MEMREQ: CONTRACT-CONDITIONAL REQUIREMENT PROFILES FOR DIAGNOSING AGENT MEMORY FAILURES

Anonymous authors
Paper under double-blind review

ABSTRACT

End-to-end memory scores do not identify whether a named operation was exercised, its prerequisite state was retained, or evidence reached the Reader. We introduce MemReq, an offline diagnostic protocol that freezes the workload, information, Reader, resource, and scoring contract; contrasts executed operations; probes prerequisites; audits frozen-state evidence; and tests a bounded, gold-free intervention. Its requirement profiles are empirical and contract-conditional, not workload-intrinsic necessities. On 58 shared multi-hop questions with byte-identical contexts and matched 16,384-token completion caps, Qwen35 and GPT-OSS-120B yield V$\times$C interactions of +27.59 and -5.17 percentage points, respectively. Reader identity and reasoning settings both differ, so this is not a model-only causal effect. In a crossed Mem0-RV diagnostic, 62/64 instances with a valid update opportunity reach latest-only state: apparent version failure can originate before replacement is possible. On frozen stores, version filtering improves Additive and A-MEM under dense and graph readout but leaves RV near zero; RV's tested graph contexts do not change under filtering. The graph intervention does not reproduce its 6k gain on the official 32k-history configuration, where query anchoring limits delivery. MemReq therefore localizes observable failures without inferring unexercised primitives or universal architecture rankings.

1 INTRODUCTION

Persistent memory is now a standard component of language-model agents. Systems store episodes, retrieve demonstrations, update user facts, organize knowledge into graphs, and distill reusable procedures (Packer et al., 2023; Park et al., 2023; Chhikara et al., 2025; Xu et al., 2025; Ouyang et al., 2025). Yet an end-to-end score collapses several questions: Was the needed fact ever written? Was the current version selected? Were retrieved fragments composed? Could the reader use a clean context? A source-level memory primitive does not guarantee that the deployed system reliably realizes the corresponding capability.

StateMem studies state-first intervention with the original transcript, backend chunks, and query (Fan et al., 2026). MemReq instead audits a frozen persistent store, controlled contexts, and evaluator-supported evidence. Their information boundaries make them complementary, not direct replications.

We organize the diagnosis around three questions. *RQ1:* Which executed memory operations change results under a fixed workload, Reader, information, resource, and scoring contract? *RQ2:* Did a deployed system first create the prerequisite state and opportunity to exercise the operation? *RQ3:* Once the store is frozen, where does candidate evidence cease to be observable, and does a bounded gold-free intervention change the Reader's outcome? Figure 1 shows the protocol, not a universal repair algorithm.

Our contributions correspond to the three questions:

1. **Contract-conditional operation profiles.** Under a frozen Reader and information contract, we separate observed response, point-estimate sufficiency, and paired non-inferiority over executed cells rather than asserting workload-intrinsic necessity.

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.

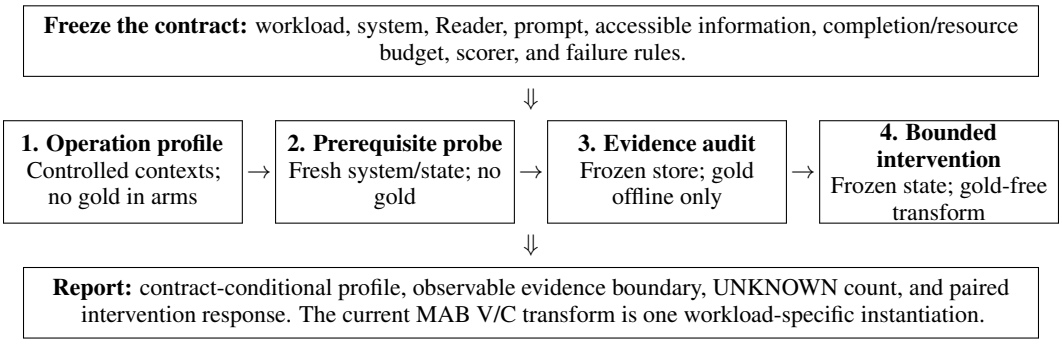

Figure 1: **MemReq workflow.** Evaluator annotations may support the offline audit but never enter system execution or the bounded transform. Fresh-state probes and frozen-state readouts answer different questions.

2. **Prerequisite-aware realization and evidence localization.** Fresh-state probes distinguish a missing update opportunity from failure after one exists; same-state traces align retained state, query readout, and answer while retaining UNKNOWN.
3. **Bounded same-state intervention.** A gold-free transform tests whether altered frozen state or readout changes outcomes. The current V/C implementation is MAB-specific, and does not reproduce the 6k gain in the 32k configuration.

2 Related Work

Long-horizon agents use external stores, retrieval, reflection, and reusable experience (Park et al., 2023; Packer et al., 2023; Shinn et al., 2023; Ouyang et al., 2025). Mem0 exposes extraction and update operations, while A-MEM builds linked notes (Chhikara et al., 2025; Xu et al., 2025). MemReq does not propose another architecture: it asks whether tested operations change outcomes under a fixed contract and whether the deployed store formed the prerequisite state.

RAGChecker decomposes retrieval and generation given a corpus/context pipeline; RGB and HippoRAG study complementary retrieval robustness and graph-based retrieval (Ru et al., 2024; Chen et al., 2024; Gutiérrez et al., 2024). We do not claim the first retriever/Reader decomposition. MemReq additionally audits persistent-state creation and retention before query-time retrieval and tests a bounded intervention on the same frozen store.

LoCoMo and LongMemEval evaluate conversational memory, while MemoryAgentBench and AMA-Bench provide agentic tasks (Maharana et al., 2024; Wu et al., 2024; Hu et al., 2025; Zhao et al., 2026). StateMem tests explicit state tracking with broader transcript and backend access; MemReq instead measures contract-conditional operation response and observable evidence boundaries under restricted information (Fan et al., 2026). MEMPROBE reconstructs hidden state from artifacts, whereas our probes test operational realization; MemDelta examines implementation-dependent rankings, whereas we align operation response, prerequisites, and same-state delivery (Ma et al., 2026; Wang, 2026). StructMemEval motivates structure-sensitive workloads, while CrossEp/AMA provide heterogeneity checks here (Shutova et al., 2026; Wang et al., 2026). Context placement and evaluator choice can also vary (Liu et al., 2024; Zheng et al., 2023); we report paired and blinded sensitivities.

3 MemReq: A Contract-Conditional Diagnostic Protocol

3.1 Scope and empirical sufficiency

In MemReq, a *requirement* is an empirical relation between an executed operation set and a measured reference under a fixed workload, Reader, prompt, accessible-information, resource, and scoring contract. Changing that contract can change the profile. A requirement is neither mathematical

Table 1: Information contract. "Template" is public workload knowledge; "order" is the adapter-persisted visible ingestion sequence, not hidden validity time. "Gold" includes evaluator current/stale and supporting-path annotations.

Stage	Store	Query	Template	Order	Raw history	Gold
Controls / system run	yes	yes	public	visible	adapter-defined	no
Isolated probe	local	local	frozen	visible	no	no
Offline stage audit	frozen	yes	public	recorded	audit-only	audit-only
Store-only transform	frozen	readout	public	recorded	no readback	no
Scoring	output	yes	no	no	no	score-only

necessity nor a universal minimal capability set: it is relative to executed cells, the empirical reference, and a stated tolerance. The Reader contract is part of this definition, not a nuisance variable.

The generic MemReq procedure freezes the evaluation and information contract, constructs controlled operation contrasts, tests whether prerequisites were realized, aligns frozen store–query–context–answer evidence, applies a bounded intervention without forbidden information, and reports observable stages, UNKNOWNs, and response. It is offline and annotation-supported, not a label-free online root-cause detector. Executable realizations of retrieval, version selection, composition, and abstraction may differ by workload. In particular, our Strict-V/C parser, slot resolver, and graph traversal use public MAB surface grammar and visible ingestion order; they are a *workload-specific instantiation*, not a general temporal database or a transform validated for arbitrary natural-language memories. Another workload would need its own independently validated executable transform under the same information rule.

We study four operational capabilities. Retrieval (R) exposes potentially relevant retained content. Version resolution (V) selects a current value while suppressing superseded values. Structured composition (C) assembles multiple retained facts or edges into an answer-supporting path. Cross-episode abstraction (A) persists a rule-like object derived from multiple experiences. A frozen task prior (P) is a non-memory control. These are tested instantiations, not an exhaustive ontology.

For a fixed workload, Reader, prompt, accessible information, resource contract, empirical reference, and tolerance, let $\mathcal{E}$ be the executed capability sets. We define point-estimate sufficiency and its inclusion-minimal frontier as

$$\begin{aligned} S \text{ is sufficient} \iff \widehat{Q}(S) \geq \widehat{Q}(\text{reference}) - \epsilon, \\ \mathcal{M}_\epsilon = \text{Min}_{\subseteq}\{S \in \mathcal{E} : S \text{ is sufficient}\}. \end{aligned} \tag{1}$$

Here $\text{Min}_{\subseteq}$ selects inclusion-minimal members. We report $\mathcal{M}_\epsilon$ only as a descriptive frontier over executed cells, not an established global or statistical minimum; we retain *unresolved* otherwise. The empirical reference is a measured comparator, the point-estimate gap is its observed score minus the candidate score, and paired non-inferiority separately asks whether the confidence bound excludes a deficit larger than the specified margin. For CrossEp we also report headroom capture relative to an explicit NoMemory baseline; its formula and denominator caveat appear in Supplementary Appendix C. The MAB-MH all-version Base condition is not an empty-set or NoMemory arm. AMA has no reference arm and therefore receives no reference-based sufficiency claim.

Algorithm 1: Generic MemReq diagnostic procedure. Input: workload/history; executed arms, reference, tolerance, system adapter, probes, support labels, and frozen budget. **Procedure:** (1) freeze permissions, cohorts, prompts, scorer, and failure rules; (2) with an empirical reference, compute executed-set point gaps and paired non-inferiority separately; without one, use factorial or response contrasts; (3) run fresh-state probes; (4) align one frozen state, query, store, context, and answer while retaining UNKNOWN; (5) record the human intervention hypothesis before outcomes; (6) apply the frozen gold-free transform; (7) pair outcomes and locate the earliest observable failure stage. **Output:** contract-conditional profile, probe fingerprint, stage evidence, UNKNOWN count, and intervention response. The earliest stage need not be the unique root cause.

3.2 CONTROLLED CONTEXT CONSTRUCTION

The MAB factorial changes context construction while holding query, actor, Reader prompt, temperature, seed, scorer, and parser fixed. MH Base exposes all versions; V keeps the latest value per surface slot; C builds a query-anchored all-version graph; and V+C composes over the latest-state graph. These arms do not use BGE top-k, so evidence counts and lengths may differ. The older SH empty-context contrast changes availability and temporal selection. The new SH control instead serializes the same complete history either as AllVersion or after one frozen, query-independent LatestGlobal filter. Equal-budget Latest-versus-Oldest separately tests direction.

3.3 ISOLATED PROBES AND ANNOTATION-SUPPORTED LOCALIZATION

Each behavioral probe uses a fresh system instance and records its snapshot and retrieval. R tests a unique marker among distractors; V requires the current marker and suppresses the stale one; C seeds top-$k = 2$ retrieval and requires all three frozen path markers in the final context; A requires one persistent object with frozen trigger, action, and nonverbatim rule markers. The A probe is marker-dependent: a semantically correct rule without those markers scores negative, so it tests artifact realization under this contract rather than unrestricted semantic abstraction. A uses a 2048-token memory-LLM cap, temperature zero, and a 300-second timeout. No probe uses an answer Judge, and A detects a rule-like artifact without establishing rule correctness or generalization.

Probe build identifiers denote repeated isolated runs across build/runtime cohorts. The probes are not executed inside the corresponding 455-fact benchmark stores, and a shared build identifier does not imply shared persistent state. Probe-to-accuracy association is therefore exploratory, not a state-level predictive-validity test.

The stage audit compares native store and retrieval traces with evaluator/gold current, stale, and supporting-path annotations. These annotations are used only for offline diagnosis and scoring; they never enter a strict transformation. Labels are operational observations: for example, WRITE_MISS means that the audited final store contains neither current nor stale evidence, not that the current fact was proved never to have been written. Multi-hop path status is multi-field and may be UNKNOWN. MemReq therefore reports annotated failure categories rather than claiming an automatic causal root cause from the first trace anomaly.

3.4 MAB-SPECIFIC EXECUTABLE INTERVENTION

The strict intervention instantiates the generic protocol on public MAB surface grammar. Here store-only is an information-source boundary: the transform may inspect the frozen store and adapter-persisted visible ingestion order, but may not reread the raw history or evaluator state. It does not imply that Original and repaired arms use the same readout. The intervention normalizes native memory text into proposition lines, parses subject/predicate/object triples, and defines a slot by normalized subject plus surface predicate. Ordering uses retained ingestion metadata (A-MEM `serial-*` timestamps or Mem0 `source_serial`) and otherwise falls back to snapshot insertion order; this is not real-world validity time. For each parsed slot, Strict-V retains the maximum $(\mathrm{order}, \mathrm{atom_id})$ and preserves every unparsed atom. The preregistered tie policy and implemented atom-ID tie-break differ. No tie occurs in the 28 valid extended or nine primary snapshots; the separate 32k transfer has one different-value tie, but its alternative changes 0/22 VC context hashes (Supplementary Appendix N).

Strict-C builds directed subject→object edges from parsed Strict-V atoms and preserves source-memory, atom, and text-span provenance. Query-time longest-subject matching seeds a stable BFS with depth below six and at most 32 edges. Strict-V readout uses BGE-M3 top-10, whereas Strict-V+C uses graph traversal; their difference is not a pure reader-reasoning effect. Global V filtering and graph construction are query-independent, but readout is query-conditioned. Across audited panels, 7,457 retained atom-view objects have source provenance; the 28 extended snapshots separately contain 7,213 parsed slots, not graph edges. Neither count proves fact recall or semantic correctness.

4 EXPERIMENTAL SETUP

Workloads and actors. MAB single-hop (SH) and multi-hop (MH) conflict tasks provide interpretable context-construction testbeds (Hu et al., 2025). The 6k MH factorial contains 58 items and three valid actors: Qwen3.5-9B, Qwen3.5-35B-A3B, and Gemma-4-12B-it (two model families). Its label "6k" is an official history-size configuration, not the Reader's input-token count. The 6k harness uses temperature zero and a fixed seed; Qwen35 and Gemma use a 256-token completion cap, while Qwen9 has the frozen 256 run plus a length-rule 512 retry. The official-history 32k confirmation uses a separate frozen harness, a fixed seed, and a 512-token completion cap. It shares the capability definitions and graph-operation family with 6k but is not byte-identical implementation evidence. The SH same-history family uses the frozen 31-item 32k and 56-item 6k panels, Qwen9/Qwen35, temperature zero, a fixed seed, and 512 completion tokens. Exact seed values and protocol lineage are reported in Supplementary Appendix N. Tokenizer checks place the longest prompt at 39,595 tokens, below the verified 99,999-token service limit; no history is cut.

CrossEp-Tool transfers prior tool-use trajectories over 64 targets from EvoMemBench (Wang et al., 2026); CrossEp-Emb contains 121 ALFWorld targets across six task families (Shridhar et al., 2021). AMA-Bench contributes 24 episodes, 288 queries, and six arms (Zhao et al., 2026). Real-system experiments compare pinned Mem0-RV[1], Mem0-Additive, and A-MEM. Their native writer/storage contracts differ; we do not rank these architectures as globally better or worse. In the extended panel, SH and MH are query lanes over one byte-identical ordered 455-event stream: each build creates one store queried by 56 SH and 58 MH items. Thus 28 valid snapshots (8 RV, 10 additive, 10 A-MEM) form 56 lane panels and $28(56 + 58) = 3{,}192$ observations, not 56 stores or independent histories. The primary panel contains nine of these snapshots (three systems by three prespecified builds), 58 shared MH queries, and Qwen9/Qwen35 Readers under one 512-token *completion* contract. Input evidence lengths and readout mechanisms are not generally matched unless an arm explicitly states a context budget. Supplementary Appendix L records the Reader, native writer, CrossEp injection, and AMA Judge prompt contracts.

Evaluation and human audit. Official exact/task scorers are used where available. AMA uses a frozen binary Judge and an episode-clustered analysis. In a blinded, deliberately stratified 100-response audit, two annotators agreed on 95% ($\kappa = 0.900$), with five third-party adjudications; Primary Judge agreement with human labels is 84% ($\kappa = 0.680$). This is an unweighted audit-sample result, not population-wide accuracy. The post-hoc raking sensitivity is incomplete because 14/77 occupied population cells are absent from the sample (Supplementary Appendix M); evaluator validity remains *qualified*.

For the MAB response panels, the frozen official normalized-substring scorer is primary: it normalizes the recorded answer text and accepts it when the normalized reference is a substring. It scores the answer field supplied by the runner; reasoning is not separately concatenated. A post-hoc normalized-equality sensitivity is a conservative text check, not independently calibrated semantic ground truth. It lowers correct counts from 535 to 484 (Qwen9) and 530 to 521 (Qwen35) across 1,566 primary outputs per Reader. It preserves large Additive/A-MEM MH VC-minus-Original directions and unresolved RV. In the extended SH panel, scorer choice changes Additive and A-MEM intervention magnitudes substantially; SH system-intervention magnitude is therefore not our core stage-localization evidence (Supplementary Tables 22–23 and 25). Five selected explicit-denial official-only cases were semantically incorrect in a blinded targeted audit; this does not estimate a population scorer-error rate.

Inference and statistical units. Statistics follow the design rather than one universal paired-bootstrap statement. Single-history factorials and SH same-history controls use item-paired resampling. For the primary Reader panel, the repair effect for system s and Reader r is

$$\widehat{\Delta}_{s,r} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{B}\sum_{b=1}^{B}\left(Y^{VC}_{s,r,b,i} - Y^{O}_{s,r,b,i}\right), \qquad N = 58,\ B = 3. \tag{2}$$

A 10,000-resample bootstrap samples query IDs after averaging each paired difference over the same three fixed builds. This interval is conditional on those builds and history. Crossed build by

[1]Our label, not a product name: pinned Mem0 commit `57f944e` with ADD/UPDATE/DELETE/NONE; the full revision is in the appendix.

query resampling is retained as a few-build sensitivity. AMA uses a 24-episode cluster bootstrap; query-level McNemar is secondary. The older extended-build analysis used nested resampling, but the archived analysis does not map every historical interval to its algorithm. We therefore report verified counts, means, and build directions without relabeling all intervals as crossed. Repeated builds of one history do not increase history coverage, and 174, 448, 560, or 580 build by query observations are not that many independent questions. We do not claim global family-wise error control over the exploratory audit; paired uncertainty is primary, and recorded Holm adjustment is confined to the frozen readout-contrast family.

5 Contract-Conditional Requirement Profiles

Controlled MAB factorial. RQ1: Operation response changes across the tested Reader contracts; it is not an intrinsic workload requirement. On the frozen 6k MAB-MH panel with Qwen3.5-9B ($n = 58$), all-version Base scores 3/58 (5.17%), C scores 11/58 (18.97%), V scores 27/58 (46.55%), and V+C scores 54/58 (93.10%). We estimate the factorial interaction as

$$\widehat{I}_{VC} = \widehat{Q}(V, C) - \widehat{Q}(V, 0) - \widehat{Q}(0, C) + \widehat{Q}(0, 0). \tag{3}$$

Qwen9 has $\widehat{I}_{VC} = +32.76$ pp [12.07,51.72]; Qwen35/256 has +27.59 pp [6.90,46.55], and Gemma +17.24 pp [-1.72,36.21]. A complete Qwen35/16k run on MH58 scores 2/13/27/54 and retains +27.59 pp [6.90,46.55] (Supplementary Appendix P). A separate complete GPT-OSS-120B single-endpoint replication under medium reasoning and a 16,384-token cap scores 43/53/47/54, with interaction -5.17 pp [-15.52,6.90]. There, C is the highest tested single-operation point estimate (53/58); adding V gains 1/58. Equality scores are 38/50/43/53, with interaction -3.45 pp [-15.52,8.62]; one empty final counts wrong. Ceiling and contract differences preclude a general design recommendation.

With byte-identical contexts and matched completion caps, the between-Reader difference between the two matched-16k risk-difference interactions is +32.76 pp [13.79,53.45] on 58 paired items. This contrast is descriptive: Reader identity and reasoning settings differ. Saturated-logit interaction odds ratios are 1.92 [0.29,12.79] for Qwen35 and 0.85 [0.37,1.96] for GPT-OSS; both within-Reader intervals include one (Supplementary Appendix P).

On frozen SH6k, AllVersion→LatestGlobal scores are 21/56→56/56 (Q35) and 19/56→55/56 (Q9); 32k Qwen counts appear in Supplementary Table 9. Oldest, random-slot, and order controls favor Qwen Latest, narrowed by reversal (Supplementary Table 11). GPT-OSS@16k scores 25/31 versus 28/31 (+9.68 pp, [-6.45,25.81]); neither a latest requirement nor 5 pp non-inferiority is established (Supplementary Table 33). On SH6k, GPT-OSS scores 56/56 in both conditions (difference 0 pp, CI [0,0]; equality 52/56 versus 53/56). In the 32k conflict subset, GPT-OSS AllVersion answers 12/18, selects stale values on four, and gives other responses on two; Qwen current selections are 1/18 each by the same matcher. Context length and stale exposure also differ. This is a Reader-contract contrast, not a model-only causal comparison.

On official-history MH ($n = 22$), Base/C/V/V+C score 2/5/5/19 of 22; the paired interaction is +50 pp [18.18, 77.27]. This separate selected history does not estimate cross-history variance (Supplementary Appendix C).

External heterogeneity checks. CrossEp-Tool favors episodic retrieval under the tested Qwen contracts, whereas CrossEp-Emb admits a non-memory task prior close to its empirical reference. Neither establishes paired non-inferiority at the 5 pp margin across the tested candidates. On AMA, retrieval improves over NoMemory for both frozen Judges, but the richer tested arms do not establish an incremental gain over retrieval alone. AMA lacks a reference arm, so retrieval is *helpful*, not declared sufficient. These checks bound the MAB interpretation rather than define the generic protocol; Supplementary Appendix C and the AMA appendix retain counts, headroom capture, and Judge-qualified uncertainty.

6 Prerequisite-Aware System Realization

RQ2: A missing update opportunity can precede an apparent version-resolution failure. Construct controls separate the four isolated probes (Supplementary Table 13). AtomicLatest realizes

Table 2: Requirement readings under executed conditions. "Sufficient" is relative to the stated reference and tolerance, not mathematical necessity.

Workload	Verified comparison	Supported reading
MAB-SH	Q35/6k: 21/56→56/56	Reader-contract dependent
	OSS/6k: 56/56→56/56	OSS ceiling
MAB-MH	Q9: 3/11/27/54; Q35/16k: 2/13/27/54	Reader-contract dependent
	OSS/16k: 43/53/47/54	
CrossEp-Tool	R > A	higher episodic-R point estimate
CrossEp-Emb	P ≈ reference	5 pp NI unresolved
AMA	R gain: +18.75/+13.19 pp	Judge-qualified

R/V, GraphLatest also realizes C, and Abstractor-Control realizes only A. The initial ten-case suites give identical $(1, 0, 0, 0)$ fingerprints for Mem0-RV and Mem0-Additive, while A-MEM scores $(1, 0, 0.7, 0)$. Mem0-RV's R score falls to 5/10 on a surface-varied v2 suite and 4/10 on its direct repeat; every failure is final-store target unavailability rather than a stored target missed by retrieval. Because v1 and v2 also differ in entities, relation wording, markers, and query wording, this establishes suite-dependent reliability, not a unique surface-form mechanism.

Identical isolated-probe fingerprints need not identify identical failure stages. Mem0-RV and Mem0-Additive share the initial profile, while store and retrieval audits locate different evidence-availability and conflict-exposure patterns. Fingerprints therefore complement, rather than substitute for, stage-level diagnosis of a workload-specific memory state.

Mem0-RV also shows why source and behavior must be separated. Its pinned source includes update/delete/no-op branches, and a minimal positive control succeeds in 36/36 trials with each Qwen write manager, whereas a distinct target suite scores 0/30. In a trace-grounded minimal success, an explicit correction replaces emergency contact "Alice Morgan" with "Bob Chen"; the final store and top-2 retrieval contain only Bob. By contrast, target case V2-01 presents old, distractor, and current markers as user-role events, yet all three extraction calls return no fact. No valid old/current update opportunity is created, so downstream no-op behavior can look like a failed version operation even though replacement was never exercised. A paired diagnostic crosses ten semantic skeletons with personal/nonpersonal wording, native-personal/generic extraction, and two writers. All 80 fresh-store instances pass the protocol-validity gate (G1). Conditional on a valid old/current opportunity, 62/64 reach latest-only state. Overall, 75/80 reach that state. The two opportunity-present failures add the current value but retain the stale old value. Of 16 cases without an opportunity, 13 succeed through a current-only write because no old target state was retained. Executor output is nonempty for 77/80 but does not by itself certify an UPDATE or DELETE. For Qwen35 on nonpersonal wording, generic extraction raises current extraction, update opportunity, and end-to-end V by 30 pp each (paired 95% CI: [0,60]). Thus upstream extraction/domain fit governs whether V is evaluable; the 0/30 is an end-to-end result, not evidence that the UPDATE primitive intrinsically fails.

The post-hoc W and write-backbone controls (Supplementary Appendix I) do not establish RV stress degradation or resolve Additive/A-MEM stale co-exposure by changing the writer alone.

7 Frozen-State Evidence and Intervention Response

RQ3: Does a same-store intervention change the candidate evidence delivered to the Reader, and does the outcome respond accordingly? These comparisons localize observable evidence states; they do not certify a unique causal failure stage or a complete multi-hop support path. Define four same-store readouts. I_D is an atomized all-version view with fixed BGE-M3 dense retrieval; V_D applies latest selection before that same dense readout. I_G parses all-version atoms and applies the same query anchor and BFS; V_G applies latest selection before that same graph readout. Their interaction is

$$J = \big[Q(V_G) - Q(I_G)\big] - \big[Q(V_D) - Q(I_D)\big]. \tag{4}$$

It measures how the observed version-filtering effect changes with readout policy, not a universal V-by-C interaction.

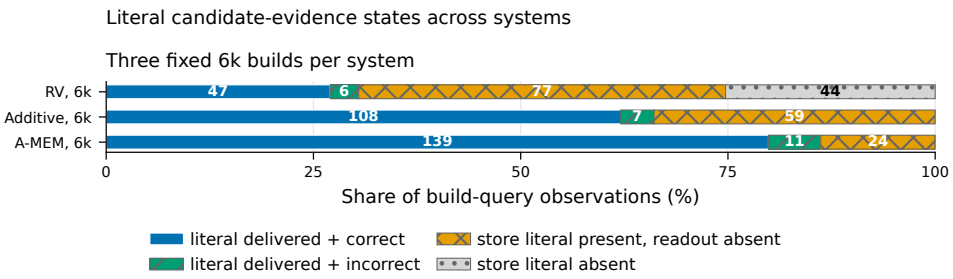

Figure 2: **Literal candidate-evidence delivery varies across systems.** Each 6k row partitions 174 build-query observations (three builds; 58 queries). Literal presence does not certify a complete semantic path; observations are not independent histories.

The GPT-OSS fixed-query intervals exclude zero for both Additive/A-MEM effects (Supplementary Table 36); no-final responses count wrong. On paired common-final cells, Additive/A-MEM remain positive and RV near zero (Supplementary Appendix P). That common-final subset is selected by response availability and is therefore secondary. Across the planned nine-store fixed-readout matrix, 277/2,088 responses have no final content; all remain in the denominator and count incorrect. No-final behavior is system/readout-dependent: RV has no final on 63/174 all-version-graph and 70/174 latest-graph cells, versus 27/174 and 28/174 for Additive and 3/174 and 8/174 for A-MEM. The logs do not isolate a reasoning mechanism, and intervals are conditional on the tested endpoint/run rather than estimates of run-to-run variance. These stores share one history and do not establish cross-history transfer. RV has only 1–2 conflicting slots per build, and $I_G = V_G$ for all 174 query contexts, versus 144/174 Additive and 165/174 A-MEM changed graph contexts. Thus the tested RV filter supplies no changed graph input. Hence its near-zero graph effect does not test a changed graph input or imply an intrinsic failure of the source UPDATE operation. Another 77 RV observations with literal candidate evidence in the store but absent from the readout expose a delivery gap under this matcher. These observations do not certify complete semantic support.

Raw retention keeps all 455 records. RV's task utility under tested V_D/V_G readouts is far below that deterministic control, while A-MEM is closer (Supplementary Table 20). Together with RV's high neither-detected rate, this suggests a larger retention/delivery bottleneck in the audited RV stores, not writer causality; storage paths and representations differ.

Annotation-supported trace diagnosis. RV has 15 WRITE_MISS among 20 SH errors in a separate 40-item audit; Additive/A-MEM often expose stale competitors despite retrieving current evidence. A blinded 68-unit audit (64/68 agreement, $\kappa = 0.882$) finds no full support in 44 RV literal-absent cells, although 41 retain partial evidence; only 12/24 literal-present controls have full support (Supplementary Table 15).

Table 3: Fixed-readout version effects and their interaction on three archived stores per system (nine stores total) and 58 shared MH queries. Values are official-score percentage points; fixed and crossed CIs appear in the appendix.

Reader	System	$V_D - I_D$	$V_G - I_G$	J
Qwen9	RV	−1.15	0.00	+1.15
Qwen9	Additive	+26.44	+50.00	+23.56
Qwen9	A-MEM	+29.31	+63.22	+33.91
Qwen35	RV	+0.57	0.00	−0.57
Qwen35	Additive	+21.26	+51.15	+29.89
Qwen35	A-MEM	+28.16	+68.39	+40.23
GPT-OSS	RV	−0.57	0.00	+0.57
GPT-OSS	Additive	+21.26	+41.95	+20.69
GPT-OSS	A-MEM	+26.44	+56.90	+30.46

Primary matched-completion-budget intervention evidence. With a shared 512-token completion cap, Qwen9 Strict-V+C minus Original is +5.75 pp for RV ([−5.75,17.24]), +52.30 pp for Additive ([39.08,65.52]), and +68.97 pp for A-MEM ([55.75,81.03]); Qwen35 gains are +5.75, +53.45, and +79.31 pp. Thus Additive/A-MEM respond strongly, while the RV effect is small and uncertain on these stores (Supplementary Appendix H). On one archived build per system, GPT-OSS-120B Original→Strict-V+C correct counts are 11→16 for RV, 8→35 for Additive, and 15→49 for A-MEM, each out of 58; paired gains are +8.62 pp [−1.72,18.97], +46.55 pp [31.03,62.07], and +58.62 pp [43.10,72.41]. Empty finals count wrong. These one-build Original/VC comparisons differ from the three-build fixed-readout result (Supplementary Appendix P).

Limited transfer to the official 32k-history setting. A fresh 2,310-fact Additive build scores 1/22 Original, 4/22 Strict-V, and 2/22 Strict-VC; 13/22 graph queries lack a seed and only 2/22 readouts retain the literal answer. Dense latest selection helps, but graph transfer does not reproduce the 6k gain: a delivery boundary, not a unique cause (Supplementary Appendix J). Prospective chronology covers only RV/Additive SH.

8 Robustness and Boundaries

Qwen Readers show positive V×C interactions under their tested contracts; Gemma's positive point estimate is uncertain. The complete single-endpoint GPT-OSS MH58 replication is our main GPT-OSS factorial result. Its earlier split-epoch estimate attenuates in the complete replication (11/58 changed arm outcomes; 8.62 pp interaction shift), providing service/run variation evidence rather than a stable negative effect. The Qwen35/16k factorial remains positive; the two-arm bridge is separate (Supplementary Appendix P). Across 48 LongMemEval-S histories, RV/Additive ordering reverses by Reader, with no consistent dense/hybrid gain (Supplementary Appendix K). These observations do not isolate Reader identity from reasoning setting or writer contract.

9 Discussion and Limitations

Profiles depend on executed cells, reference, and Reader contract. Same-state intervention evidence comes from one MAB history; repeated builds do not estimate cross-history transfer. Strict-V/C uses MAB templates and visible ingest order, not validity time or a validated general-language resolver. The official scorer is primary, yet equality sensitivity changes some SH magnitudes. Literal presence marks *candidate* evidence, not a complete semantic path; the selected human audit does not estimate prevalence.

GPT-OSS fixed-readout intervals concern one endpoint/run with material no-final behavior; selected common-final analysis is secondary. Two of nine stores lack build-local provenance, and all lack exact checkpoint/template hashes. Context hashes validate within-store pairs, not build equivalence. The failed 32k graph transfer bounds 6k intervention portability. Cross-history replication and independently validated non-template V/C resolvers remain future tests (Supplementary Appendix R).

A null intervention has bounded meaning. Without an old/current opportunity, a version probe cannot establish how replacement would behave when invoked. With byte-identical readouts, as in the tested RV graph pairs, a near-zero contrast does not test changed graph input. A positive response shows recoverable headroom for the executed transform under its Reader and readout contract, not every implementation of the named capability. MemReq therefore reports the earliest observable evidence boundary while leaving mechanisms beyond the recorded trace unresolved.

10 Conclusion

MemReq separates contract-conditional operation response from prerequisite formation, retained evidence, and query-time delivery. Apparent failure can precede execution of the named primitive. Large intervention gains occur where the frozen transform changes the evidence delivered to the Reader; a null contrast with byte-identical contexts does not test the value of changed input. These are tested-contract diagnoses, not architecture rankings. Before attributing an end-to-end score to a capability, audit its opportunity to execute and the state or context actually changed.

AI USE STATEMENT

The authors used ChatGPT and Codex for research discussion and hypothesis refinement; feedback on experimental design; coding and debugging assistance; data cleaning and analysis support; assistance in interpreting results; literature discovery and metadata checks; figure creation; and manuscript drafting, editing, formatting, and consistency checks. Model actors and judges used as experimental components are separately documented in the methods. The authors froze the scientific protocols, reviewed and tested generated code and analyses, interpreted the evidence, and retain responsibility for the final content.

ETHICS STATEMENT

The study uses public or synthetic benchmarks, model outputs, and expert annotation of benchmark responses. It involves neither deployment decisions about individuals nor collection of new personal data. Annotators reviewed blinded benchmark responses; arm, system, and Judge verdict were hidden during labeling, and five disagreements were adjudicated by a third annotator. The paper reports annotations only in aggregate. A separate 68-unit semantic-evidence audit used two blinded annotators who reviewed query readouts and the complete normalized store view; a third annotator adjudicated four disagreements. These selected strata are not a population-prevalence sample. Separately, one annotator reviewed five scorer cases in a blinded sheet that hid system, arm, build, and prior deterministic category; this targeted audit is also reported only in aggregate.

REPRODUCIBILITY STATEMENT

The paper and appendices specify the operational definitions, algorithm, information permissions, prompts and context-construction rules, model roles, completion contracts, scorer behavior, failure handling, cohort sizes, statistical units, intervention provenance, and blinded human-audit design. Aggregate counts and sensitivity analyses are reported with their denominators. Upon acceptance, we will release all experimental code and configurations required to reproduce the reported experiments, including prompts, adapters, analysis scripts, probe definitions, and figure-generation code, subject to third-party licenses. This commitment does not include raw model outputs or private annotation files; historical provenance limitations remain as reported.

REFERENCES

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (16):17754–17762, 2024. doi: 10.1609/aaai.v38i16.29728. URL https://ojs.aaai.org/index.php/AAAI/article/view/29728.

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. URL https://arxiv.org/abs/2504.19413.

Xinyi Fan, Miri Liu, Ruozhen Yang, Siru Ouyang, and Jiawei Han. Can agent memory systems track evolving state? arXiv preprint arXiv:2608.19652, 2026. URL https://arxiv.org/abs/2608.19652.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1902. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/6ddc001d07ca4f319af96a3024f6dbd1-Abstract-Conference.html.

Yuanzhe Hu, Yu Wang, and Julian McAuley. MemoryAgentBench: Evaluating memory in LLM agents via incremental multi-turn interactions. arXiv preprint arXiv:2507.05257, 2025. URL https://arxiv.org/abs/2507.05257.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. URL https://arxiv.org/abs/2307.03172.

Enze Ma, Yufan Zhou, Wei-Chieh Huang, Jie Yang, Huanhuan Ma, Zixuan Wang, Chengze Li, Chunyu Miao, Philip S. Yu, and Zhen Wang. MEMPROBE: Probing long-term agent memory via hidden user-state recovery. arXiv preprint arXiv:2606.24595, 2026. URL https://arxiv.org/abs/2606.24595.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. LoCoMo: Evaluating very long-term conversational memory of LLM agents. In *Proceedings of ACL*, 2024. URL https://arxiv.org/abs/2402.17753.

Siru Ouyang, Jun Yan, I-Hung Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long T. Le, Samira Daruki, Xiangru Tang, Vishy Tirumalashetty, George Lee, Mahsan Rofouei, Hangfei Lin, Jiawei Han, Chen-Yu Lee, and Tomas Pfister. Reasoningbank: Scaling agent self-evolving with reasoning memory. arXiv preprint arXiv:2509.25140, 2025. URL https://arxiv.org/abs/2509.25140.

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

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

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0692. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/27245589131d17368cccdfa990cbf16e-Abstract.html.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2303.11366.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Cote, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=0IOX0YcCdTn.

Alina Shutova, Alexandra Olenina, Ivan Vinogradov, and Anton Sinitsin. Evaluating memory structure in LLM agents. arXiv preprint arXiv:2602.11243, 2026. URL https://arxiv.org/abs/2602.11243.

Kuan Wang. MemDelta: Controlled baselines and hidden confounds in agent memory evaluation. arXiv preprint arXiv:2606.29914, 2026. URL https://arxiv.org/abs/2606.29914.

Yuyao Wang, Zhongjian Zhang, Mo Chi, Kaichi Yu, Yuhan Li, Miao Peng, Bing Tong, Chen Zhang, Yan Zhou, and Jia Li. EvoMemBench: Benchmarking agent memory from a self-evolving perspective. arXiv preprint arXiv:2605.18421, 2026. URL https://arxiv.org/abs/2605.18421.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. arXiv preprint arXiv:2410.10813, 2024. 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. arXiv preprint arXiv:2502.12110, 2025. URL https://arxiv.org/abs/2502.12110.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, and Jishen Zhao. AMA-Bench: Evaluating long-horizon memory for agentic applications. arXiv preprint arXiv:2602.22769, 2026. URL https://arxiv.org/abs/2602.22769.

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