCoMo and LongMemEval test multi-session conversational memory, with LongMemEval further separating indexing, retrieval, and reading through oracle-evidence conditions (Maharana et al., 2024; Wu et al., 2025). MemoryAgentBench is our primary evaluation anchor: it presents information incrementally and evaluates learning, retrieval, understanding, and forgetting; its FactConsolidation task includes ordered edits for which later information should take precedence (Hu et al., 2026). Broader suites cover interdependent tasks, multiple evolving entities, and evolving preferences (He et al., 2026; Jung et al., 2026; Li et al., 2026). These benchmarks establish that storing or retrieving an update does not guarantee its correct use. Our focus is narrower: once a correct current record and its historical counterpart are both visible, does an explicit lifecycle relation change the structured action? Transformed MemoryAgentBench and LongMemEval items test this contrast separately from their official evaluations.

Forming and representing evolving state. Implicit-personalization methods infer latent preferences from dialogue or user behavior (Peng et al., 2025; Tan et al., 2025; Guo et al., 2026), while concept-drift models characterize how such targets change (Gama et al., 2014; Caroprese et al., 2025). Belief-revision research asks when new evidence should alter an existing belief base (Alchourrón et al., 1985; Hase et al., 2024; Wilie et al., 2024); temporal memory methods instantiate related ideas with timestamps, learned operations, or structured temporal reasoning (Ge et al., 2025; Du et al., 2025; Su et al., 2026; Xu et al., 2026). STALE studies implicit invalidation, distinguishes update recognition from downstream adaptation, and explores lifecycle-aware adjudication and readout (Chao et al., 2026). Against this background, we measure how much a supplied lifecycle relation changes the first structured action, conditional on a correct, jointly visible record pair. This design separates use of the relation from its upstream assignment, without assuming that a deployed manager can construct it automatically.

Control points in memory systems. MemGPT/Letta, Mem0, A-MEM, and LangMem combine extraction, storage, evolution, and retrieval differently (Packer et al., 2023; Chhikara et al., 2025; Xu et al., 2025; LangChain, Inc., 2025); their captured outputs let us audit whether evidence is retained, exposed, or related. StateAuditor checks drafts against dated evidence and state dependencies (Sun & He, 2026), while lifecycle governance constrains readout before generation (Chao et al., 2026; Lam et al., 2026). We compare control points rather than rank architectures: native captures diagnose upstream paths, our paired intervention isolates relation use before action, and response audits operate afterward.

3 From Evidence to Action: State Arbitration at the Action Boundary

Let $z_T^a \in \mathcal{V}_a$ denote the generator-defined current value of user attribute a at decision time T. We represent a memory system as an evidence-to-action pipeline:

$$E_{1:T}^a \longrightarrow M_T \longrightarrow R_T \longrightarrow P_T \longrightarrow Y_T, \tag{1}$$

where M_T is the auditable store, R_T the retrieved records, P_T the exact context given to the host model, and Y_T the generated answer or artifact. We audit four checkpoints:

Checkpoint	Governance question	Observable signal
Retention	Does source evidence survive in the store?	Evidence-bearing record in M_T
Retrieval	Does the evidence reach the host?	Evidence span in P_T
Materialization	Does the context expose a usable current state?	Normalized record with target value
Arbitration	Which jointly visible state controls the action?	Action uses current or historical value

The first three checkpoints concern whether a state reaches the action boundary. Arbitration concerns which available state the output follows. Thus, retaining and retrieving evidence need not prevent a stale action.

At the action boundary, the context may contain an old record $r^{old} = (z^{old}, m^{old})$ and a current record $r^{cur} = (z^{cur}, m^{cur})$, with $z^{old} \neq z^{cur}$. Here m denotes lifecycle metadata, including status, validity, supersession, and provenance fields. A neutral rendering presents the records as peers with no authority relation. A lifecycle-labeled rendering keeps record content, order, task, host instruction, model, and scorer fixed, but marks r^{cur} as `ACTIVE` and r^{old} as `SUPERSEDED`. The robustness panel expresses the same relation through ordinary validity/provenance fields, explicit supersession links, and timestamps.

We say that a visible record has *action authority* when the structured output uses its value. State arbitration maps the visible records and their lifecycle representation to the action selected by the host. Figure 1 shows the paired intervention: the old and current strings remain visible, and only their lifecycle representation changes.

To separate the mechanisms along this path, we use three paired estimands. Each contrast keeps the scenario, task, candidate space or action evaluator, and host fixed within a pair. The declaration contrast measures content-level anchoring by comparing an old declaration with a token-matched neutral replacement while the later behavioral trace is identical:

$$\Delta_{decl} = \mathbb{E}[\mathbf{1}(Y^{neutral} = z^{cur}) - \mathbf{1}(Y^{old} = z^{cur})]. \tag{2}$$

The lifecycle contrast isolates authority transfer after the current state has been materialized and made visible beside its historical counterpart. Let $C = 1$ denote this eligibility gate: the context contains a correct normalized current record and a jointly visible old record. The read-time representation effect is

$$\Delta_{life} = \mathbb{E}[\mathbf{1}(Y^{lab} = z^{cur}) - \mathbf{1}(Y^{neu} = z^{cur}) \mid C = 1]. \tag{3}$$

The stable endpoint applies the same authority contrast when the underlying value z does not change. Stable controls label the records `ACTIVE` and `HISTORICAL_VALID`:

$$\Delta_{stable} = \mathbb{E}[\mathbf{1}(Y^{lab} = z) - \mathbf{1}(Y^{neu} = z) \mid stable]. \tag{4}$$

The three estimands address successive questions. Δ_{decl} measures whether earlier content anchors the model against later evidence; Δ_{life} measures whether lifecycle labels redirect action once both states are visible; and Δ_{stable} tests preservation when the state has not changed. The primary intervention neither removes historical evidence nor changes the current record; it specifies which record applies at action time. Native audits examine whether the required evidence and relation were formed upstream.

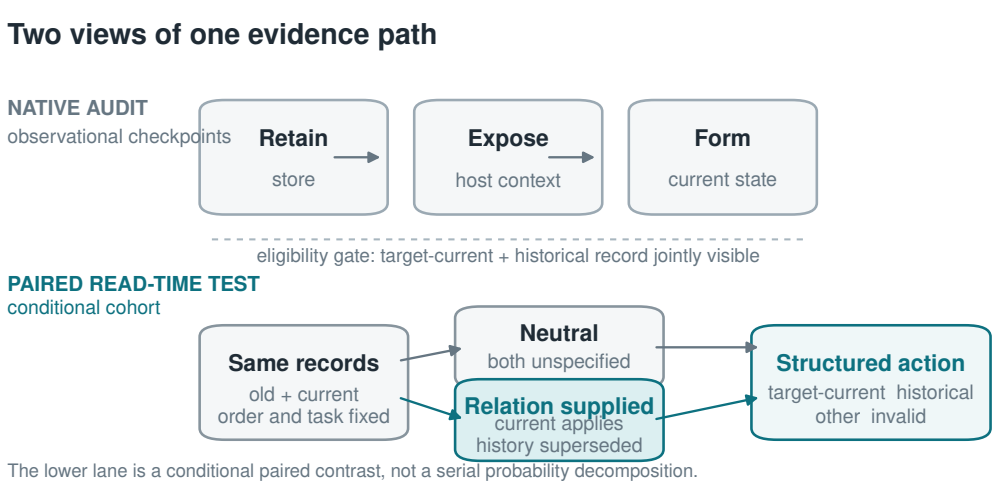


Figure 1: State arbitration at the action boundary. The old and current strings remain jointly visible; the paired intervention changes only their lifecycle roles.

4 Benchmark and Study Design

We construct the benchmark as a controlled data-generating process that separates evidence availability from action authority. Code fixes the latent state trajectory, renders the behavioral evidence observed by the model, and generates the final decision task. This construction yields deterministic counterfactual pairs in which a declaration or lifecycle relation changes while the underlying trajectory, task, candidate space, host, and evaluator remain fixed. We can therefore identify where a fixed body of evidence ceases to guide the action, rather than merely enumerate stale-memory failures.

4.1 Controlled state trajectories

Each scenario instantiates one user attribute over 15 sessions. Sessions 0–5 establish and reinforce an initial value, sessions 6–11 provide behavioral evidence for a different value without an explicit revision announcement, and sessions 12–14 probe the state that should govern the final decision. This establishment–drift–probe structure makes the historical state salient and well supported, while only the later behavioral trace matches the generator-defined state at decision time. Success therefore requires enough evidence to distinguish the temporal roles of the two periods.

We cross 50 profiles with eight attributes and three drift patterns, producing 1,200 scenarios from roughly 170 behavioral templates. The attributes span programming and naming choices, serialized output formats, tool and interaction preferences, communication style, and error handling. They are chosen so that a latent state can be expressed through ordinary behavior and later tested through a concrete choice or executable action. Abrupt trajectories switch cleanly to the new value, gradual trajectories distribute the transition across later sessions, and noisy trajectories interleave limited contradictory evidence. These regimes vary the shape of the post-change evidence while leaving the target semantics and scoring rule unchanged.

Profiles determine surface details, while the attribute value and drift pattern determine the state trajectory. A scenario is the basic unit of generated evidence; a model–scenario pairing is the unit of response; and the same generated item across two intervention conditions forms the unit of paired comparison. Profile-disjoint panels are used where stated below so that robustness estimates do not depend on reusing the same surface identities.

4.2 Outcome semantics and deterministic scoring

For each drifting scenario, the generator draws distinct pre- and post-drift values from a closed, attribute-specific value space. We call the post-drift value *target-current*, the exact pre-drift value *historical-state*, another member of the valid value space *other-valid*, and an output outside the task

specification *invalid*. Here "current" is an experimental role defined by the scenario graph, not a claim that real human preferences admit a unique latent truth or can be inferred without ambiguity.

Models receive only the rendered history and task; generator state is never exposed to them. Deterministic parsers map both value probes and structured actions into the four outcome classes. For an action to count as target-current or historical-state, it must both select the corresponding value and satisfy the task's structural constraints. Malformed code, invalid serialization, and other task failures therefore remain invalid rather than being credited as stale behavior. This distinction prevents state selection from being conflated with the ability to produce a well-formed artifact.

No LLM generates benchmark truth or judges a primary outcome. The same parser and task-specific validity checks score both members of every pair. Thus a paired effect reflects a change in the model's response to the manipulated representation, rather than a change in the reference answer or evaluator.

4.3 Experimental populations

The studies form a staged evaluation of the evidence-to-action path rather than a single pooled benchmark score. The *declaration study* first tests content anchoring: it evaluates the same 240 histories from 50 profiles with three model families, then repeats the contrast on a separate 120-history action panel (Appendix F). This stage asks whether an early explicit statement can dominate a later behavioral trace even before the failure is attributed to a particular memory implementation.

The *lifecycle cohort* is the primary test of conditional read-time arbitration. For each host, it contains 192 drift contexts selected by prespecified rank from the 494 of 576 generated candidates that satisfy the eligibility gate $C = 1$: a correct normalized current record and its historical counterpart are jointly visible at the action boundary. Conditioning on this gate is intentional. It holds upstream availability fixed and asks whether making the records' authority relation explicit changes the action. A separate set of 128 history-preserving stable controls per host tests whether the same mechanism preserves, rather than disrupts, an unchanged state (Appendix D).

The *representation robustness panel* evaluates whether the lifecycle effect depends on one privileged notation. It uses a prespecified, profile-disjoint set of 96 drift and 32 stable contexts per host under six schemas, selected without reference to action outcomes. The *native audit* then moves upstream: for each of Mem0 and LangMem, it processes 400 drift histories and 400 matched stable STALE histories with Qwen and Llama as hosts, measuring whether current evidence and lifecycle relations are formed by the systems' own write and retrieval paths before any controlled read-time rendering.

Finally, the *external suite* tests the same intervention on different source tasks. It uses all 400 STALE items, 50 of 78 LongMemEval knowledge-update pairs, and the 74 of 100 MemoryAgentBench FactConsolidation questions whose truth and action targets can be grounded mechanically. These external results are treated as validation under different source tasks, not pooled with the controlled benchmark estimands. The next section specifies the interventions, instrumentation, and statistical analysis; the appendices give construction rules, selection procedures, and full protocols.

5 Operationalizing Evidence-to-Action Governance

Section 3 traces memory behavior from evidence availability to action authority. We measure each stage separately and use paired contrasts that change one control point while keeping the downstream evaluator fixed.

Stage-level observability. For each history, the store inventory measures retention, retrieved records measure retrieval, and the exact final host input defines the decision context P_T for evaluating materialization and arbitration. The generated response and deterministic action parse identify which visible state controls the output. Source-provenance rules distinguish behavioral current evidence from the opening declaration. These observables locate failures along the evidence-to-action path; paired conditions retain the same task, host instruction, model, and scorer.

Paired intervention suite. The primary read-time intervention keeps old/current records, order, task, host instruction, model, and scorer fixed while changing only their lifecycle relation from neutral peers to `ACTIVE`/`SUPERSEDED`; its paired difference estimates Δ_{life}. The declaration intervention

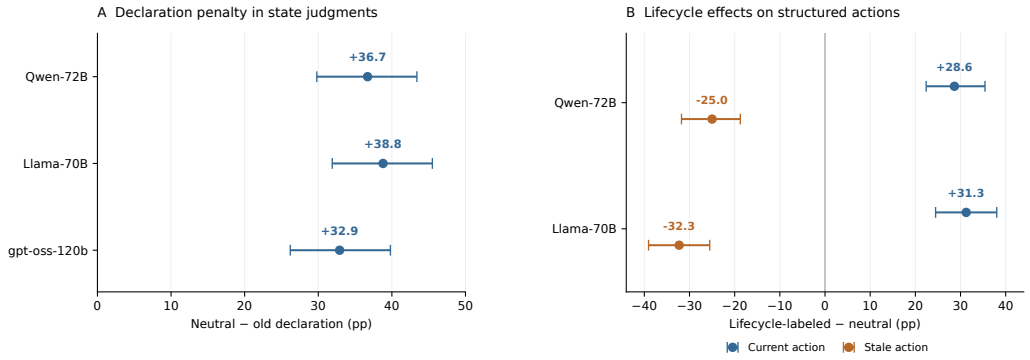


Figure 2: The two causal contrasts. Earlier declarations still shift judgments when later behavior is fixed; lifecycle labels then redirect action while old and current records remain unchanged.

instead replaces the early statement with token-matched neutral text while holding later sessions fixed, with constrained-choice, turn-ablation, positive-control, and action panels testing content anchoring (Appendix F). An availability $\times$ stale-suppression factorial appends a correct current record or masks old-value spans (Table 11), separating evidence availability from the authority contrast.

Robustness and transfer. A prespecified, profile-disjoint panel tests six lifecycle schemas with a fixed parser and neutral baseline. Native captures preserve each memory system's complete write/retrieval context before paired rendering, allowing separate measurement of evidence availability, relation formation, and downstream sensitivity. External items keep evidence, task, and evaluator fixed while lifecycle roles vary, testing transfer beyond the synthetic cohort.

Scoring and inference. Primary outcomes are paired differences in current-state action accuracy, with exact stale-action changes reported alongside them. Profile- or scenario-cluster bootstrap intervals use 10,000 replicates; the two host-specific lifecycle effects receive exact paired tests with Holm correction. Stable controls use a prespecified -5-point noninferiority margin with conservative exact paired bounds. Scenario truth, action parsing, and primary scoring are deterministic; parser sensitivity and protocol audits are reported in the appendix.

Closed-schema behavioral aggregation measures whether the current state is inferable from the history before memory and host arbitration, separating unavailable evidence from evidence that loses authority at the action boundary.

6 Results: Read-Time Representation Controls Action

The experiments ask one question: when old and current state are both visible, what makes one of them control the action? Figure 2 previews the two paired contrasts. We first measure declaration anchoring, then test lifecycle labels across tasks, schemas, and native paths. Stable, reversed-label, and external tests assess preservation, failure under incorrect labels, and transfer.

6.1 Phenomenon: explicit declarations remain authoritative

In the paired declaration intervention, later behavioral sessions, the probe, and candidate order are identical; only the early declaration is retained or replaced with token-matched neutral text. Across 240 histories and three model families, retaining the declaration lowers current-state accuracy by 32.9–38.8 points (Table 8). The penalty persists when assistant turns are removed (34.2–46.2 points) or neutralized (34.2–45.8), and an explicit current declaration restores accuracy by 50.4–71.7 points. In a separate 120-history action panel, retaining the declaration increases outputs that follow the old preference by 20.0–25.0 points across code, naming, format, and tool tasks. A closed-schema behavioral aggregator recovers the target-current value on 80.9% of the 1,200 histories, so the behavioral signal is usually present before arbitration.

Table 1: Lifecycle-label intervention on 192 paired contexts per host. Counts are current/exact-stale actions; other and invalid outcomes remain separate. Gains use paired 95% bootstrap intervals.

Host	Neutral	Lifecycle	Δ current (pp, 95% CI)
Qwen-72B	47/123	102/75	$+28.6$ [22.4, 35.4]
Llama-70B	46/104	106/42	$+31.3$ [24.5, 38.0]

Table 2: Representation robustness on the six-schema panel (96 drift contexts per host, pooled over two hosts). The panel has its own neutral baseline, so its rates are not comparable to Table 1.

Representation	Current	Exact stale	Current Δ
Neutral	49.5%	32.8%	–
Literal status	58.3%	22.4%	+8.9
Timestamp only	49.0%	33.3%	-0.5
Supersedes relation	52.6%	28.6%	+3.1
Natural validity/provenance	61.5%	18.8%	+12.0
Host-level recency rule	70.8%	9.4%	+21.4

6.2 Primary causal result: lifecycle representations control action

On the 192 contexts per host that pass the eligibility gate, the neutral condition produces 47 current-state actions for Qwen and 46 for Llama. Labeling the unchanged records `ACTIVE`/`SUPERSEDED` raises these counts to 102 and 106. The paired gains are $+28.6$ points [22.4, 35.4] and $+31.3$ [24.5, 38.0] (Table 1); both tests remain significant after Holm correction ($p < 10^{-15}$). Lifecycle labels convert a substantial fraction of stale decisions into current-state actions: Qwen has 55 favorable and no adverse current-action flips, while Llama has 61 favorable and one adverse flip. Exact stale actions fall by 25.0 and 32.3 points, respectively. Residual stale actions remain, so the labels help but do not replace downstream reasoning.

Task heterogeneity. The naming and programming families (96 contexts) show current-action gains of 43.8 points [34.4, 54.2] for Qwen and 45.8 [35.4, 56.3] for Llama. Output-format gains are 25.0 and 22.9 points, and tool preference is directionally positive for both hosts (2.1 points [0.0, 6.3] for Qwen and 10.4 [2.1, 18.8] for Llama). Gains are largest for naming and programming and positive across all four action families; pilot details appear in Appendix D.

Encoding beyond literal status labels. In a separate panel of 96 contexts per host with its own neutral baseline (Table 2), natural validity/provenance fields add 12.0 points pooled over the two hosts (+19.8 Qwen, +4.2 Llama; paired $p = 5.65 \times 10^{-6}$), whereas timestamp-only metadata changes action by -0.5 points. The gain comes mainly from Qwen: Llama's neutral accuracy in this panel is already 63.5%, and no schema moves it by more than 5.2 points. The host-level recency rule is exploratory because it changes instruction guidance; full schema and stable-control results appear in Appendix E. Single-field ablations and component interactions are reported in Appendix D.

6.3 Native validation and upstream diagnosis

Conditional native closure. The native-manager test asks whether the representation effect remains when a memory manager produces the record. In 120 independent LangMem drift histories, the manager itself forms and retrieves a correct current/previous relation in 24. Holding each complete context fixed byte-for-byte, the manager's own lifecycle-labeled record yields 22/24 current actions for Qwen and 21/24 for Llama, versus 17/24 and 16/24 when that record is rendered neutrally: +20.8 points for each host ([8.0, 34.8] and [4.3, 40.0]). Exact stale actions fall from 3 to 0 for Qwen and from 1 to 0 for Llama. This contrast is conditional on correctly formed native relations; its unconditional coverage is 24/120.

Cross-system native paths. We next audit the complete native path for Mem0 and LangMem on 400 STALE drift histories. The managers receive no benchmark truth, candidate values, or lifecycle labels. We measure evidence availability and native action unconditionally.

Table 3: Native-path diagnostics. New-evidence and native-relation counts use fixed lexical and token-overlap diagnostics. Native-action entries are current-action accuracy for Qwen/Llama using each manager's native memory context.

System	New evidence	Native relation	Native action (Qwen/Llama, %)
LangMem	5/400	0/400	0.8/2.5
Mem0	387/400	0/400	42.0/87.8

After the native run, we render the benchmark's own old/new observations (authored evidence) in neutral and lifecycle-labeled form to measure downstream sensitivity.

Under these diagnostics, the two systems fail at different upstream stages: the later observation is detectable in 5/400 LangMem drift histories but 387/400 Mem0 histories, and neither forms a grounded current/previous relation under the audited schema. Because the diagnostics match surface text, they indicate where evidence is detectable rather than everything a manager encodes. Replacing each manager's retrieved context with authored evidence raises current-state action accuracy by 40.3 points [37.6, 42.9], and adding lifecycle labels to that same evidence adds a further 24.7 points [21.6, 27.8] on the held-out evaluation set ($n = 390$). The broader stage-localization comparison appears in Appendix G.

6.4 Stable-state preservation

On 128 history-preserving stable controls, Qwen changes from 112/128 to 109/128 (-2.3 points [-5.5, 0.0]) and Llama from 97/128 to 98/128 ($+0.8$ [0.0, 2.3]). Qwen's three adverse flips are not significant under a two-sided McNemar test ($p = .25$), but Qwen does not establish the prespecified 5-point noninferiority bound: its conservative lower bound is -7.4 points. Llama passes with a lower bound of -3.4 points (Table 5).

6.5 External transfer and reversed-label sensitivity

We apply the read-time intervention to STALE (400 items), LongMemEval knowledge-update (KU) pairs (50 of 78), and MemoryAgentBench FactConsolidation (FC) questions (74 of 100), keeping evidence, task, and action evaluator fixed. Figure 3 shows the results. Correct lifecycle roles improve all six benchmark–host cells, although the gains vary widely: MemoryAgentBench gives +75.0/+70.3 points for Qwen/Llama, LongMemEval +3.0/+5.0, and STALE +63.7/+0.5 (Qwen/Llama; Llama is near ceiling on STALE drift items, 796/800 correct without labels). Reversing the labels lowers accuracy in all six cells, and wrong labels are costly: on STALE controls whose state never changed, marking the valid record `SUPERSEDED` lowers Qwen's action readiness (acting on the still-valid state) from 398/400 to 368/400, and Llama's from 241/400 to 8/400.

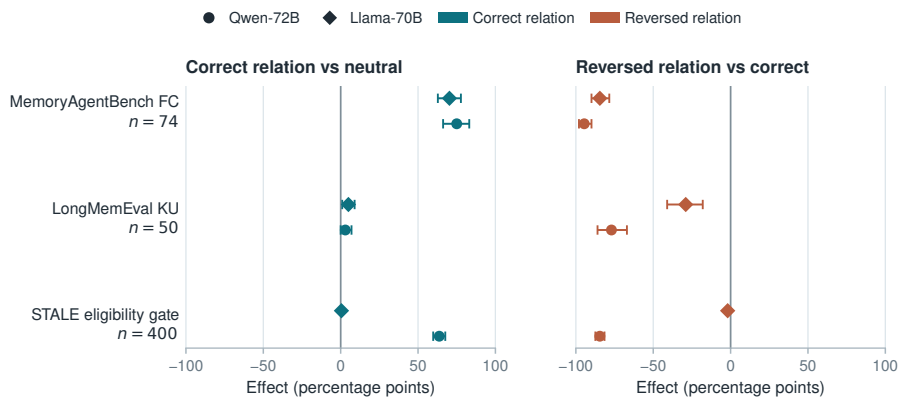


Figure 3: Lifecycle effects and reversed-label sensitivity on three external update-task families. Effects vary substantially across benchmarks and hosts.

7 Discussion and Conclusion

Three findings connect stale action to different stages of the memory path. An explicit declaration can retain influence after later behavior supports another state. Given both records, lifecycle representation transfers authority between jointly visible states without adding or removing factual evidence. Native memory paths may fail earlier by omitting new evidence or leaving it unlinked to the older record. Evidence availability and the choice among available records are therefore separate sources of error.

An evidence plane and an authority plane. An evolving memory needs source-level history for provenance and audit, but the final context should also indicate which record currently applies and which record it replaces. Destructive overwrite loses provenance; presenting every retrieved record as an undifferentiated peer leaves the choice to wording and model priors. A lifecycle relation can preserve the historical record while clarifying its present applicability. Recency alone cannot do this when observations are noisy, constraints persist, or records govern different scopes.

Lifecycle metadata is a governance control plane. The labels that correct stale action can also impose an incorrect state when the relation is wrong. Assignments should therefore carry provenance and confidence, be logged when conflicts occur, and have restricted write access. When the evidence does not justify a unique current state, a manager should preserve competing hypotheses and defer, ask for confirmation, or scope the relation to a particular task. The reversed-label results show why these safeguards matter: lifecycle fields directly influence action selection.

Implications for evaluation. End-to-end accuracy alone cannot distinguish an update that was never retained from one that was visible but overruled. Memory evaluations should report evidence availability, whether the newer record is linked to the older one, the final context, and the resulting action, together with stable-state preservation and wrong-label sensitivity. The positive effects on STALE, LongMemEval, and MemoryAgentBench suggest that the same issue appears beyond our generated trajectories, but natural histories, open-set and multi-user state, interacting attributes, and independently implemented lifecycle interfaces remain to be tested.

Long-horizon memory is not only a problem of remembering more. It is a problem of governing which remembered state is allowed to control action. That decision should be explicit, inspectable, and safe to override.

Ethics Statement

Behavioral preference inference creates privacy and autonomy risks. Temporary activity may be mistaken for durable state, and silently inferred preferences may reveal or impose information a user did not intend to store. Systems should provide provenance, inspection, correction, deletion, and opt-out controls and should not silently override explicit constraints. Because supplied lifecycle labels strongly steer action, write access to such fields is security-relevant; deployments should restrict and log who can set them. All benchmark histories are synthetic.

Reproducibility Statement

Section 4 and the appendices describe scenario generation, scoring, serving settings, and study-specific protocols. We release an anonymized artifact containing the deterministic benchmark generator, templates, evaluation and analysis code, system adapters, exact final prompts, raw model responses, and material protocol deviations. Scenario generation requires no LLM call, primary scoring is deterministic, and all model calls use temperature zero.

AI USE STATEMENT

We used generative AI tools to provide feedback on research framing, conceptual presentation, methodology, and experiments; assist with implementation and artifact auditing; help clean, prepare, and format data; and edit author-drafted prose for readability. We did not use generative AI to generate synthetic benchmark instances or benchmark truth, formulate mathematical claims or proofs, propose hypotheses, translate research material, perform qualitative or thematic analysis, judge primary outcomes, interpret or explain experimental results, or draft paper sections. Models used as systems under test and, in one labeled control, as session-level evidence extractors are described in the paper. The authors selected the protocols, verified the code, results, and citations, reviewed all AI-assisted material, and take responsibility for the final text, claims, and artifacts.

REFERENCES

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *The Journal of Symbolic Logic*, 50(2):510–530, 1985.

Luciano Caroprese, Francesco Sergio Pisani, Bruno M. Veloso, Maximilian König, Giuseppe Manco, Holger H. Hoos, and João Gama. Modelling concept drift in dynamic data streams for recommender systems. *ACM Transactions on Recommender Systems*, 2025. doi: 10.1145/3707693. Published online December 2024.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. STALE: Can LLM agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready AI agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Yiming Du, Baojun Wang, Yifan Xiang, Zhaowei Wang, Wenyu Huang, Boyang Xue, Bin Liang, Xingshan Zeng, Fei Mi, Haoli Bai, Lifeng Shang, Jeff Z. Pan, Yuxin Jiang, and Kam-Fai Wong. Memory-T1: Reinforcement learning for temporal reasoning in multi-session agents. *arXiv preprint arXiv:2512.20092*, 2025.

João Gama, Indre Zliobaite, Albert Bifet, Mykola Pechenizkiy, and Abdelhamid Bouchachia. A survey on concept drift adaptation. *ACM Computing Surveys*, 46(4):44:1–44:37, 2014. doi: 10.1145/2523813.

Yubin Ge, Salvatore Romeo, Jason Cai, Raphael Shu, Yassine Benajiba, Monica Sunkara, and Yi Zhang. TReMu: Towards neuro-symbolic temporal reasoning for LLM-agents with memory in multi-session dialogues. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18974–18988, 2025. doi: 10.18653/v1/2025.findings-acl.972. URL https://aclanthology.org/2025.findings-acl.972/.

Qianyun Guo, Yibo Li, Yue Liu, and Bryan Hooi. Towards realistic personalization: Evaluating long-horizon preference following in personalized user-LLM interactions. *arXiv preprint arXiv:2603.04191*, 2026.

Peter Hase, Thomas Hofweber, Xiang Zhou, Elias Stengel-Eskin, and Mohit Bansal. Fundamental problems with model editing: How should rational belief revision work in LLMs? *Transactions on Machine Learning Research*, 2024. arXiv:2406.19354.

Zexue He, Yu Wang, Churan Zhi, Yuanzhe Hu, Tzu-Ping Chen, Lang Yin, Ze Chen, Tong Wu, Siru Ouyang, Zihan Wang, Jiaxin Pei, Julian McAuley, Yejin Choi, and Alex Pentland. MemoryArena: Benchmarking agent memory in interdependent multi-session agentic tasks. *arXiv preprint arXiv:2602.16313*, 2026.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.05257.

Seokwon Jung, Alexander Rubinstein, Arnas Uselis, Sangdoo Yun, and Seong Joon Oh. MEME: Multi-entity & evolving memory evaluation. *arXiv preprint arXiv:2605.12477*, 2026.

Chingkwun Lam, Jiaxin Li, Lingfei Zhang, and Kuo Zhao. Governing evolving memory in LLM agents: Risks, mechanisms, and the stability and safety governed memory (SSGM) framework. *arXiv preprint arXiv:2603.11768*, 2026.

LangChain, Inc. LangMem: Long-term memory for LangChain agents. https://github.com/langchain-ai/langmem, 2025.

Shuyue Stella Li, Bhargavi Paranjape, Kerem Oktar, Zhongyao Ma, Gelin Zhou, Lin Guan, Na Zhang, Sem Park, Lin Chen, Diyi Yang, Yulia Tsvetkov, and Asli Celikyilmaz. HorizonBench: Long-horizon personalization with evolving preferences. *arXiv preprint arXiv:2604.17283*, 2026.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. arXiv:2402.17753.

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

Bo Peng, Zhiheng Wang, Heyang Gong, and Chaochao Lu. IP-Dialog: Evaluating implicit personalization in dialogue systems with synthetic data. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 17007–17040, 2025. doi: 10.18653/v1/2025.findings-emnlp.923. URL https://aclanthology.org/2025.findings-emnlp.923/. arXiv:2506.02449.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. *arXiv preprint arXiv:2601.07468*, 2026.

Haofei Sun and Lin He. When memory updates but behavior does not: Repairing implicit stale dependencies in personalized agent responses. *arXiv preprint arXiv:2608.01619*, 2026. First posted 3 August 2026.

Zhaoxuan Tan, Zheng Li, Tianyi Liu, Haodong Wang, Hyokun Yun, Ming Zeng, Pei Chen, Zhihan Zhang, Yifan Gao, Ruijie Wang, Priyanka Nigam, Bing Yin, and Meng Jiang. Aligning large language models with implicit preferences from user-generated content. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7792–7820, 2025. doi: 10.18653/v1/2025.acl-long.384. URL https://aclanthology.org/2025.acl-long.384/.

Bryan Wilie, Samuel Cahyawijaya, Etsuko Ishii, Junxian He, and Pascale Fung. Belief revision: The adaptability of large language models reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 10480–10496, 2024. doi: 10.18653/v1/2024.emnlp-main.586. URL https://aclanthology.org/2024.emnlp-main.586/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10813.

Derong Xu, Shuochen Liu, Pengfei Luo, Pengyue Jia, Yingyi Zhang, Yi Wen, Yimin Deng, Wenlin Zhang, Enhong Chen, Xiangyu Zhao, and Tong Xu. Learning how and what to memorize: Cognition-inspired two-stage optimization for evolving memory. *arXiv preprint arXiv:2605.00702*, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2502.12110.

A LIMITATIONS

The implicit-drift benchmark is synthetic, English-only, single-user, and closed-valued. Its deterministic truth enables paired interventions but omits open-set ambiguity, cross-attribute causal propagation, shared-device behavior, and explicit user correction. Behavioral evidence may also reflect temporary constraints, exploration, or noise rather than a durable preference; deployment therefore requires provenance, calibrated uncertainty, user inspection, deletion, and a defer/confirm path.

Our read-time estimates are conditional on generated histories that pass prespecified materialization and co-visibility checks. The primary contrast uses 192 of 576 candidate contexts, and the native LangMem contrast uses the 24 of 120 histories in which the manager forms and retrieves a correct current/previous relation. Lifecycle roles are supplied or made semantically transparent rather than inferred automatically. Estimates also depend on a deterministic action parser, and their intervals quantify variation over sampled histories rather than models or deployment users. The smaller representation panel shares this eligibility-gated population, so its timestamp null result is specific to this protocol. Qwen does not establish the prespecified five-point stable noninferiority bound.

Native-system coverage is limited. Capture uses one Qwen host, proxy rules for text state, and only Mem0 and LangMem in the larger write–action audit; Letta could not be run reproducibly in our replay setup. Neither audited manager forms a grounded lifecycle relation on the 400 drift items, so the supplied-role intervention isolates downstream use rather than autonomous assignment. We also enlarged the native-audit sample from 48 to 240 histories after finding an error in our bootstrap code and recomputed all intervals from raw outputs. Because this expansion was not planned, we treat it as a precision update, not an independent replication.

External transfer is evaluated through structured action transformations rather than official end-to-end leaderboard protocols. The suite uses curated or mechanically grounded subsets of LongMemEval ($50/78$) and MemoryAgentBench ($74/100$), with reference-supplied lifecycle roles; effect magnitudes therefore need not match official benchmark results. HorizonBench data were inaccessible, so we do not report an independent long-horizon score.

B BEHAVIORAL TEMPLATES AND PROBE EXTRACTORS

The released artifact contains two relevant components: the complete template library and the deterministic probe-side extractors used by the evaluation code. We summarize the design here to make the benchmark auditable from the paper.

Template constraints. For each attribute-value pair, templates must satisfy three checks: (i) they express the value behaviorally without an explicit statement of preference change; (ii) they contain at least one feature recoverable by a deterministic structural recognizer; and (iii) they do not contain lexical cues for another value in the same attribute's value space. The generator samples templates under these constraints with a fixed seed.

Probe extraction. Probe answers are mapped to values by a deterministic vocabulary matcher over $\mathcal{V}_a$. If multiple values are mentioned, the extractor uses the latest occurrence in the response, reflecting the common pattern "not X, use Y." If no value from $\mathcal{V}_a$ appears, the response is counted as no-answer. This rule is intentionally simple and is applied identically to every system.

Behavioral recognizer scope. The closed-schema recoverability control uses only structural and lexical signals available in the transcript: code-block language tags, identifier casing, fenced output formats, command-line vocabulary, GUI vocabulary, defensive-programming keywords, message length, and order of code versus prose. No LLM is used inside the benchmark generator, primary evaluator, or deterministic action evaluator.

C RUNTIME AND ACTION-EVALUATION DETAILS

Host serving. Qwen-2.5-72B-Instruct and Meta-Llama-3.1-70B-Instruct are served with vLLM, tensor parallelism 4, bf16 weights, eager execution, and temperature zero. The original native captures

and value interventions use Qwen. The action validation reuses those exact rendered contexts and calls both hosts through the same OpenAI-compatible interface. The gpt-oss-120b declaration runs use the official MXFP4 checkpoint with vLLM 0.17.1 on one 80 GB GPU and a $4{,}096$-token limit.

Native adapters. Mem0, A-MEM, LangMem, and retrieval baselines use repository snapshots pinned in the released environment manifests. Simple RAG uses `bge-m3` with top-$k = 10$. Captures save adapter inputs, store inventory, retrieved pieces, exact host messages, response, and parsed value. No memory system is rerun for the action validation.

Action task construction. Action tasks reuse complete native-audit captures whose context, scenario truth, and stale mask pass provenance checks. We select 30 scenarios for each of four action attributes within each architecture and instantiate all four factorial conditions; the task template is assigned deterministically by hashing the scenario identifier. The resulting $1{,}440$ prompts were fixed before any response was collected; their exact messages, answer truth, and parser version are released.

Deterministic scoring. Programming-language tasks require a syntactically characteristic function or command in the target language; naming tasks inspect identifier case; format tasks parse a single JSON, YAML, CSV, Markdown, or XML artifact; tool tasks match the required CLI or GUI tool and reject conflicting alternatives. Required semantic fields are checked independently of surface form. The primary parser was revised only after a balanced 48-response pilot to admit mechanical equivalents (for example, `typescript` as JavaScript and a single fenced YAML document), and was then fixed before the main runs. A post-hoc sensitivity parser additionally accepts common naming-pattern synonyms, case-insensitive scalar keys, Lens GUI wording, and a generic Rust signature; it is used only for sensitivity analysis.

Statistics and integrity checks. All accuracy contrasts are paired within scenario. Bootstrap replicates sample profiles with replacement and retain every system, action, and condition for a sampled profile. We use $10{,}000$ replicates and percentile 95% intervals. The analysis checks that every cell is present exactly once and that each response matches its prompt and ground truth.

D Lifecycle study: cohort construction and supplementary results

D.1 Primary cohort construction

A 26-context pilot motivated the main study. The main cohort size, selection rule, primary contrasts, and scoring procedure were fixed before any main-cohort response was collected, and pilot responses are not pooled with the primary estimates.

A fixed, released seed generates 576 drift candidates (48 per four attributes $\times$ three patterns), plus 128 stable histories (16 per four attributes $\times$ clean/noisy behavior). Each trajectory has a unique synthetic profile ID. We verified that these histories do not overlap with those used in the other experiments. Shared behavioral templates remain a limit on distributional generalization.

Capture and eligibility. Every drift candidate passes through the same LangMem-based store adapter and deterministic materializer used in the pilot. The estimator sees only session turns and timestamps; benchmark truth is used subsequently for eligibility, not to repair its output. An eligible capture has (i) a correct final estimate, (ii) one materialized record with the correct value in the actual read context, and (iii) exactly one other record matching a consolidated pre-drift session. Of the 576 candidates, 494 satisfy these criteria; a prespecified within-cell rank selects the first 16 distinct old/current contexts per cell, producing 192 pairs. Selection never uses a host action outcome.

Each primary drift and stable item has one generated history per synthetic profile. The current-only diagnostic contains 168 distinct prompts because some current records coincide after the old record is removed; it does not increase the effective sample size of the primary comparison. Independence therefore refers to generated histories rather than deployment users.

Table 4: All drift conditions in the main lifecycle cohort, $n = 192$ per host. Each cell lists current/stale/other/invalid counts. The merged contrast changes both field semantics and serialization.

Condition	Qwen	Llama
Native records, no status	52/120/1/19	46/102/0/44
Neutral (both `UNSPECIFIED`)	47/123/2/20	46/104/0/42
Current `ACTIVE` only	65/109/1/17	71/81/0/40
Old `SUPERSEDED` only	59/108/2/23	77/75/0/40
Both labels (primary)	102/75/0/15	106/42/0/44
Merged, neutral	27/139/2/24	15/114/0/63
Merged, lifecycle	108/70/1/13	135/18/0/39
Current record only	175/0/0/17	136/3/0/53

Table 5: Stable controls on 128 histories per host. Noninferiority (NI) uses a prespecified -5pp margin and conservative one-sided bounds adjusted for two hosts.

Host	Neutral	Lifecycle	Δ current	NI conclusion
Qwen-72B	112/128	109/128	-2.3pp	Not established
Llama-70B	97/128	98/128	$+0.8$pp	Established

D.2 Interventions and stable controls

Eight drift conditions are run: native records without a status wrapper; neutral (both `UNSPECIFIED`); current labeled `ACTIVE` only; old labeled `SUPERSEDED` only; both labels (primary); merged neutral; merged lifecycle; and current only. The four factorial status cells change only the status field; old/current source strings, record order, host instruction, and task remain fixed. Merged roles also change field names, nesting, and serialization, so their contrast is reported separately.

The stable arm retains the session-0 transcript as one record and sessions 1–11 as the other. Both neutral and lifecycle conditions receive identical text, with `UNSPECIFIED`/`UNSPECIFIED` versus `HISTORICAL_VALID`/`ACTIVE` labels. Stable truth remains the same for every session, including the noisy controls. All 128 histories are kept without estimator- or outcome-based filtering. This design was fixed before the main runs.

Statistics. Each history has its own profile. Paired profile bootstrap uses 10,000 replicates with a fixed seed. The primary family contains the two host-specific current-action effects for lifecycle minus neutral, with exact McNemar tests and Holm correction. Adjusted p-values are approximately 5.55×10^{-17} and 5.46×10^{-17}. Attribute, pattern, component, merged, and naming/programming contrasts are secondary; the naming/programming stratum was prespecified.

Stable noninferiority. Let p_+ and p_- denote favorable and adverse current-action flip probabilities; the paired accuracy effect is $p_+ - p_-$. We use a conservative one-sided 97.5% lower bound per host, formed from a lower exact Clopper–Pearson bound on p_+ and an upper exact bound on p_-, each with one-sided tail 0.0125. The union bound gives tail at most 0.025 per host; Bonferroni across two hosts controls the family at 0.05. Noninferiority requires the resulting difference bound to exceed the prespecified -0.05 margin. This exact bound remains informative when all observed paired differences are zero, unlike a degenerate bootstrap interval. Qwen's bound is -7.4 points (not established); Llama's is -3.4 (established). A nonsignificant Qwen decline ($p = .25$) does not reverse that decision.

D.3 Heterogeneity and failure interpretation

The main cohort shows a large naming/programming effect. Output-format transfer is supported on both hosts; tool-preference magnitude is weaker and differs by host. Current-action factorial interactions are $+13.0$ points [6.3,19.8] for Qwen and $+2.1$ [-5.2,9.4] for Llama. The component interaction therefore does not generalize uniformly across hosts, although the primary effect is clear.

Table 6: Primary current-action effect by prespecified attribute and by the naming/programming stratum. Per-attribute cells are descriptive, without separate multiplicity claims.

Stratum	n	Qwen: pp [95% CI]	Llama: pp [95% CI]
Naming/programming	96	$+43.8$ [34.4, 54.2]	$+45.8$ [35.4, 56.3]
Naming	48	$+58.3$ [43.8, 70.8]	$+62.5$ [47.9, 75.0]
Programming	48	$+29.2$ [16.7, 41.7]	$+29.2$ [16.7, 43.8]
Output format	48	$+25.0$ [12.5, 37.5]	$+22.9$ [10.4, 37.5]
Tool preference	48	$+2.1$ [0.0, 6.3]	$+10.4$ [2.1, 18.8]

Table 7: Representation robustness on 96 drift contexts per host. Pooled columns combine the two hosts; per-host columns give the paired current-action change relative to each host's neutral rendering (neutral current accuracy: Qwen 35.4%, Llama 63.5%).

	Pooled			Δ current by host	
Representation	Current	Stale	Δ current	Qwen	Llama
Neutral	49.5%	32.8%	–	–	–
Literal status	58.3%	22.4%	+8.9	+14.6	+3.1
Timestamp only	49.0%	33.3%	-0.5	-4.2	+3.1
Supersedes relation	52.6%	28.6%	+3.1	+1.0	+5.2
Natural validity/provenance	61.5%	18.8%	+12.0	+19.8	+4.2
Host-level recency rule	70.8%	9.4%	+21.4	+42.7	0.0

The three adverse Qwen stable flips comprise two task-constraint failures and one parser-coverage failure; none switches to a historical value. Primary scoring and the noninferiority conclusion are unchanged.

E Representation robustness ablation

This appendix tests whether the read-time effect depends on the literal `ACTIVE`/`SUPERSEDED` strings. We fixed 96 drift and 32 history-preserving stable contexts from the correct-and-visible population before collecting any responses. The six renderings use the paired controls specified in Section 5, including a fixed deterministic parser.

The six renderings are neutral records, literal status fields, timestamp-only metadata, an explicit supersedes relation, ordinary validity/provenance fields, and a host-level recency rule stated in the system instruction. The last condition is exploratory because it changes host guidance rather than only record representation.

The natural-field gain is +19.8 points for Qwen and +4.2 for Llama (pooled paired $p = 5.65 \times 10^{-6}$). Qwen drives the pooled robustness result; Llama's neutral accuracy in this panel is already 63.5%, and no schema changes it by more than 5.2 points, including the literal labels (+3.1). Timestamp-only metadata has no detectable pooled effect ($p = 1$), while the supersedes relation is positive but smaller and descriptive ($p = .0703$). The host-level recency rule has a large but heterogeneous effect (+42.7 for Qwen and 0.0 for Llama), so it is not used for the primary lifecycle conclusion. Stable controls show no adverse pooled flips for the natural-field condition; their small +1.6-point change is a preservation check rather than a new noninferiority claim.

The panel weakens a literal-token explanation in the pooled data and for Qwen: ordinary validity and provenance fields retain the downstream effect, while a timestamp without a validity or supersession relation does not.

F Declaration studies

F.1 Judgments over a closed candidate set

The 240-history declaration panel uses 50 source profiles and one scored probe per scenario. Each model receives the same scenario identities and paired conditions. The neutral replacement and the retained declaration occupy the same construction slots; the subsequent behavioral trace is unchanged. The strict-choice protocol fixes candidate order within a scenario and counterbalances old/current positions to within one per attribute. The inference prompt presents candidates without old/current role labels. Invalid and runtime-error outcomes remain incorrect in the denominator.

The main-text estimates come from constrained-choice runs on three hosts; the full 240-history estimates and control contrasts are reported below.

Table 8: Declaration effects. Judgment columns use 240 histories per model; the final column uses a separate 120-history action study. Accuracy is a percentage; penalties and increases are percentage points with study-specific profile-cluster 95% intervals.

Host	Neutral judgment	Old declaration	Judgment penalty	Stale-action increase
Qwen-72B	55.0	18.3	+36.7 [29.8, 43.4]	+20.8 [13.3, 28.5]
Llama-70B	57.9	19.2	+38.8 [31.9, 45.5]	+20.0 [12.8, 26.7]
gpt-oss-120b	60.4	27.5	+32.9 [26.2, 39.8]	+25.0 [16.9, 32.8]

Table 9: Additional declaration controls on the 240-history panel. Values are current-state accuracy differences in percentage points. User-only and assistant-neutral columns measure the declaration penalty; explicit revision is the benefit of declaring the new value.

Host	User-only penalty	Assistant-neutral penalty	Revision benefit
Qwen-72B	+46.2	+45.8	+71.7
Llama-70B	+34.2	+38.3	+69.2
gpt-oss-120b	+35.8	+34.2	+50.4

The four declaration hypotheses use study-specific profile-cluster inference and Holm correction. User-only and assistant-neutral comparisons are controls for alternative sources of declaration influence, while declaring the new state is a positive-control condition.

F.2 Separate action study

The separate action analysis uses 120 histories from 46 source profiles and algorithmic current/old labels. It compares neutral, exact-old-declaration, and candidate-free-paraphrase conditions under a prespecified six-contrast Holm analysis; invalid outputs and explicit deferrals remain in the denominator.

The outcomes measure conformity to the scenario generator's state. The action cohort is reported separately from the 240-history forced-choice study because the outcome space, task constraints, and profile composition differ.

G Supporting Analyses

This appendix reports the native-path decomposition and the availability $\times$ stale-suppression intervention. Protocol and data provenance details for the external suite and native audit are provided in the released artifact.

The external action suite and native STALE audit use transformed public-data tasks rather than official leaderboard evaluations; full provenance and parser records are released with the artifact.

Table 10: Native-system localization on 240 scenarios per system (Qwen host). Ordered pipeline labels are observational diagnostics; causal claims use the paired factorial interventions. Brackets are profile-cluster bootstrap 95% intervals.

System	Native acc.	Retention	Retrieval	Conflict exposed	Error | exposed	Unresolved
full context	38.8	0.0	0.0	128/240; 53.3 [47.7, 58.8]	54/128; 42.2 [31.7, 53.0]	38.8
context window	27.1	0.0	72.9	0/240; 0.0 [0.0, 0.0]	–	0.0
simple RAG	24.6	0.0	1.7	68/240; 28.3 [23.0, 33.6]	49/68; 72.1 [60.3, 82.1]	53.3
Mem0	19.6	37.9	13.8	80/240; 33.3 [28.2, 38.7]	55/80; 68.8 [57.6, 79.8]	8.3
A-MEM	21.7	0.0	1.7	64/240; 26.7 [21.0, 32.1]	52/64; 81.2 [70.2, 90.5]	55.0
LangMem	1.7	0.0	81.2	0/240; 0.0 [0.0, 0.0]	–	17.1

Table 11: Availability $\times$ stale-suppression factorial on captured native contexts, both hosts. P: native context; $+C$: a correct current-state record is appended; $-S$: old-value spans are masked. Accuracy columns are percentages; $I_{A\times S}$ is the interaction, with profile-cluster bootstrap 95% intervals.

Host	Context source	P	$P{+}C$	$P{-}S$	$P{+}C{-}S$	$I_{A\times S}$
Qwen-2.5-72B	Mem0	9.2	52.5	21.7	67.5	+2.5 [-5.8, 10.4]
Qwen-2.5-72B	A-MEM	15.0	52.5	21.7	70.0	+10.8 [1.7, 19.7]
Qwen-2.5-72B	LangMem	0.0	15.0	2.5	64.2	+46.7 [37.4, 55.7]
Qwen-2.5-72B	pooled	8.1	40.0	15.3	67.2	+20.0 [14.8, 25.2]
Llama-3.1-70B	Mem0	12.5	65.0	24.2	73.3	-3.3 [-11.5, 4.1]
Llama-3.1-70B	A-MEM	18.3	67.5	26.7	75.8	-0.0 [-8.2, 8.8]
Llama-3.1-70B	LangMem	0.8	27.5	3.3	60.0	+30.0 [21.1, 38.9]
Llama-3.1-70B	pooled	10.6	53.3	18.1	69.7	+8.9 [3.9, 13.9]