# COVER: CAUSALLY ORDERED, VERIFIED EVENT RECORDS FOR SCHEDULE-ROBUST LLM MEMORY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Stateful language-model memory can interpret the same event differently depending on which other events have already arrived. Deduplicating or sorting records after extraction cannot repair this semantic change. We introduce COVER, which compiles each event from its causal ancestors, retains source-bound records, and joins them independently of delivery order. Under authenticated causal dependencies and a fixed deterministic compiler, COVER produces identical terminal states and answers across legal schedules and retries. Our analysis characterizes the local conditions for this guarantee and the information that compression must preserve for future queries. Experiments separate structural reliability from semantic utility. In a controlled action relay, COVER improves success across all tested schedules by 2.50 percentage points over typed FIFO with the same content and budget. Source-retaining readers also improve complete-history LongMemEval and LoCoMo accuracy in selected configurations. Strong canonical controls account for much of the reliability benefit, however, and restrictive typed adapters lose utility on natural dialogue. COVER provides a causal interface for reliable memory construction; its semantic gains depend on the compiler, reader, and evidence budget.

## 1 INTRODUCTION

A memory system should distinguish a change in the world from a change in the order in which its events arrive. Retries, asynchronous tools, and delayed messages make this distinction difficult for an online learned writer: its current store can contain information that the original event never observed.

Consider a profile stating "I live in Paris," a hotel request issued from that profile, and an independent update to Rome (Figure 1). Both delivery orders are valid. A writer that interprets the request using its current store may bind it to Paris in one order and Rome in the other. Sorting the resulting records preserves the mistaken binding. Exact-ID deduplication helps repeated deliveries but leaves these distinct events unchanged.

COVER places the causal boundary *before* learned interpretation. Each event is compiled from the evidence in its causal past; immutable records and retained sources are then joined into a stable view. In the example, the hotel request remains bound to Paris while the updated profile records Rome. This separation matters even when the storage layer already converges: MELD, for example, distinguishes a convergent claim-status CRDT from its order-sensitive learned merge graph (Lovén et al., 2026).

**Contributions.** COVER combines a causal evidence boundary, source-preserving validation, and order-independent materialization in one learned-memory interface. We characterize its structural guarantee through classical commutation and future-query equivalence, separating exact convergence from semantic correctness. We then evaluate both properties: controlled schedule experiments measure reliability beyond retries, while complete-history dialogue and official-code comparisons test utility across representations and readers.

Sections 3–4 describe the construction and its guarantee. Table 1 measures schedule robustness with deduplication and canonical controls; Table 2 evaluates source-retaining readers. Figure 2 connects



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.

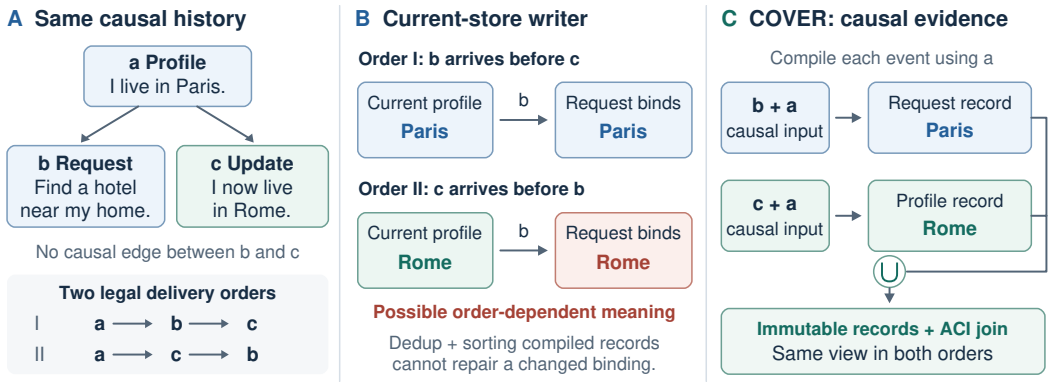


Figure 1: Conceptual illustration. A: causal parents are supplied by the runtime, not inferred from arrival order. B: sorting compiled records cannot repair an order-dependent binding. C: with fixed compilation, COVER joins immutable source-bound records, preserving the Paris request and Rome profile under both orders.

utility, execution, and mechanism evidence. Tables 3 and 4 examine reader transfer and readout choices.

## 2 Problem: Stability and Utility Are Different Targets

Let $P = (E, \prec)$ describe events and their causal dependencies. A legal schedule is a linear extension of $P$. Across $m$ schedules of root instance $i$, let $A_{is}$ be the normalized answer, $Y_{is} \in \{0, 1\}$ its correctness, and $Z_{is}$ the memory state or reader-visible view. We measure exact convergence of $Z$, the fraction of roots whose correctness changes, and pairwise answer disagreement

$$D = \frac{1}{n} \sum_i \frac{1}{\binom{m}{2}} \sum_{s<t} \mathbf{1}[A_{is} \neq A_{it}].$$

Failures and opaque outputs are retained in the denominators. We report task utility or coverage alongside every stability result to distinguish useful stability from constant abstention.

Mean accuracy $\bar{A} = \mathbb{E}_i \mathbb{E}_s Y_{is}$ and all-schedule accuracy $A_\forall = \mathbb{E}_i \min_s Y_{is}$ answer different questions. The latter requires a root to succeed in every tested order. If $k_i$ of $m$ schedules succeed, pairwise correctness disagreement is $D_i^Y = 2k_i(m - k_i)/(m(m-1))$. The gap $\Gamma = \bar{A} - A_\forall$ satisfies

$$\tfrac{1}{2} D^Y \leq \Gamma \leq \tfrac{m-1}{2} D^Y. \tag{1}$$

Both bounds are sharp (Appendix). Unchanged mean accuracy can therefore conceal a loss of reliability across schedules. Empirical $A_\forall$ measures success across the tested schedules only.

## 3 COVER: Fix Evidence Before Interpretation

**1. Compile an event from its causal past.** For event $e$, the learned input is $x_e = \phi(e, V(\mathrm{Anc}(e)))$: the event and the active records of its exact ancestors. Incomparable siblings, arrival rank, and mutable native state are excluded. In Figure 1, both children can read the Paris profile; the request cannot read the independent Rome update. The runtime must supply immutable event IDs and truthful causal dependencies. Arrival order alone cannot reconstruct which profile version a request observed.

The relevant dependency is the version an operation actually read. A runtime that issues the hotel request from the Paris profile records that profile's identity as its parent. An independent profile update does not become a parent merely because it reaches the memory first. This metadata connects the application's execution history to the compiler's evidence boundary. When the runtime cannot supply it, the construction cannot certify the intended interpretation from timestamps or textual similarity alone.

**2. Bind records to sources and preserve failed extractions.** The model proposes semantic fields; transport assigns event and evidence IDs. Literal-span and type checks bind accepted fields to source text. Unaccepted material remains available as an *opaque record*: retained source whose meaning has not been certified by the extractor. The typed implementation also checks same-clause and valid-time constraints and allows one constrained span lookup. The long-history implementation validates fact pointers independently and preserves complete chunks alongside unparsed residue. Quotation validation establishes textual provenance without establishing the truth or completeness of the extracted interpretation.

**3. Materialize once and read a stable view.** Authenticated event identities bind immutable payloads. For identity-consistent deliveries, compatible event-to-record maps merge by union, which is associative, commutative, and idempotent (ACI). Conflicting bindings under one ID are outside this guarantee; first-arrival rejection alone cannot resolve them independently of order. A record becomes readable only when its predecessors are active; pending records cannot influence compilation or answers. Repeated activation until no further record is ready handles child-first delivery. Deterministic retrieval uses source identities to break ties and applies a fixed context-packing rule.

For a delayed-parent delivery in the running example, the request remains pending until the Paris profile is available. Activating the profile makes the request eligible for compilation; the Rome update can neither supply its missing evidence nor alter its interpretation. Retransmitting the request adds no new record. This separates transport arrival, causal readiness, and reader visibility: an event may have arrived without being ready to influence an answer. The declared-parent stream extension formalizes this behavior without requiring the runtime to know future events.

**One construction principle, several evaluated readouts.** The *typed adapter* composes relations through declared entity types. The *source-retaining reader* uses learned facts or record hints alongside literal source passages and a raw fallback. The *bounded raw specialization* disables learning and retains event-time top-$C$ records. We identify these variants in every comparison: performance of one readout is not automatically performance of the others. The six-benchmark comparison evaluates the typed adapter; the full-history gains use source-retaining readers.

## 4 What the guarantee requires

**Theorem 1: robustness has two local requirements.** Consider deterministic transitions on reachable causal prefixes. Two states are equivalent when all legal future continuations and required queries yield the same answers. Schedule-and-retry robustness holds exactly when *retry absorption* and *independent-event commutation* hold modulo this equivalence at every reachable prefix. The former removes repeated events; the latter permits swapping two jointly enabled incomparable events. The full quantified statement and proof are in Appendix B.1. Exact-ID enforces the first condition, not the second.

The proof first deletes retries using absorption. Any two remaining linear extensions can then be connected by adjacent swaps of incomparable events. Each swap preserves future behavior by the local commutation law, establishing equivalence of the complete traces. Necessity follows by comparing a reachable prefix with its retried version, or its two enabled extensions, followed by any common legal continuation. Quantifying over those continuations is essential: agreement on today's query can conceal a difference that a later event or query exposes.

**Theorem 2: compression must preserve future distinctions.** For an ACI representation of causal event sets, identify two histories only when union with every causal ideal gives the same required answer to every query. This relation is a join congruence, and its quotient is the smallest reachable behaviorally exact state space within that ACI class. If $N_Q$ classes are reachable, at least $\lceil \log_2 N_Q \rceil$ bits are needed. Membership queries over $n$ independent events require $n$ bits even before storing payloads. These are applications of classical trace and future-context quotient ideas (Mazurkiewicz, 1995; Myhill, 1957; Nerode, 1958; Adámek et al., 2018). Their role is to expose what an order-robust learned interface may discard.

**COVER's structural guarantee.** For an authenticated finite DAG, a fixed total compiler, exact-ancestor inputs, deterministic validation, ACI admission, fair activation, and a deterministic reader

of the active view, every legal schedule produces the same terminal map, view, and answer. Rank induction fixes each ancestor record before its descendants. An extension to streams with unknown future events provides equal-prefix and pointwise convergence under authenticated finite parent declarations and decreasing natural ranks (Appendix B.2). This guarantee requires a fixed deterministic compilation pipeline; model aliases and temperature zero alone are insufficient.

**Consequences for practical controls.** For byte-exact convergence of immutable per-event maps, compilation must be invariant to incomparable events. Exact compression must preserve distinctions needed by future queries and by read-only execution after any common legal continuation (Theorems 3–6). Retaining all source bytes preserves information in the memory store, although a finite reader prompt may still displace useful evidence. Source preservation therefore ensures information availability without guaranteeing a utility gain for a fixed reader.

For two deterministic selectors $C, C'$ of literal evidence, fix a deterministic reader, a common distribution of histories, queries, and targets, and a loss in $[0, 1]$. Their risks satisfy $|R(C') - R(C)| \leq \Pr\{C' \neq C\}$, because losses agree whenever the contexts agree. The bound is tight even when both selectors are schedule invariant (Appendix B.2). Adding a passage can help or hurt by displacing evidence in a fixed prompt. This motivates separate representation, retrieval, and reader-call controls in the utility experiments.

At capacity $C < n$, an exact-ID store of raw events without a sufficient summary either retains different events across schedules or omits some events in every schedule. FIFO and canonical top-$C$ realize these alternatives. For three events and capacity two, both average $2/3$ membership accuracy; over all six permutations, their all-schedule accuracies are respectively zero and $2/3$.

**Canonical replay and cost.** Trusted total-order keys also permit canonical rebuilding. A suffix-replay policy costs $n + I(\pi)$ calls, where $I$ counts arrival inversions. Cached canonical insertion using the same event-local compiler instead needs one compilation per event, matching COVER. Both are relevant controls; the quadratic expected suffix-replay cost is policy-specific.

## 5 Experiments: separate robustness from semantic utility

Experiments test schedule robustness, semantic utility, and transfer. Native schedule tests use four systems, three models, and six legal schedules, with 24-root screens, a 60-root MemoryOS extension, and 1,000 synthetic COVER roots. Natural tasks use LongMemEval (Wu et al., 2025), LoCoMo (Maharana et al., 2024), and MemoryAgentBench tasks (Hu et al., 2026); broader panels add RULER (Hsieh et al., 2024) and CL-Bench (Asawa et al., 2026). Writers operate without access to target answers or answer provenance. To account for dependence, paired uncertainty estimates resample whole source histories where available, treating questions and fixed-reader evaluations from the same history as dependent. We distinguish held-out evaluation from repeatedly inspected development. Panels evaluated with local judges use protocols that differ from the benchmark-official APIs. Figure 2 separates utility, execution, and mechanism evidence. Each panel retains its own tasks and controls, distinguishing natural dialogue, fixed-plan execution, and synthetic probes.

### 5.1 Order sensitivity exists; deduplication solves only part of it

Native state diverges in 11/12 system–model cells; answer divergence also occurs in 11/12, and correctness flips in 10/12. Identical-order replays vary in two Mem0 cells, so scheduling is not the only source of variation. In the 60-root extension all three MemoryOS models produce divergent states on every root. On the separate 1,000-root synthetic workload, COVER achieves zero state and view divergence and 99.3% accuracy, with a simultaneous risk upper bound of 1.44%. These synthetic results are not directly comparable to natural-data accuracy (Appendices C–D).

**Which components matter?** On the synthetic workload, removing transport-bound identity lowers accuracy from 99.3% to 97.4%, and removing the one-shot pointer repair lowers it to 98.5%. Removing typed entity-reference normalization reduces multi-hop accuracy to 0.1%, identifying composition as essential for this generator. Dropping opaque records or allowing unsafe passthrough raises accuracy to 100% by removing seven abstentions, with no observed unsafe-wrong answers;

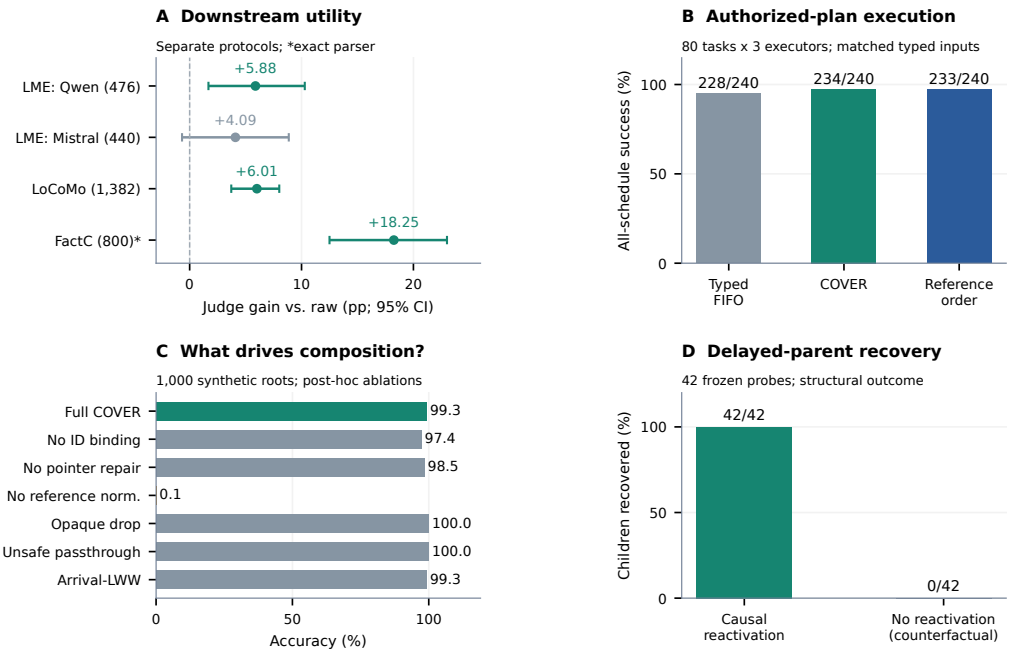


Figure 2: **Utility, execution, and mechanism evidence.** A: paired gains on complete-history LongMemEval (compiler indicated), LoCoMo, and exact-parser FactConsolidation; protocols and sample sizes differ. B: identical typed authorizations across six schedules; COVER gains 2.50 points over FIFO, while the reference-order control is close. C: frozen-output synthetic ablations, including the two opacity variants that reach 100% by removing abstentions. D: causal reactivation recovers delayed children; the no-reactivation control is a mechanical counterfactual. Panels are separate experiments, not a cross-dataset ranking.

arrival-order replacement leaves accuracy unchanged. A separate child-first probe recovers 42/42 delayed children with reactivation, versus 0/42 under a mechanical no-reactivation counterfactual. Figure 2C–D separates composition from delivery handling; the finite-sample safety benefit of opaque retention remains unestablished.

**Event-local compilation versus replay.** Across 1,000 synthetic roots, 5,000 initial extractions and 25 selective repairs require 1.005 model calls per event. Replaying every affected suffix would average 8.50 calls per five-event schedule, compared with COVER's 5.025. Canonical insertion can also reuse the same event-local outputs. With caches keyed by event identity, local-input hash, and fixed compiler configuration, it matches all 30,000 nonempty prefix maps and all 6,000 terminal maps, using the same 5.025 first-use lookups per schedule. This frozen-output replay makes no new model calls and does not compare runtime. Avoided recomputation depends on the reusable compilation boundary, which both implementations share (Appendix H).

**Retries: a strong simple baseline.** On complete LongMemEval histories with controlled 10% retry, exact-ID removes all 102 correctness-flip roots (Table 1A). Canonical raw materialization additionally aligns states, retrieval order, contexts, and answers. Relative to flat append, its mean-accuracy difference is $-0.30$ points ($[-1.87, 1.27]$): the observed benefit is schedule consistency, with no significant gain in average accuracy. The 10% retry rate defines a controlled stress test; production retry rates remain unmeasured.

**Bounded state: stable retention matters after deduplication.** The duplicate-free all-500 test retains 192 rounds under six schedules that preserve within-session order. Event-time top-$C$ improves all-schedule accuracy over exact-ID FIFO from 13.2% to 42.8%: $+29.6$ points ($[25.6, 33.6]$), while the mean gain of 2.2 points remains uncertain ($[-0.37, 4.70]$). The mechanism is canonical retention, with learned compilation off; it is a simple specialization of the stable-view design.

Table 1: **What remains after deduplication?** Matched comparisons within each panel separate retry handling, bounded retention, and action ordering. Percentages are used unless a denominator is shown. Bold highlights the reported all-schedule gains and exact convergence. Learned compilation is disabled in the raw panels; reference plans are supplied in the action relay.

**A. Unbounded LongMemEval-S: 500 histories, reorder + 10% retry**

| Materialization | Mean accuracy | Answer $D$ | Flip roots / 500 |
|---|---|---|---|
| Flat append | 47.90 | 81.61 | 102 |
| Exact-ID deduplication | 47.60 | 0.24 | 0 |
| Canonical raw COVER | 47.60 | **0.00** | **0** |

**B. Bounded LongMemEval-S: the same 500 histories, no retries,** $C = 192$

| Retention after exact-ID | Mean accuracy | All-schedule | Answer $D$ | Flip rate |
|---|---|---|---|---|
| Arrival FIFO | 40.6 | 13.2 | 94.7 | 49.4 |
| Event-time top-$C$ | 42.8 | **42.8** | **0.0** | **0.0** |

**C. Reference-plan relay: 80 tasks × 3 executors, six schedules**

| View / order | Mean success (%) | All-schedule (/240) | Plan flips (/240) | Unsafe (/240) | Max. exposure (simulator $) |
|---|---|---|---|---|---|
| Raw arrival FIFO | 25.5 | 10 | 230 | 139 | 12,490 |
| Raw canonical replay | 23.3 | 56 | 0 | 59 | 6,810 |
| Typed arrival FIFO | 96.9 | 228 | 8 | 5 | 6,173 |
| COVER typed union | 97.5 | **234** | **0** | 4 | 2,902 |
| Typed reference order | 97.1 | 233 | 0 | 5 | 6,173 |

**Action relay: measure the increment beyond compression.** Public $\tau$-bench and AgentDojo tasks (Yao et al., 2025; Debenedetti et al., 2024) supply explicit authorized reference plans to three downstream executors. Across 80 roots, success requires exact simulator database equality. The large raw-FIFO deficit in Table 1C combines capacity and representation effects. Holding typed content, prompt, model, and budget fixed, COVER versus typed FIFO improves all-schedule success by **2.50 points** ($[0.83, 4.58]$) and removes eight plan flips. Typed reference order reaches 233/240 versus COVER's 234/240; a practically important advantage over that canonical control is unestablished. In airline task 41, raw FIFO can cancel protected reservations despite a no-write plan, exposing $12,490 of simulator value; COVER makes no call across all six schedules. This relay evaluates execution fidelity to authorized reference plans. Autonomous planning and realized financial losses are outside its scope (Appendix J).

### 5.2 Source-retaining readers can recover utility, with important limits

**Natural dialogue exposes the cost of a restrictive reader.** On Fresh83, original COVER F1 is 9.76%, versus Mem0's 20.16% (Table 2A). A shared set reader raises COVER raw-memory F1 to 40.09%, compared with Mem0's 35.15%, although the difference remains uncertain. The 16.27-point increase in coverage accompanies a 10.64-point increase in covered-wrong mass. Thus, schedule stability alone does not ensure safer selective answers.

**Complete histories support selective semantic gains.** LongMemEval-S contains 500 questions, 23,867 sessions, and 246,750 turns. Each learned compiler processes all 38,432 chunks query-independently, retaining sources and validated fact pointers. A fixed cascade uses the fact-reader answer when it parses successfully and is not `INSUFFICIENT`; otherwise, it uses the raw-reader answer. On roots excluded from each compiler's pilot, Qwen and Mistral compilation improve F1 by **7.51** and **5.32** points (Table 2B). Qwen's judge gain is positive; Mistral's judge interval crosses zero. Opacity is 3.42% and 10.86%, respectively. The preregistered Mistral facts-only judge contrast is $+0.68$ points ($[-4.32, 5.68]$); its cascade results are secondary.

Source retention also changes resource costs. In the Qwen-compiled complete-history experiment, the fact-plus-evidence payload occupies $1.75\times$ the source size; index compression is unmeasured. The 1,314 opaque chunks remain stored, and all 72 reader parse failures across 1,500 evaluations remain in the denominators. Opaque sources preserve information extraction missed, but storage

Table 2: **Semantic utility depends on the representation and reader.** Panels use separate protocols and are not a pooled ranking. Values are percentages; bracketed effects have paired 95% bootstrap intervals. Panel A retains the original utility loss. In B–C, bold marks effects whose intervals exclude zero; primary/secondary status is specified in the text. $J$: local-judge accuracy.

**A. Fresh83: robustness and utility must be measured separately**

| Memory / reader | Exact | F1 | Covered | Wrong | Answer $D$ | Flip |
|---|---|---|---|---|---|---|
| Mem0 / original | 15.26 | 20.16 | 55.42 | 40.16 | 18.88 | 16.87 |
| COVER / original | 8.43 | 9.76 | 30.12 | 21.69 | 0 | 0 |
| Mem0 / shared set reader | 26.91 | 35.15 | 75.30 | 48.39 | 27.71 | 21.69 |
| COVER raw / shared set reader | 32.53 | 40.09 | 91.57 | 59.04 | 0 | 0 |

**B. Complete-history LongMemEval-S: two learned compilers, fixed Qwen reader**

| Compiler ($n$) | Raw $J$ | Facts $J$ | Cascade $J$ | $\Delta J$ vs. raw | $\Delta$F1 vs. raw |
|---|---|---|---|---|---|
| Qwen-7B (476) | 32.98 | 35.92 | 38.87 | **+5.88 [1.68, 10.29]** | **+7.51 [4.04, 10.96]** |
| Mistral-7B (440) | 32.50 | 33.18 | 36.59 | +4.09 [−0.68, 8.86] | **+5.32 [1.57, 9.08]** |

**C. Transfer: learned LoCoMo compilation and a separate exact-parser test**

| Dataset / fixed view | Exact | F1 | $J$ | $\Delta J$ vs. raw |
|---|---|---|---|---|
| LoCoMo (1,382): canonical raw | 12.59 | 29.76 | 31.11 | — |
| Semantic facts | 16.86 | 36.54 | 34.66 | **+3.55 [0.87, 5.91]** |
| Raw + facts (secondary) | 18.67 | 41.06 | 41.10 | **+9.99 [8.16, 11.66]** |
| Frozen facts/raw cascade | 18.09 | 39.24 | 37.12 | **+6.01 [3.72, 8.01]** |
| FactConsolidation (800): raw | 28.25 | 29.43 | 25.88 | — |
| Exact typed materialization | 54.25 | 54.82 | 44.13 | **+18.25 [12.50, 23.00]** |

Table 3: **Reader transfer under matched two-call limits.** All 476 LongMemEval histories use fixed Qwen compilation and a Qwen-7B judge. Raw-2 verifies its first draft against the same canonical raw evidence. COVER uses the facts/raw cascade. This matches reader-call limits, not compilation, retrieval, or token costs. Accuracy is reported in percent; paired differences have 95% root-bootstrap intervals. Bold marks intervals excluding zero in either direction.

| Reader | Raw-1 | Raw-2 | COVER | COVER − Raw-2 |
|---|---|---|---|---|
| Qwen2.5-7B | 32.98 | 25.42 | 38.87 | **+13.45 [9.03, 18.07]** |
| Mistral-7B | 28.57 | 24.58 | 30.67 | **+6.09 [1.89, 10.29]** |
| Llama-3.2-3B | 12.39 | 25.00 | 17.65 | **−7.35 [−11.55, −3.36]** |

does not ensure that retrieval selects them. These costs are separate from the cascade's additional answering calls.

These comparisons use canonical raw retrieval, already controlling IDs and order. They measure semantic/readout utility beyond that consistency baseline. The Qwen cascade uses 577 reader calls, $1.21\times$ raw-only, plus compilation.

**Does another reader call explain the gain?** Table 3 fixes compilation and varies readers on the same 476 roots. The cascade exceeds Raw-2 for Qwen and Mistral but loses for Llama. Raw-2 itself underperforms Raw-1 for Qwen and Mistral. For Llama, the cascade selects raw on 406/476 roots. Routing thus interacts with reader reliability. The control matches reader-call limits, not total pipeline cost, and adds no independently retrieved evidence. Appendix K includes the separate 72B scale probe.

**Transfer is conditional on the extraction interface.** The frozen Qwen cascade improves LoCoMo judge accuracy by **6.01** points and F1 by **9.48** on 1,382 questions in ten conversations (Table 2C). On FactConsolidation, judge accuracy improves by **18.25** points with matched retrieval. This configuration uses an exact benchmark parser and source-ordinal replacement in place of the learned dialogue compiler. At 262k multi-hop, accuracy increases from 4% to 6% but remains low. The two extraction mechanisms have distinct generalization limits.

### 5.3 UTILITY ACROSS DATASETS AND BACKBONES

The official-code comparison uses five static datasets with 100 questions each and 125 source banks, plus sequential CL-Bench. The static panel fixes the Qwen-7B writer and varies five readers; CL100 varies both. Pinned revisions and budget-bounded configurations are detailed in Appendix A.

Across static datasets and readers, typed COVER averages 26.12%, versus RAW's 44.20%; its adjusted difference is $-18.08$ points ($[-22.76, -13.63]$). It exceeds MemoryOS, while intervals against the other three native systems cross zero. On CL100, COVER averages 47.0%, RAW 48.8%, and Mem0 49.2%. On the downloaded 884-task complete-context subset of CL-Bench, COVER scores 48.64%, compared with RAW's 47.62% and Mem0's 49.77%; all five adjusted COVER contrasts cross zero.

**Source-preserving readout does not remove the utility gap.** From one static compilation cache, typed replay, source expansion, and raw/typed fusion score 25.24%, 31.84%, and 37.88%, respectively. Fusion still trails RAW. A protected-RAW two-call variant scores 49.4% against matched two-call RAW's 49.0%, with F1 difference $-0.11$ points ($[-0.73, 0.49]$). Its accepted revisions fix three judged errors and damage five correct answers. Detailed source-span, projection-filter, and residual controls appear in the appendix. The matched residual comparison provides no established utility advantage over two-call RAW.

The official-code comparison and readout ablations diagnose different stages of the pipeline. Reusing one compilation cache holds the writer's outputs fixed, exposing what is lost when the reader sees typed fields alone and what source expansion recovers. The protected-RAW comparison then tests whether learned evidence improves an already competitive raw reader. Its small, uncertain difference shows why recovery from a restrictive adapter is insufficient evidence of superiority over raw retrieval. The same distinction applies when interpreting improvements from a stronger downstream model.

### 5.4 READOUT VARIANTS AND MATCHED CONTROLS

The lexical+dense DeepSeek reader has higher point accuracy than all four native configurations on four static development subsets, but not LongMemEval. Removing learned hints changes scores by $-1, -4, +2, 0, -2$ points: the benefits are selective, and source retrieval accounts for much of the recovery.

On complete-source CL100, lexical-bundle COVER scores 47/100 against native scores of 43–45, although paired comparisons do not establish superiority. Semantic ranking lowers COVER to 40/100; the query-program variant scores 37/100, matching MemOS with the same program layer and trailing MemoryOS's 42/100. In the source-first pilot, pre-indexing documents for every system yields 12/26 for COVER and 13/26 for NLP-complete Mem0; COVER without precompilation also scores 12/26. Budgets are matched within each protocol, while the program experiment changes the planning/final output allocation.

Each candidate's memory evolves from its own generated answers, while the evaluation reuses previously inspected development questions. These comparisons characterize readout choices on those data; independent confirmation of one frozen method's superiority across all six datasets remains outstanding. Appendix A.1 reports variants, budgets, and failure accounting.

## 6 RELATED WORK

Learned memory supports extraction, evolution, and tiered storage (Chhikara et al., 2025; Xu et al., 2025; Li et al., 2025; Kang et al., 2025). Other approaches optimize utility, admission, budgets, and fallback (Yan et al., 2026; Zhang et al., 2026; Zhang & Li, 2026; Bhargava & Sobral Barrento, 2026; Zhu et al., 2026). Incremental and cross-episode benchmarks vary task progression (Hu et al., 2026; Wang et al., 2026; Asawa et al., 2026); schedule evaluation holds logical events fixed and varies their legal delivery order. Event-sourced and chronology-preserving memory (Brito dos Santos Filho, 2026; Liu et al., 2026) complements the causal compilation boundary. Causality, ACI/CRDTs, and CALM provide the classical foundations (Lamport, 1978; Shapiro et al., 2011;

Table 4: **Source-preserving readouts and DeepSeek development controls.** All panels reuse development data and a Flash-family judge. Compare methods within A, within each matched row of B, and within C: source visibility, writer configuration, and output budgets differ across these comparisons. Bold in A marks the best point estimates across all rows, including ties, not significance. B–C report correct answers.

**A. Static subsets: 100 questions each; 12,000-character source-event prefix**

| Method / readout | LME | LoCoMo | FactC | EventQA | RULER |
|---|---|---|---|---|---|
| Mem0 | 59 | 67 | 11 | 84 | 23 |
| A-MEM | 65 | 56 | 10 | 83 | 28 |
| MemOS | **67** | 54 | 10 | 79 | 28 |
| MemoryOS | 7 | 29 | 12 | 77 | 22 |
| COVER: typed | 60 | 46 | 13 | 88 | 46 |
| COVER: literal-source retrieval | 66 | 75 | 18 | 89 | 78 |
| COVER: lexical+dense evidence | 66 | **78** | 19 | **92** | **85** |
| Same evidence reader, no learned hints | 65 | 74 | **21** | **92** | 83 |

**B. Complete-source CL100: 100 tasks / 56 contexts; fresh COVER trajectories**

| Protocol / COVER variant | COVER | Mem0 | A-MEM | MemOS | MemoryOS |
|---|---|---|---|---|---|
| 2,048/6,144 tokens: lexical bundles | 47 | 43 | 45 | 45 | 45 |
| Semantic ranking; same controls | 40 | 43 | 45 | 45 | 45 |
| 4,096/4,096 tokens: query programs | 37 | 32 | 34 | 33 | 42 |

**C. Source-first CL pilot: 26 tasks / 16 contexts; all native states rebuilt**

| Protocol / COVER variant | COVER | Mem0$^{+}$ | A-MEM | MemOS | MemoryOS |
|---|---|---|---|---|---|
| Compile provided sources before answering | 12 | 13 | 12 | 12 | 12 |

B: token pairs are planning/final output limits. Giving MemOS the same query program layer yields 37/100; withholding COVER's execution results yields 34/100. C: COVER without pre-query compilation also scores 12/26. $^{+}$Mem0's optional spaCy models are installed and loaded in C; A-MEM uses native metadata analysis. Earlier rows retain their original native configurations.

Hellerstein, 2026). Small official StateFuse and MELD probes separate operation-level convergence from learned admission (Appendix, Table 9).

Storage convergence specifies how replicas combine accepted records; learned memory also decides what those records mean. A convergent store can retain a mistaken profile binding consistently. COVER constrains evidence before extraction, then preserves records through order-independent updates. Canonical methods can inherit this property by adopting the same evidence boundary, as the cached control demonstrates.

## 7 Limitations and conclusion

COVER requires truthful causal provenance and fixed compilation; its stream extension cannot infer missing dependencies or resolve absent parents. Empirical scope is limited by narrow synthetic coverage, small history counts, local or same-family judges, bounded native configurations, and repeatedly inspected development data. Only MemoryOS has 60-root cross-model confirmation, and Mistral exhibits frequent parser failures. Natural utility varies with compiler, reader, and budget. The action relay evaluates fixed upstream plans and is retail-heavy. Production retry rates, isolated deployment latency, and the safety value of opaque retention remain unmeasured. Deployment evaluation should measure naturally occurring reorderings, task losses, and the full cost of writing, updating, retrieving, and answering under matched budgets.

COVER establishes a causal boundary for schedule-robust memory construction. Its measurable value is stable behavior across equivalent deliveries, with selective natural-data utility when sources and readout are designed together. Canonical controls explain much of the robustness gain, and broad accuracy superiority remains unresolved. These boundaries determine where the method is useful and what further evidence is needed.

## Reproducibility Statement

Appendix A specifies native repository versions, budget-bounded configurations, sampling, and paired uncertainty. Appendix B.1 gives the formal assumptions and proof details; Appendices J and K describe the execution and reader controls. Appendix A.1 separates reused development data from held-out comparisons. Native failures remain in reported denominators, and source visibility and compiler variants are identified for each comparison.

## Ethics Statement

The experiments use public benchmark material and synthetic event streams. Source-retaining memory can expose personal information or untrusted instructions present in its inputs; source attribution does not establish privacy, truth, or instruction authority. Deployment therefore requires access controls, retention policies, and separation of evidence from executable instructions. Action-relay outcomes and monetary exposure are measured in simulators and do not represent realized financial harm or deployment safety.

## AI Use Statement

The research idea, methodological design, and core algorithms were developed by the human authors. AI tools were used to assist with implementation, and all AI-generated code was reviewed and verified by the authors. GPT-6 Astra and Gemini 3.8 Flash were additionally used for language polishing and improving the clarity and readability of the manuscript.

## References

Jiří Adámek, Stefan Milius, and Henning Urbat. A categorical approach to syntactic monoids. *Logical Methods in Computer Science*, 14(2:9):1–34, 2018. doi: 10.23638/LMCS-14(2:9)2018.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control, 2021.

Parth Asawa, Christopher M. Glaze, Gabriel Orlanski, Ramya Ramakrishnan, Benji Xu, Asim Biswal, Vincent Sunn Chen, Frederic Sala, Matei Zaharia, and Joseph E. Gonzalez. Continual learning bench: Evaluating frontier AI systems in real-world stateful environments. *arXiv preprint arXiv:2606.05661*, 2026. URL https://arxiv.org/abs/2606.05661.

Nishant Bhargava and Rodrigo Sobral Barrento. MEMAUDIT: An exact package-oracle evaluation protocol for budgeted long-term LLM memory writing. *arXiv preprint arXiv:2605.02199*, 2026. URL https://arxiv.org/abs/2605.02199.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-24/issue-2/Equivalent-Comparisons-of-Experiments/10.1214/aoms/1177729032.full.

Elzo Brito dos Santos Filho. ESAA-Conversational: An event-sourced memory layer for continuity, handoff, and curation across heterogeneous LLM coding agents. *arXiv preprint arXiv:2606.23752*, 2026. URL https://arxiv.org/abs/2606.23752.

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.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramèr. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems Datasets and Benchmarks Track*, 2024. URL https://openreview.net/forum?id=m1YYAQjO3w.

Joseph M. Hellerstein. Complete CALM: A coordination criterion for specifications. *arXiv preprint arXiv:2602.09435*, 2026. URL https://arxiv.org/abs/2602.09435.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963. doi: 10.1080/01621459.1963.10500830.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024. URL https://arxiv.org/abs/2404.06654.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2507.05257.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*, 2025. URL https://arxiv.org/abs/2506.06326.

Leslie Lamport. Time, clocks, and the ordering of events in a distributed system. *Communications of the ACM*, 21(7):558–565, 1978. doi: 10.1145/359545.359563.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, et al. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*, 2025. URL https://arxiv.org/abs/2507.03724.

Yifan Simon Liu, Liam Gallagher, Faeze Moradi Kalarde, Jiazhou Liang, Armin Toroghi, and Scott Sanner. Temporal order matters for agentic memory: Segment trees for long-horizon agents. *arXiv preprint arXiv:2606.04555*, 2026. URL https://arxiv.org/abs/2606.04555.

Lauri Lovén, Jaakko Sauvola, Jukka Riekki, and Sasu Tarkoma. MELD: A protocol for merging knowledge across distributed agentic memories. *arXiv preprint arXiv:2608.16357*, 2026. URL https://arxiv.org/abs/2608.16357.

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

Antoni W. Mazurkiewicz. Introduction to trace theory. In Volker Diekert and Grzegorz Rozenberg (eds.), *The Book of Traces*, pp. 3–41. World Scientific, 1995. doi: 10.1142/9789814261456_0001.

John Myhill. Finite automata and the representation of events. Technical Report WADD Technical Report 57-624, Wright Air Development Division, 1957.

Anil Nerode. Linear automaton transformations. *Proceedings of the American Mathematical Society*, 9(4):541–544, 1958. doi: 10.1090/S0002-9939-1958-0135681-9.

Anthony T. Nixon. The Myhill–Nerode theorem for bounded interaction: Canonical abstractions via agent-bounded indistinguishability. *arXiv preprint arXiv:2603.21399*, 2026. URL https://arxiv.org/abs/2603.21399.

Marc Shapiro, Nuno Preguiça, Carlos Baquero, and Marek Zawirski. A comprehensive study of convergent and commutative replicated data types. Technical Report RR-7506, INRIA, 2011. URL https://inria.hal.science/inria-00555588.

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. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

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.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 12805–12825, 2026. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. $\tau$-bench: A benchmark for tool-agent-user interaction in real-world domains. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=roNSXZpUDN.

Kehao Zhang, Shangtong Gui, Sheng Yang, Wei Chen, and Yang Feng. Learning to remember: End-to-end training of memory agents for long-context reasoning. *arXiv preprint arXiv:2602.18493*, 2026. URL https://arxiv.org/abs/2602.18493.

Yan Zhang and Shibo Li. ConsistencyGate: Preventing memory contamination in LLM agents via self-consistency admission control. *arXiv preprint arXiv:2607.22962*, 2026. URL https://arxiv.org/abs/2607.22962.

Qiming Zhu, Shunian Chen, Rui Yu, Zhehao Wu, and Benyou Wang. From lossy to verified: A provenance-aware tiered memory for agents. *arXiv preprint arXiv:2602.17913*, 2026. URL https://arxiv.org/abs/2602.17913.

Mingxi Zou, Zhihan Guo, Langzhang Liang, Zhuo Wang, Qifan Wang, Qingsong Wen, Irwin King, Lizhen Qu, and Zenglin Xu. Remember the decision, not the description: A rate-distortion framework for agent memory. *arXiv preprint arXiv:2605.10870*, 2026. URL https://arxiv.org/abs/2605.10870.

## A Official comparison protocol and uncertainty

The design selects 100 label-blind roots from each of LongMemEval, LoCoMo, FactConsolidation, EventQA, RULER-QA, and CL-Bench by stable hash ranking within source/category strata. The first five benchmarks share a Qwen2.5-7B writer and judge across six conditions, with five fixed readers. CL-Bench repeats the conditions with Qwen2.5-3B/7B/72B, Llama-3.2-3B, and Mistral-7B writers/readers, holding the Qwen2.5-7B judge fixed. The 100 CL-Bench tasks belong to 56 contexts; paired analysis resamples contexts jointly across conditions and fixed readers. Neither tasks within a context nor reader-specific copies of a task are treated as independent roots. Before aggregation, we check code revisions, public-input hashes, common answer formatting, and failure accounting. Earlier local-reimplementation values are excluded. The static comparison changes readers, not writers; CL-Bench changes both.

**Static tasks and paired analysis.** Each of the five static datasets contributes 100 questions. Their memory-bank counts are 100, 10, 8, 5, and 2, respectively (125 banks, 7,404 source-event occurrences in total). RAW uses BM25 retrieval of source events; the COVER adapter retrieves typed records or opaque fallback. Each reader receives at most ten retrieved items and 6,200 context tokens, with a 192-token answer cap. One fixed Qwen-7B judge uses temperature 0, top-$p$ 1, seed 0, a ten-token output cap, and a yes/no decision. All 15,000 task outcomes are scored without judge errors. Mem0 returns answers for 496 questions per reader; the four native failures receive zero scores, without substitute answers or reruns.

For 20,000 bootstrap draws, we resample whole banks within each dataset, retaining their questions and sharing weights across all methods and readers. Contrasts weight the five datasets and five fixed readers equally. Five primary COVER contrasts use 99% percentile intervals (Bonferroni nominal familywise 95%). Against RAW, Mem0, A-MEM, MemOS, and MemoryOS, respectively, the differences are $-18.08$ $[-22.76, -13.63]$, $-0.52$ $[-4.36, 3.38]$, $-2.96$ $[-7.24, 1.47]$, $-2.88$ $[-7.14, 1.61]$, and $+10.24$ $[6.02, 14.41]$ points. This post-outcome bootstrap is approximate; the

Table 5: Readout development on 500 static questions: five-reader accuracy (%) at 6,200 context tokens. Bold/underline mark best/second-best within columns, not significance. Original RAW/COVER remain separate references; compiler cost is additional. The protected-RAW residual comparison is consolidated in Table 10.

| Readout | Qwen3 | Llama3 | Qwen7 | Mistral7 | Qwen72 | Avg. |
|---|---|---|---|---|---|---|
| RAW source text | **40.8** | **39.0** | **46.0** | **43.2** | **52.0** | **44.20** |
| COVER (original) | 22.6 | 22.6 | 27.0 | 25.0 | 33.4 | 26.12 |
| Typed replay (new compiler outputs) | 21.6 | 20.0 | 28.0 | 24.0 | 32.6 | 25.24 |
| Source expansion + diversification | 29.6 | 26.0 | 33.6 | 31.4 | 38.6 | 31.84 |
| Raw/typed source fusion | <u>35.4</u> | <u>31.4</u> | <u>39.0</u> | <u>38.8</u> | <u>44.8</u> | <u>37.88</u> |

two RULER and five EventQA banks particularly limit its reliability. The fixed datasets and readers are not samples from a model or benchmark population.

All 3,000 CL-Bench answers were graded by one fixed compact local Qwen2.5-7B judge (temperature 0, top-$p$ 1, seed 0, a 128-token output cap, JSON output), with no judge errors. All methods share the 2,048-token answer cap and 13,000-character augmented-input setting. RAW supplies the bounded original messages without cross-task memory; it is not raw append. The COVER adapter compiles each trajectory into evidence-bound records, using opaque source fallback when no typed record is accepted. It is distinct from the raw-and-fact reader used in the full-history experiments. The 100 selected tasks provide 44 cross-task transitions; 43 of the 56 contexts omit earlier source tasks. This limits the sequential-memory estimand without negating its negative result.

Mem0 is evaluated with a 1,000-token writer cap, temperature 0, local MiniLM embeddings, and a local Qdrant store. Its pinned vLLM configuration defaults to 2,000 tokens. The optional fastembed BM25 and spaCy features are unavailable in this runtime, so the comparison measures these budget-bounded configurations rather than each system's full default.

An outcome-aware sensitivity study increases only Mem0's writer output cap to 2,000 tokens across all 125 static banks. Accuracy averages 26.24%, compared with 26.64% under the original cap: $-0.40$ points, with a secondary pointwise 95% bank-bootstrap interval of $[-2.68, 1.81]$. Five native failures per reader receive zero scores. Three banks were preselected for reuse at the same cap independently of outcome; the remaining banks were rebuilt. Optional NLP/BM25 features remain unavailable, so the study isolates the writer output budget under the existing configuration rather than reproducing full defaults.

We resample all 56 contexts jointly across six methods and five fixed backbones for 20,000 draws. Five primary contrasts average task-weighted accuracy equally across those backbones. Their 99% percentile intervals use Bonferroni adjustment for nominal familywise 95% coverage; bootstrap coverage is approximate, not finite-sample exact. COVER minus RAW, Mem0, A-MEM, MemOS, and MemoryOS is, respectively, $-1.80$ $[-5.00, 1.29]$, $-2.20$ $[-5.90, 1.84]$, $-0.40$ $[-5.51, 3.92]$, $+0.80$ $[-3.45, 4.48]$, and $+1.20$ $[-3.23, 6.30]$ points. Every interval crosses zero. Reader-specific 95% intervals are secondary and pointwise. The analysis plan was specified after collection began and is not an outcome-independent preregistration. Shared serving supports accuracy comparison but not isolated latency or cost estimates; A100 and A800 timings are not pooled.

**Complete episodes.** The outcome-aware extension includes every task in the downloaded `CL-bench_context_ge5` slice: 884 tasks in 120 complete contexts, with 764 within-context transitions. It is not the entire CL-Bench release. All six methods are run afresh with Qwen-7B writers/readers and the same answer interface and judge; earlier 100-task checkpoints are not reused. All 5,304 grades succeed. RAW, Mem0, A-MEM, MemOS, MemoryOS, and COVER solve 421, 440, 436, 425, 439, and 430 tasks, respectively. Paired analysis resamples 120 whole contexts for 20,000 draws with shared weights across methods. The five task-weighted COVER contrasts use 99% percentile intervals (Bonferroni nominal familywise 95%): $+1.02$ $[-2.22, 4.38]$, $-1.13$ $[-4.41, 2.14]$, $-0.68$ $[-3.81, 2.65]$, $+0.57$ $[-2.69, 3.78]$, and $-1.02$ $[-4.37, 2.49]$ points, in the same baseline order. Secondary accuracy on the 764 later tasks is 46.99%, 49.21%, 48.56%, 47.25%, 49.48%, and 48.30%. These secondary outcomes do not replace the full-task comparison or establish equivalence.

**Source-preserving readout ablations.** The three new variants in Table 5 share one fixed Qwen-7B compiler configuration: 7,206 unique cached source compilations cover the 7,404 input occurrences. Typed replay retrieves the new typed outputs. Source expansion deduplicates the typed ranking by source identity and returns up to ten full source events; it changes both text retention and source diversity. The hybrid combines raw and typed source rankings with equal-weight reciprocal-rank fusion, offset 60, and returns ten source events. All use the same 6,200-token cap, five readers, 192-token answer cap, and static judge. All 7,500 new answers are graded without judge errors. Original RAW and COVER answers are separate references, not regenerated with this new compiler configuration.

Shared within-dataset bank resampling uses 20,000 draws. The two primary hybrid contrasts use 97.5% percentile intervals for nominal familywise 95% coverage: hybrid minus typed replay is $+12.64$ $[7.77, 17.71]$, and hybrid minus RAW is $-6.32$ $[-9.36, -3.42]$ points. Expansion minus typed replay is $+6.60$ $[3.92, 9.38]$ with a secondary pointwise 95% interval. These outcome-aware comparisons and the small bank counts limit inferential strength. They measure readout utility under fixed compilation, not causal attribution to text length alone, compiler replication, or schedule robustness.

**DeepSeek evidence-reader protocol.** Table 4 uses the existing 100-task subsets and a non-thinking DeepSeek Flash writer, reader, and method-blind judge. All native scores come from completed pinned official-code runs, with the same optional Mem0 dependency omissions and A-MEM schema translation described above. The hosted model alias is not an immutable checkpoint. This panel is separate from the fixed-Qwen-writer, five-reader comparison; differences between their absolute scores cannot be assigned to the memory method alone.

Every static method sees at most the first 12,000 characters of each source event. Hints previously compiled from longer events are excluded, including hints whose quotation falls inside that prefix, because their wording may depend on text the reader cannot see. Source windows contain 512 Qwen-tokenizer tokens at stride 384. The v11 reader fuses lexical source ranks and evidence-bound record hints with reciprocal-rank weights $1, 0.5$ and offset 60. The v12 reader additionally ranks the first 96 lexical candidates with local MiniLM embeddings and adds that rank with weight one. Each window is split into two fixed character halves; their normalized embeddings are averaged and renormalized. Similarities are rounded to six decimals before immutable source-ID tie-breaking. Both readers return at most ten nonoverlapping literal windows, within 6,200 Qwen-tokenizer tokens, to the unchanged 192-output-token answer interface. The source-only ablation keeps the same candidate pool, dense encoder, and budget while removing record hints. Each variant is fixed across all five static datasets.

The CL100 interface preserves original messages within an augmented ceiling of 13,000 characters and a 2,048-token answer limit. The rule-index variant (v11) adds a query-independent generated index. The source-line variant (v12) instead appends at most eight exact system-source lines selected using the public question and a fixed constraint-word detector, without an additional compiler call. Both variants trail the earlier typed adapter on CL100. These prompt interventions differ from the static retrieval experiments, so the six-dataset results do not isolate one common intervention.

All 1,100 v12 condition-by-task outcomes were generated before labels were loaded for grading, and every outcome is retained. Paired nominal 95% cluster-bootstrap intervals for v12 minus the strongest native point estimate are, in dataset order, $-1$ $[-9, 7]$, $+11$ $[-1, 24]$, $+7$ $[1.96, 12.24]$, $+8$ $[3, 14]$, $+57$ $[54, 60]$ against A-MEM, and $-2$ $[-8.25, 4.26]$ points. MemOS ties A-MEM on RULER; that paired interval is $[52, 62]$. These intervals are exploratory, unadjusted for multiple comparisons and post-selection, with respectively 100, 10, 8, 5, 2, and 56 history clusters. The subsets have been inspected during earlier development. Consequently, these comparisons do not instantiate Proposition 3 or establish improvement on unseen histories. In particular, the two RULER histories cannot support a broad population-level claim despite the large observed margin.

Removing record hints yields 65/74/21/92/83% on the five static datasets. The LoCoMo difference is four points (nominal interval $[1, 9]$), whereas FactConsolidation loses two points with hints ($[-5, 0]$). A generally beneficial learned-index effect is therefore unestablished. The computational checks verify literal provenance and budgets for all 1,000 static contexts, and identical contexts in ten real-encoder reorder/duplicate probes spanning all five static datasets. These finite checks concern the source index and do not replace the causal-DAG assumptions of the full memory guarantee.

### A.1 COMPLETE-SOURCE AND SOURCE-FIRST DEVELOPMENT

The later Flash experiments retain every original message of the selected 100 CL tasks in 56 contexts. Each method writes its own generated answer before the next selected task; grading begins after all generation and writes finish. The base reader uses a 2,048-token plan, a 6,144-token final answer, and up to 8,000 additional memory characters. Writers have an 8,192-token output ceiling. All these tasks were previously used in development. Exact-request response reuse and any completed-context reuse are documented separately from fresh memory-state construction; incremental API cost is not isolated deployment cost.

**Complete-source representation and reader controls.** The lexical-bundle candidate scores 47/100, with 26/56 first tasks and 21/44 later tasks correct. Mem0, A-MEM, MemOS, and MemoryOS score 43, 45, 45, and 45. The native A-MEM path with its own metadata analysis also scores 45; an author-paper memory-layer adaptation scores 40 and does not replace the stronger library result. Dependency-complete retrieval scores 44, a fresh split-bundle trajectory scores 45, and a source-bound quoted planner scores 47. The split readout on a frozen earlier trajectory scores 49, illustrating why readout interventions and fresh own-answer trajectories must be distinguished. First and later tasks differ in difficulty; their accuracy difference is not a causal estimate of memory's benefit.

**Query-program layer.** This experiment gives every condition 4,096 planning and 4,096 final output tokens. Conditions with the program layer additionally receive a source-derived catalog and execute a bounded set of source-bound arithmetic operations. COVER, Mem0, A-MEM, MemOS, and MemoryOS score 37, 32, 34, 33, and 42 out of 100. Withholding COVER's execution results gives 34; attaching the same program layer to MemOS gives 37. COVER minus the latter is zero points with nominal interval $[-9.09, 9.09]$. Five COVER plans fail whole-program parsing or program limits, and 198 catalog-ID references are unresolved; all task outcomes remain in the evaluation. An earlier configuration was stopped before grading because plans were frequently truncated. The smaller final-answer budget and changed trajectories prevent attributing differences from the base protocol solely to the program layer. This comparison evaluates the mechanism under matched conditions while allowing prompts to differ.

**Semantic source ranking.** Lexical/semantic rank fusion uses a frozen local MiniLM encoder and literal whole evidence bundles, with the base compiler and shared reader unchanged. A fresh 100-task COVER trajectory scores 40/100; the four complete base controls are reused in full, not selected by outcome. The difference from lexical COVER is $-7$ points with nominal context-bootstrap interval $[-14.86, -0.83]$. Three final answers reach the output cap. Exact source binding and reproducible fixed-state selection do not prevent the accuracy decline. The experiment does not support promoting semantic ranking.

**Source-first pilot.** Six fresh conditions use all 26 selected tasks in 16 contexts chosen by a fixed context-ID hash. Every method ingests each newly visible system document before answering; future tasks are unavailable. COVER compiles that document into literal evidence bundles, while its control only stores the source before the first query. Both still compile completed interactions afterward. All four native systems receive the same early source opportunity. Mem0 runs in an isolated environment with spaCy 3.8.16 and `en_core_web_sm` 3.8.0 verified through native loaders; A-MEM uses native metadata analysis. This configuration change does not repair or relabel earlier Mem0 results.

COVER scores 12/26, NLP-complete Mem0 13/26, and each other native method and the no-precompilation COVER control 12/26. All 156 final answers completed normally, with six plans reaching their cap. The 16 COVER sources produced 176 accepted evidence bundles, which were selected a total of 27 times during planning and final retrieval. The intervention therefore affected retrieval. Its correctness vector is identical to the three 12/26 native vectors and the ablation. Resampling all-zero paired differences gives a degenerate empirical bootstrap interval; this establishes neither population equivalence nor zero disagreement risk. COVER minus Mem0 is $-3.85$ points, with nominal interval $[-20.00, 6.52]$ and four-comparison-adjusted interval $[-23.81, 7.89]$.

Table 6: Qwen7 span-packing accuracy (%), 100 questions per static dataset. Original RAW/COVER are separate reference runs; all four new conditions are shown.

| Dataset | RAW | COVER | Raw spans | Event spans | Typed spans | Hybrid spans |
|---|---|---|---|---|---|---|
| LongMemEval | 40 | 17 | 43 | 48 | 27 | 32 |
| LoCoMo | 51 | 32 | 53 | 51 | 37 | 45 |
| FactConsolidation | 22 | 3 | 19 | 20 | 6 | 14 |
| EventQA | 63 | 48 | 73 | 69 | 56 | 63 |
| RULER | 54 | 35 | 63 | 60 | 36 | 53 |

**Scope of the exploratory analyses.** The query-program and semantic-ranking trials and the source-first pilot evaluate individual components on development data without establishing superiority. The reserved 64 contexts containing 445 tasks are disjoint from recent development but overlap historical 884-task experiments. Full-source ingestion on the five static datasets has only a technical prefix pilot of 25 states and 300 event writes, without task-accuracy evaluation. A six-dataset comparison of one fixed COVER version with complete rebuilt source histories remains incomplete. The older static results in Table 4A retain their original source limits.

**Earlier CL100 evaluation with bounded messages.** Under the earlier bounded-original-message CL100 protocol, Mem0, A-MEM, MemOS, and MemoryOS score 16/15/11/14, respectively; typed, source-backed, and evidence-ranked COVER score 16/13/14. Source visibility and reader budgets differ from the later complete-source results of 47 and 40, preventing direct attribution of those differences to the memory method. These development-specific versions remain separate and do not form a confirmed six-dataset winning method.

**Granularity-matched source spans.** The Qwen7 follow-up reuses the compilation cache and all 500 questions. Windows contain 256 tokens at stride 192. RAW spans rank all windows by BM25; event spans use only RAW's ten selected events. Typed spans project learned rankings onto source quotations, retaining opaque source text but excluding 43 records without a literal match. Hybrid fuses raw and typed rankings equally at offset 60. All four use nonoverlapping greedy packing and a 6,200-token cap including headers. All 2,000 prepared contexts reached the reader without further truncation or decoding changes.

The fixed Qwen7 reader and judge retain their original prompts and decoding budgets. Maximum serving batch size is 64 rather than the earlier 16; runtime batching is therefore an additional difference from the original RAW/COVER references. All 2,000 answers and grades complete without errors. Equal-weight dataset means and within-dataset whole-bank bootstrap use 20,000 draws. The two planned primary contrasts use 97.5% percentile intervals: hybrid spans minus RAW spans is $-8.80$ $[-12.40, -5.20]$ points; RAW spans minus original RAW is $+4.20$ $[0.59, 8.00]$. Event spans score 49.6%, only $0.60$ points below all-event RAW spans (secondary 95% interval for RAW minus event spans: $[-1.39, 2.60]$). These fixed-Qwen7 results remain outcome-aware development on the same questions; they do not establish a causal decomposition, model transfer, or learned-memory superiority.

RAW and hybrid both answer 183 questions correctly and 225 incorrectly; 68 are RAW-only correct and 24 hybrid-only correct. Hybrid introduces 5,804 span–question occurrences absent from RAW's selection: 2,455 (42.3%) have projection support only from positive-score opaque records, and 85 (1.5%) only from zero-score records. Opaque projection can promote source without an accepted typed fact. In 22 of the 68 RAW-only-correct cases, a reference string appears in RAW's input but not hybrid's. These descriptive counts neither measure semantic recall nor identify either projection rule's causal effect.

**Does stricter projection eligibility repair the deficit?** Table 8 tests three modifications against fresh RAW and unfiltered hybrid controls: exclude opaque projection, require strictly positive BM25 query scores for projected records, or apply both. The index, BM25 statistics, stored records and sources remain fixed; every source span remains eligible through RAW. Only projection eligibility changes, and positive lexical overlap does not imply semantic relevance. Windowing, equal fusion

Table 7: Post-hoc decomposition of the fresh 500-question comparison. Presence refers to a literal reference string in the actual reader input; counts are questions or correct answers, not percentages. Every question is included. Net is hybrid minus RAW.

| RAW presence | Hybrid presence | Questions | RAW correct | Hybrid correct | Net |
|---|---|---|---|---|---|
| Present | Absent | 45 | 25 | 3 | −22 |
| Present | Present | 214 | 116 | 110 | −6 |
| Absent | Absent | 234 | 108 | 94 | −14 |
| Absent | Present | 7 | 0 | 1 | +1 |
| All | | 500 | 249 | 208 | −41 |

Table 8: Fresh projection-filter accuracy (%), 100 questions per dataset. "No opaque" excludes opaque projection; "Positive" requires positive lexical scores; "Both" combines them. Every column is from the same run.

| Dataset | RAW spans | Hybrid | No opaque | Positive | Both |
|---|---|---|---|---|---|
| LongMemEval | **43** | 32 | 26 | 31 | 22 |
| LoCoMo | **52** | 45 | 40 | 42 | 39 |
| FactConsolidation | **19** | 15 | 14 | 17 | 17 |
| EventQA | **73** | 63 | 64 | 63 | 64 |
| RULER | **62** | 53 | 44 | 53 | 44 |
| Mean | **49.8** | 41.6 | 37.6 | 41.2 | 37.2 |

at offset 60, nonoverlapping packing, and the 6,200-token budget are unchanged. Before fresh inference, we verified that both controls reproduced the earlier reader inputs byte for byte.

All 2,500 answers were generated and graded successfully, using 1,576 distinct prompts and 924 instances of identical-prompt output reuse. Reader-visible contexts matched the prepared text without further truncation. Fresh RAW and hybrid score 49.8% and 41.6%, against 50.2% and 41.4% in the earlier run. Excluding opaque projection scores 37.6%; requiring positive lexical scores yields 41.2%; both filters yield 37.2%. The two predefined primary contrasts use 20,000 whole-bank bootstrap draws and 97.5% percentile intervals: both filters minus fresh RAW is $-12.60$ points ($[-16.80, -8.40]$), and minus fresh hybrid is $-4.40$ ($[-7.59, -1.19]$). The corresponding secondary token-F1 differences are $-2.79$ ($[-3.66, -1.95]$) and $-0.21$ ($[-0.95, 0.55]$), with pointwise 95% intervals. Exact match is 15.2%, 13.4%, 13.6%, 13.4%, and 13.6% in the table's column order; mean token F1 is 25.43%, 22.85%, 22.69%, 22.80%, and 22.64%.

Removing opaque projection reduces rather than improves judged accuracy on these questions. The result does not support filtering failed extractions as a remedy for the retrieval deficit. RAW still wins in all five datasets, and the utility of a source record cannot be inferred from its extraction status alone. The same-question development design, fixed Qwen7 models, local judge, and few EventQA/RULER banks limit generalization and interval reliability.

**Stored source versus reader-visible source.** Table 7 partitions all 500 questions from the *fresh* RAW/hybrid pair by literal reference-string presence. Hybrid answers 208 correctly versus RAW's 249. Losing string presence coincides with 22 of these 41 net losses. In this subgroup, hybrid inputs retain on average 41.9% of RAW's selected source-character intervals, measured by unioned interval overlap within each source event. The two-filter variant loses string presence on 57 questions, with 36 correct RAW answers versus two filtered answers. These post-hoc associations are not causal attribution: literal presence and character overlap do not measure semantic relevance, and RAW answers 108 questions correctly when both inputs lack the string. The findings distinguish stored source from budgeted reader input, consistent with Section 4's counterexample. The residual readout below tests an intervention that protects RAW's context and independent answer.

**Protecting RAW before adding learned evidence.** The residual readout keeps the original 6,200-token RAW context and its independent answer. Learned, source-bound records retrieve additional passages that do not overlap RAW's selection; these have a separate 6,200-token budget. The sec-

Table 9: Official consistency mechanisms on small, separate diagnostic workloads. We sample 100 orders per workload; the four-message case covers all 24 distinct orders. State counts are within-row comparisons, not a ranking across systems. No answer accuracy is measured here.

| System / workload | Inputs | Unique orders | Distinct states |
|---|---|---|---|
| StateFuse / fixed operations | 16 | 100 | 1 |
| MELD / running example | 4 | 24 | 1 |
| MELD / semantic-merge example | 16 | 100 | **95** |

Table 10: Protected-RAW residual readout: per-dataset local-judge accuracy and aggregate token F1 (%), 100 questions per dataset. Bold marks the best value, including ties, not significance. The two residual conditions use matched extra context and two calls, sharing the initial RAW answer.

| Dataset | RAW once | RAW residual | COVER residual | Long RAW |
|---|---|---|---|---|
| LongMemEval | **43** | **43** | **43** | 41 |
| LoCoMo | 52 | 50 | **53** | 51 |
| FactConsolidation | 19 | 19 | 19 | **20** |
| EventQA | **73** | 71 | 70 | 72 |
| RULER | **62** | **62** | **62** | 58 |
| Mean | **49.8** | 49.0 | 49.4 | 48.4 |
| Mean token F1 | 25.43 | 25.20 | 25.09 | **25.62** |

ond reader either keeps the original answer or proposes a revision supported by one to four literal quotations, at least one of which is taken from the added evidence. Programmatic checks bind every quotation to a source event and to passage bounds; a proposal that fails those checks is discarded and the original answer stands. These checks establish provenance, not entailment. The matched two-call RAW control uses the same prompts and checks but selects its supplement by BM25. A third control reads the longer RAW context once. All four conditions use Qwen7; the two-call conditions have a 768-token revision cap, and the one-call readers a 192-token cap. The original RAW answers are shared paired inputs. All 500 questions and 125 banks are retained.

Table 10 reports the complete results. COVER's F1 is 25.09%, versus 25.20% for two-call RAW, 25.43% for RAW once, and 25.62% for long RAW. The primary COVER minus two-call RAW interval uses 20,000 whole-bank bootstrap draws: $-0.11$ $[-0.73, 0.49]$ points. COVER accepts 16 revisions and two-call RAW 12. Relative to RAW once, COVER corrects three judged errors and introduces five; the other eight accepted changes do not change judged correctness. The original answer was kept after reader timeouts in two COVER calls and five two-call RAW calls. All 2,000 judge calls completed without errors.

Analysis of failed revisions identifies 149 proposals with invalid JSON and 239 with unbound citations. Among the latter, 136 have every quotation somewhere in the displayed evidence, 102 contain an absent quotation, and one has a malformed later citation. These failures motivate simpler source references, but the diagnosis adds no successful answers. The two-call readers consume approximately 9.2 million logical tokens, compared with 3.12 million for single-call RAW; COVER also reuses 7,206 compiler outputs. On these reused development banks, neither net utility, a cost advantage, nor noninferiority is established.

# B Proof details

## B.1 Formal causal trace and compression statements

**Legal causal traces.** Let $\mathcal{J}(P)$ be the down-closed ideals of the finite event poset $P$. A retry trace is legal when deleting every occurrence after the first leaves a linear extension of its support $A \in \mathcal{J}(P)$. Let a deterministic online memory have initial state $s_0$, event transitions $T_e$, and decoder $d(s, q)$. We compare support-tagged configurations with the same ideal modulo $\approx_{P,Q}$: equality of answers after every legal continuation from that ideal and query in $Q$. Write $s_A$ for an arbitrary reachable configuration whose support is $A$; every local law below is quantified over every such

configuration, not one chosen representative. The memory is $P$-schedule-and-retry robust when all legal retry traces with the same support end in equivalent configurations.

**Theorem 1 (causal trace characterization).** $P$-schedule-and-retry robustness holds if and only if, modulo $\approx_{P,Q}$, both local laws hold on every reachable ideal $A$:

$$e \in A: \quad T_e(s_A) \approx_{P,Q} s_A \qquad \text{(retry absorption)},$$
$$e, f \notin A,\ A \cup \{e\}, A \cup \{f\} \in \mathcal{J}(P): \quad T_f T_e(s_A) \approx_{P,Q} T_e T_f(s_A) \qquad \text{(independent swap)}.$$

The second law concerns only jointly enabled, hence incomparable, events. Consequently terminal behavior factors through ideal support. Theorem 2 gives the corresponding future-compatible ACI quotient of the ideal semilattice. The unconstrained set-support theorem is the special case in which $P$ is an antichain. The proof uses retry deletion and connectivity of linear extensions by adjacent incomparable swaps (Mazurkiewicz, 1995).

**Theorem 2 (minimal causal future-query state).** Let $\mathcal{H}_P = (\mathcal{J}(P), \cup)$ and required answer behavior be $a : \mathcal{J}(P) \times Q \to Y$. Define, for $A, B \in \mathcal{J}(P)$,

$$A \equiv_{P,Q} B \iff \forall K \in \mathcal{J}(P),\ \forall q \in Q:\ a(A \cup K, q) = a(B \cup K, q).$$

Define a behaviorally exact future-compatible ACI memory as $(M, \sqcup, \mu, d)$ whose reachable representation satisfies $\mu(A \cup B) = \mu(A) \sqcup \mu(B)$ and $d(\mu(A), q) = a(A, q)$. Then $\equiv_{P,Q}$ is a join congruence. Its quotient is behaviorally exact and refined by every such memory on reachable ideals: it is the unique smallest reachable state space, up to isomorphism, within this ACI class for fixed $(P, E, Q, a)$.

Behavioral exactness can preserve factually wrong answers. Finite benchmark queries cannot certify it for all future agent queries.

This is a future-context Myhill–Nerode/syntactic-semilattice quotient (Myhill, 1957; Nerode, 1958; Adámek et al., 2018); related policy/probe-relative quotients appear in decision memory and bounded interaction (Zou et al., 2026; Nixon, 2026). Exact-ID enforces retry absorption, not independent-event commutation. Full source-set union is both robust and information-preserving; compression, evidence loss, and reading can separately reduce utility.

**Consequences.** Within this ACI class, $N_Q$ reachable classes require $\lceil \log_2 N_Q \rceil$ bits. Membership queries over an enabled $n$-event antichain distinguish all $2^n$ extensions, requiring $n$ bits before payloads. Retaining $C < n$ raw events without a sufficient summary cannot be universally exact, even without retries. A smaller semantic quotient creates room for exact compression.

**Implications for compilation and execution.** The Appendix derives four consequences: indistinguishable worlds with $k_j$ winners per key require $\lceil \log_2 \prod_j k_j \rceil$ provenance bits (Theorem 3); byte-exact convergence of compile-once event maps requires invariance to incomparable events (Theorem 4); exact reading requires preserving future-query distinctions (Theorem 5). The same equivalence preserves every deterministic read-only-query executor after every common legal continuation (Theorem 6).

If $c = p \circ c'$, decoder composition gives $R^*(c') \leq R^*(c)$ for every task/loss. Prompt truncation breaks this order: risk can rise from zero to $1/2$ at every budget $B$. Canonical raw therefore receives its own fallback budget. Invariant selection preserves robustness; an optimal selector allowing constant choices can match the better reader, but fixed-selector utility is empirical. The Appendix checks 187,648 refinements/tasks and reversals for $B = 1, \ldots, 128$.

**Corollary (bounded exact-ID dichotomy).** For $n > C$ jointly enabled, exactly-once events, retaining $C$ raw identities without a sufficient aggregate forces a choice. Order-dependent retention makes some membership answer schedule-dependent; order-independent retention omits some answer uniformly. FIFO and canonical top-$C$ respectively realize these branches in the all-500 test.

With $p = C/n$, uniform permutations give FIFO membership disagreement $2p(1-p)$ and $m$-schedule flip probability $1 - p^m - (1-p)^m$. Over all $n!$ permutations, uniform-query mean/all-schedule utility is $p/0$ for FIFO and $p/p$ for fixed top-$C$. For $m$ independent sampled permutations, FIFO's expected all-sampled-schedule utility is $p^m$ (Appendix).

The algebra is classical; its role here is to characterize the information that learned memory must preserve and the capacity limit of exact-ID retention.

### B.2 PROOFS

**Theorem 1.** For a legal retry trace $w = e_1 \cdots e_k$, let $\rho(w)$ delete every occurrence of an event after its first and let $\mathrm{supp}(w) = \{e_1, \ldots, e_k\}$. By definition, $\rho(w)$ is a linear extension of the ideal $\mathrm{supp}(w)$. For configurations with the same support $A$, future equivalence $\approx_{P,Q}$ means equality after every word that is a legal continuation from $A$ and every $q \in Q$. It is a right congruence along enabled events.

Necessity is local. Let any legal retry trace reaching the particular configuration $s_A$ be fixed. If $e \in A$, append one retry of $e$; robustness of the two same-support traces gives $T_e(s_A) \approx_{P,Q} s_A$. If $e$ and $f$ are jointly enabled after $A$, then they are incomparable and appending either $ef$ or $fe$ to that same trace is legal. Robustness gives $T_f T_e(s_A) \approx_{P,Q} T_e T_f(s_A)$.

For sufficiency, delete repeated deliveries one at a time at their positions; the prefix support already contains the repeated event, so retry absorption preserves the equivalence class, and the right-congruence property carries that equivalence through the untouched suffix. The remaining trace is a linear extension. Any two linear extensions of a finite poset are connected by adjacent swaps of incomparable elements. At each such swap, the common prefix is an ideal and the two events are jointly enabled, so the independent-swap law preserves the equivalence class. Therefore all legal retry traces with one support are equivalent.

Define $\mu(A)$ as that support-tagged class for any linear extension of $A$. It is well defined by the preceding argument, so terminal behavior factors through ideal support. Theorem 2 constructs the future-compatible join quotient of those ideals. When $P$ is an antichain, $\mathcal{J}(P) = 2^E$ and all first deliveries are jointly enabled, recovering the unrestricted set-support result.

**Theorem 2.** Suppose $A \equiv_{P,Q} B$. For arbitrary ideals $D, K \in \mathcal{J}(P)$ and $q \in Q$, choose the ideal continuation $D \cup K$:

$$a((A \cup D) \cup K, q) = a(A \cup (D \cup K), q) = a(B \cup (D \cup K), q) = a((B \cup D) \cup K, q).$$

The union of ideals is an ideal, so $A \cup D \equiv_{P,Q} B \cup D$ and quotient union is well defined. The decoder $d([A], q) = a(A, q)$ is well defined by setting $K = \varnothing$, proving existence. For minimality, let a behaviorally exact ACI representation on $\mathcal{J}(P)$ satisfy $\mu(A) = \mu(B)$. For every ideal $K$,

$$\mu(A \cup K) = \mu(A) \sqcup \mu(K) = \mu(B) \sqcup \mu(K) = \mu(B \cup K).$$

Exact behavioral decoding gives equal answers for all $q$, hence $A \equiv_{P,Q} B$. Distinct quotient classes therefore require distinct reachable memory states, and the quotient construction attains the bound.

The executable checker independently enumerates every transitively closed four-event relation compatible with one fixed topological labeling (40; every four-event poset admits such a relabeling), every assignment of the four two-state transition functions to its events, and both initial states. Across 20,480 transducer cases and 275,744 reachable support–state configurations, global same-support robustness and the two local laws agree exactly (2,480 positive cases each, zero mismatches). Separately, all 104,288 Boolean answer behaviors on the corresponding ideal lattices induce well-defined ACI joins; 8,372 distinct quotient algebras confirm the same correspondence. These finite checks target specification and proof-transcription errors; the general result follows from the preceding argument.

**Theorem 3.** For key $j$, let its $k_j$ distinct updates have any chosen one as the unique maximum of a realizable per-key total order. The disjoint union over keys realizes every one of the $W = \prod_j k_j$ winner vectors while leaving the unordered payload set fixed. A deterministic invariant reader sees the same input in all worlds, hence emits one vector and is exact in at most one. A randomized reader is a mixture of deterministic vectors, each with uniform-world success $1/W$, so its average is also at most $1/W$. If a certificate maps two different winner vectors to one message, their certified reader inputs coincide although their required answers differ. An exact scheme therefore needs at least $W$ messages and $\lceil \log_2 W \rceil$ bits. Returning all maximal candidates avoids a false assertion but leaves an ambiguous key unresolved.

**Theorem 4.** This statement concerns byte-exact per-event maps, with an injective event key and no rewriting. Each emitter is a deterministic function of the event and its preceding immutable-record map, without access to delivery order or serialization order. Let $r_e(H)$ denote its output on such a map $H$. For a reachable prefix map $H$ at which $e, f$ are jointly enabled, suppose $r_e(H) \neq r_e(H \cup \{f\})$. Extend the two prefixes to linear extensions with $e$ respectively before and after $f$. The runs emit different immutable bytes at the injective key for $e$; later ACI admission cannot erase the difference. Conversely, require every emitter to be invariant to inserting or removing incomparable records in any compatible, predecessor-closed input map. Remove those records in reverse topological order. Induction on event height fixes the remaining ancestor records, so the emitter sees the same ancestor map and immutable source in every schedule. The emitted per-event maps and their ACI unions therefore agree. A projection that erases the differing bytes weakens the guarantee to an observational quotient and may lose query utility.

**Bounded exact-ID dichotomy.** Let $R_\pi \subseteq E$ be the at-most-$C$ identities retained after permutation $\pi$. If $R_\pi \neq R_\sigma$, choose $e \in R_\pi \triangle R_\sigma$; the membership answer for $e$ differs. If all $R_\pi$ are equal to one set $R$, then $|R| \leq C < n$, so any $e \in E \setminus R$ is absent despite occurring. The schedules contain each identity once, hence a retry filter is never invoked.

**Bounded FIFO schedule law.** For a fixed $C$-subset $S$, exactly $C!(n-C)!$ permutations place precisely $S$ in the last $C$ positions. Thus FIFO's retained set is uniform over $\binom{n}{C}$ possibilities. Two independent draws agree in state with probability $\binom{n}{C}^{-1}$. A fixed event is retained with probability $p = C/n$, so two membership answers differ with probability $2p(1-p)$ and $m$ answers contain both truth values with probability $1 - p^m - (1-p)^m$. Because $C < n$, every event is omitted by some legal permutation, giving zero all-schedule point-query utility. A fixed-priority top-$C$ join retains one invariant $C$-subset, hence has uniform-query mean and all-schedule utility $C/n$. The executable check enumerates all 46,232 permutations, 2,493,928 membership rows, and 56,944 top-$C$ ACI state triples through $n = 8$ (ACI triples through $n = 5$).

**Theorem 5.** If $c(A) = c(B)$, the homomorphism law gives $c(A \cup K) = c(A) \sqcup c(K) = c(B) \sqcup c(K) = c(B \cup K)$ for every ideal $K$. A behaviorally exact reader therefore implies $a(A \cup K, q) = a(B \cup K, q)$ for every $K, q$, hence $K_c \subseteq \equiv_{P,Q}$. Conversely, under that containment set $d(c(A), q) = a(A, q)$. Equality of compiler states implies $A \equiv_{P,Q} B$, so the decoder is well defined; the homomorphism law gives the same statement after every legal future union. An opaque refinement augments rather than erases source identity, so its new kernel $K'$ satisfies $K' \subseteq K_c$ and cannot add a false merge. This conclusion is conditional on retaining the rejected source; dropping it is a merge, and evidence-span validation alone does not establish semantic correctness of accepted records. The executable checker enumerates all 4,140 kernels over the eight histories of three events, filters 61 join congruences, and verifies the equivalence on all 15,616 semantics–representation pairs in the three-event antichain specialization. Fact-local totalization is the product of the opaque source state and admitted valid facts; projection recovers the former, so its kernel is a subset of the whole-chunk opaque kernel and its joins remain ACI. For the utility statement, let $c = p \circ c'$ and let $d$ be any decoder on $c$. Then $d' = d \circ p$ on $c'$ induces exactly the same action and loss for every history–query pair. Taking the infimum over decoders on each side gives $R^*(c') \leq R^*(c)$ for any distribution and loss. This Blackwell-style information-order result (Blackwell, 1953) concerns optimal decoders; implementation by a particular finite-capacity neural reader is an empirical question. The executable check enumerates 733 ordered refinement pairs among 61 join congruences and all 256 Boolean tasks.

**Proposition 1 (budgeted refinement reversal and selector envelope).** For any token budget $B \geq 1$, let the binary history $h$ be uniform and let $c(h)$ serialize as the one token $h$. Define the refined state $c'(h) = (c(h), z^B)$, with projection onto its first component. Before serialization, $c'$ therefore refines $c$. Suffix-$B$ truncation preserves $c(h)$ but maps both serialized $c'(h)$ states to $z^B$; under 0–1 loss their Bayes risks are zero and $1/2$. Thus Blackwell dominance need not survive a non-information-monotone resource operator. Conversely, let both separately budgeted component decoders and their public signals be schedule invariant. Any deterministic selector over those signals is then schedule invariant. If the selector class contains *always coarse* and *always fine*, its optimal risk is at most the smaller component risk for arbitrary loss because both constant policies are feasible. A particular

public-signal selector need not attain that optimum. The checker verifies this strict reversal and envelope for 128 budgets.

**Proposition 2 (literal readout and the limit of source preservation).** Let $S$ be the terminal source-and-record map of a confluent memory, with a fixed source-visibility operator applied before compilation and retrieval. For a public query $q$, enumerate source intervals by immutable identities, rank them with fixed functions of $(S, q)$, resolve ties by those identities, and pack them under a fixed budget with a deterministic procedure. Render each selected interval literally, together with its source identity and offsets. Then the resulting context $C(S, q)$ is schedule invariant, and every rendered evidence interval is a substring of an authorized visible source. The answer is invariant if the reader is also a deterministic function of this context and the public query.

For any two such context functions $C, C'$ and the *same* fixed reader $f$, define $R(C) = \mathbb{E}[\ell(f(q, C), y)]$, where $0 \le \ell \le 1$. Under a common distribution of histories, queries, and targets,

$$|R(C') - R(C)| \le \Pr\{C' \ne C\}. \tag{2}$$

This bound is tight, even when both contexts contain only literal evidence from the same visible source. No smaller distribution-free bound follows solely from source retention or schedule invariance.

*Proof.* Confluence fixes $S$. Every subsequent operation is a deterministic function of $S$ and $q$, including selection and serialization, so contexts agree across schedules. Literal extraction establishes substring provenance; composition with a deterministic reader establishes answer agreement. For the risk bound, losses coincide whenever $C' = C$ and differ in absolute value by at most one otherwise. Taking expectations proves Equation 2. For tightness, let a visible source contain two distinct one-token intervals $u, v$, with target zero. Let the fixed reader output zero on $u$ and one on $v$. Use $u$ in both contexts except on histories of probability $p$, where $C'$ selects $v$. Both views are literal and invariant, but their risks differ by $p$.

The premises require fixed numerical scores and tie handling, not merely an embedding model name. An API model alias or temperature zero alone does not establish a fixed deterministic reader. Substring provenance certifies where text came from, not its truth, completeness, or relevance. Likewise, adding verbatim instructions to a prompt can alter model behavior adversely; it is an experimentally testable intervention rather than a semantic guarantee.

**Proposition 3 (simultaneous cross-benchmark superiority).** Fix $M$ candidate readers, $D$ datasets, and $K$ baselines *before* observing fresh calibration outcomes. For dataset $d$, let the independent units be $n_d$ source-history clusters. Questions within a cluster may be arbitrarily dependent. Let $X_i^{m,d,k} \in [-1, 1]$ be candidate-minus-baseline loss averaged within cluster $i$, with the within-cluster weighting specified in advance. Fix nonnegative cluster weights $w_{di}$ with $\sum_i w_{di} = 1$.

$$\widehat{\Delta}_{mdk} = \sum_{i=1}^{n_d} w_{di} X_i^{m,d,k}, \qquad \Delta_{mdk} = \sum_{i=1}^{n_d} w_{di} \mathbb{E} X_i^{m,d,k},$$
$$\varepsilon_d = \sqrt{2\left(\sum_i w_{di}^2\right) \log \frac{MDK}{\delta}}, \qquad U_{mdk} = \widehat{\Delta}_{mdk} + \varepsilon_d. \tag{3}$$

For $0 < \delta < 1$, with probability at least $1 - \delta$, every one of the $MDK$ inequalities $\Delta_{mdk} \le U_{mdk}$ holds simultaneously. Consequently, selecting any candidate $m$ for which $U_{mdk} < 0$ for *all* $(d, k)$ certifies lower expected loss against every baseline on every dataset at this joint confidence level. Selection among the fixed $M$ candidates may use the calibration outcomes. The statement does not cover additional candidates tuned using those same outcomes.

*Proof.* For a fixed $(m, d, k)$, independent summands $w_{di} X_i^{m,d,k}$ have interval widths $2w_{di}$. Weighted Hoeffding concentration (Hoeffding, 1963) therefore gives

$$\Pr\{\Delta_{mdk} - \widehat{\Delta}_{mdk} > t\} \le \exp\left(-\frac{t^2}{2\sum_i w_{di}^2}\right).$$

Substitution of $\varepsilon_d$ makes this probability at most $\delta/(MDK)$. A union bound proves simultaneous validity; no independence between candidates, baselines, or datasets is required. On that event, any candidate passing every upper bound has all its expected loss differences strictly negative.

This is an application of classical concentration and multiple testing, related to calibration-based risk control (Angelopoulos et al., 2021), not a new concentration inequality. Equal cluster weights estimate macro-average history risk. Fixed unequal weights estimate the stated weighted target; they must not be silently substituted for a different benchmark population. The effective sample size is $n_{\mathrm{eff},d} = 1/\sum_i w_{di}^2$. For a target margin $\gamma$, even a measured improvement of $\gamma$ requires $n_{\mathrm{eff},d} > 2\log(MDK/\delta)/\gamma^2$ to pass this conservative bound. More questions from the same history do not create more independent histories. In particular, a benchmark subset with only two source histories cannot support this certificate merely by adding many questions.

Our repeatedly inspected development subsets do not satisfy fresh calibration, and we do not apply this certificate to them. Their point estimates identify promising or failing mechanisms; they do not establish joint population-level superiority across six datasets and four systems.

**Theorem 6.** For configurations with the same causal support, future-query equivalence is precisely the relation preserving every deterministic read-only-query executor after every common legal continuation. Fix a common continuation $w$ and let the executor choose each query from its transcript, with an identical initial executor state in both runs. If $s \approx_{P,Q} t$, the first answers after $w$ agree; determinism gives the same next query, and induction gives the same transcript and external action whenever the executor terminates. If $s \not\approx_{P,Q} t$, by definition some legal continuation $w$ and query $q$ give distinct answers $y, y'$. Apply $w$ to both memories before execution. A read-only executor that asks $q$ once and emits distinct actions $a_y, a_{y'}$ then separates them. Assigning loss $L$ to one action and zero to the other realizes a gap $L$ for any $L > 0$. This result characterizes an interface-level possibility; real-world loss requires an external deployment model. The executable checker enumerates all 256 Boolean semantics over eight three-event histories, all ordered history pairs, all future unions, and all Boolean-to-binary-action maps. It verifies the equivalence in 16,384 pairs; 4,740 separated pairs agree on the current answer, showing why one-shot accuracy cannot certify safe composition.

**Schedule-risk identity.** For a root with $k$ correct schedules, pair counting gives $D_i^Y = k(m-k)/\binom{m}{2}$. Its guarantee gap is zero for $k \in \{0, m\}$ and $k/m$ otherwise. For $0 < k < m$, use $1 \le m-k \le m-1$ to obtain $D_i^Y/2 \le \Gamma_i \le (m-1)D_i^Y/2$; averaging proves Equation 1. The same calculation gives $F/m \le \Gamma \le (m-1)F/m$ for correctness-flip fraction $F$. Exhaustion over all $(m, k)$ through $m = 64$ checks 2,142 cases and the equality witnesses.

**COVER instantiation.** Let $A_e = \mathrm{Anc}_P(e)$ and $r_e = g_\Theta(e, V(A_e))$. Induct on event height. Root records are schedule independent because they use empty ancestor views. If all lower-height records agree, $V(A_e)$ agrees, hence $r_e$ agrees. Fair saturation therefore produces the same complete record map. Deterministic ACI materialization and a behavior-only reader give identical terminal views and answers. Child-first events remain inactive until closure, after which they compile from the same exact ancestors.

Causal confluence and semantic correctness are independent: a constant writer is confluent and generally wrong. If two possible worlds yield identical observable evidence but require different records, no evidence-only compiler can be correct in both; explicit conflict or opacity is necessary unless an additional trusted assumption is introduced.

**Declared-parent stream extension.** Advance knowledge of a finite terminal event set is unnecessary. Let events form an at most countable set, each with an immutable authenticated payload, finitely many declared parents, and a natural-number rank strictly exceeding every parent's rank. Identities bind these fields. The fixed total compiler $g_\Theta$ reads only the event and its exact ancestor records. For a finite received set $S$, let $D(S)$ be its greatest parent-closed subset. A materializer buffers events until their parents activate, then compiles them once and fairly saturates. Its active support is exactly $D(S)$, and its record map depends only on $S$, even with child-first delivery and retries. Future arrivals never revise an active record.

Table 11: Core schedule-robustness results. Native entries count system–model cells with a positive root-level rate; COVER entries count affected roots. Bold highlights native prevalence and exact COVER convergence; failures remain in the denominator.

| System | Evaluation scale | State fork | Answer fork | Correctness flip | Accuracy | Coverage |
|---|---|---|---|---|---|---|
| Four native writers | 3 readers × 4 writers | **11/12** | **11/12** | **10/12** | – | – |
| COVER | 1,000 roots × 6 schedules | **0/1,000** | **0/1,000** | **0/1,000** | **99.3%** | **99.3%** |

*Proof.* Finite branching and decreasing natural rank give each event a finite ancestor closure. Rank induction defines a unique $r_e = g_\Theta(e, V(\mathrm{Anc}(e)))$. Activation includes only events in $D(S)$, because every activated ancestor has been received. Conversely, fair saturation activates every member of $D(S)$ by rank induction. Each compilation uses the unique lower-rank records, so it produces $r_e$ regardless of arrival order. Every ancestor is fixed before activation; future descendants and incomparable events cannot alter $r_e$. Retained records absorb retries.

If every event is eventually delivered and eligible work is processed fairly, each event activates in finite time: its finite ancestor closure has a finite maximum delivery time. Replicas then agree pointwise, and deterministic readers over the same finite ideal agree. Answers over unequal, ever-growing sets need not stabilize. This is a classical well-founded extension, not a new confluence principle. Missing parents can block indefinitely; false declarations can give stable but wrong outputs. Purity and a shared, fixed compiler identity and configuration are required; local caching alone does not ensure them. The compile-once count is per replica.

A CPU implementation receives no complete graph or future event catalog and is checked against an independent finite reference. The test covers 64 four-vertex DAGs under a fixed topological labeling and their 24 delivery orders: 1,536 traces, 6,144 received-prefix comparisons, and 6,144 retry/no-recompilation checks. A deterministic witness compiler tests structural behavior independently of LLM task accuracy. The theorem assumes binding identities; the implementation relies on SHA-256 collision resistance, which does not guarantee an injective finite digest over an infinite universe.

**Canonical suffix-replay cost.** Fix canonical ranks $1 < \cdots < n$ and arrival permutation $\pi$. Consider the algorithm that unconditionally replays the affected suffix after insertion. Every event is applied once. For each pair $i < j$, event $j$ is replayed exactly when it arrived before $i$, because the algorithm rebuilds the suffix containing $j$. Hence its calls equal $n + I(\pi)$. Every pair is inverted with probability $1/2$ under a uniform permutation, giving expectation $n + n(n-1)/4$; reverse order contains all $n(n-1)/2$ inversions and costs $n(n+1)/2$. A writer that exposes reusable outputs, unchanged intermediate states, or commutative patches can permit fewer calls. The expression is an algorithm-specific count, not a universal black-box lower bound.

# C SYNTHETIC AGGREGATE

Table 12: Synthetic results. Native coverage and COVER's formally defined semantic coverage use different criteria and are reported separately.

| System | Roots | Accuracy | Coverage | $D$ | Flip |
|---|---|---|---|---|---|
| Native Mem0 | 24 | 38.19 | – | 21.39 | 25.00 |
| A-MEM | 24 | 92.36 | – | 2.50 | 4.17 |
| MemOS | 24 | 79.17 | – | 1.39 | 0 |
| MemoryOS | 24 | 93.06 | – | 11.94 | 29.17 |
| COVER | 24 | 100.00 | 100.00 | 0 | 0 |
| COVER | 173 | 100.00 | 100.00 | 0 | 0 |
| COVER | 1,000 | 99.30 | 99.30 | 0 | 0 |

Table 13: Cross-model native evaluation on the identical 24 roots and six legal schedules per root. All entries are percentages. $D$ is answer disagreement; Flip is the fraction of roots whose correctness changes; State $D$ is exact semantic-state disagreement; Replay $D$ uses identical chronological replays. Parse failures remain in primary denominators.

| Model | System | Acc. | Answer D | Flip | State D | Replay D | Parse fail |
|---|---|---|---|---|---|---|---|
| Qwen-7B | Mem0 | 38.2 | 21.4 | 25.0 | 100.0 | 8.3 | 0.0 |
| Qwen-7B | A-MEM | 92.4 | 2.5 | 4.2 | 100.0 | 0.0 | 0.0 |
| Qwen-7B | MemOS | 79.2 | 1.4 | 0.0 | 62.2 | 0.0 | 0.0 |
| Qwen-7B | MemoryOS | 93.1 | 11.9 | 29.2 | 88.1 | 0.0 | 0.0 |
| Llama-3B | Mem0 | 22.9 | 10.0 | 16.7 | 0.0 | 16.7 | 0.0 |
| Llama-3B | A-MEM | 84.7 | 10.0 | 25.0 | 100.0 | 0.0 | 0.0 |
| Llama-3B | MemOS | 54.9 | 2.8 | 4.2 | 12.2 | 0.0 | 7.6 |
| Llama-3B | MemoryOS | 94.4 | 10.6 | 29.2 | 90.8 | 0.0 | 0.0 |
| Mistral-7B | Mem0 | 18.1 | 60.3 | 58.3 | 100.0 | 0.0 | 43.8 |
| Mistral-7B | A-MEM | 60.4 | 7.2 | 12.5 | 100.0 | 0.0 | 38.9 |
| Mistral-7B | MemOS | 62.5 | 0.0 | 0.0 | 28.9 | 0.0 | 29.2 |
| Mistral-7B | MemoryOS | 59.0 | 26.1 | 58.3 | 91.1 | 0.0 | 38.9 |

Table 14: All-500 raw complete-history LongMemEval-S control.

| Materialization | Accuracy | Recall any | Recall all | F1 |
|---|---|---|---|---|
| Flat BM25 top-10 | 47.60 | 90.40 | 74.60 | 9.284 |
| COVER canonical | 47.60 | 90.40 | 74.60 | 9.276 |

## D COMPLETE NATIVE CROSS-MODEL EVALUATION

## E RAW-TURN COMPLETE-HISTORY CONTROL

Under the raw-turn control, the paired contexts are byte-identical on 494 of 500 roots. The six differences arise from score ties; five yield different answer strings, but none changes the binary judge decision. Truncation affects 12 of 1,000 reader rows, and no retrieval, reader, or judge request fails. Multi-session and temporal accuracy is 27.82% and 36.09% despite 90.4% answer-session recall, localizing residual error beyond retrieval alone. This control uses raw-turn totalization and excludes learned compilation.

## F BOUNDED-STATE CAPACITY SENSITIVITY

All physical retries are removed before retention. The mean evidence advantage is small at every capacity; the large effect is that arrival FIFO's lucky and unlucky schedules separate, while trusted event-time selection has zero schedule disagreement. Thus the $C = 192$ reader experiment is the predefined behavioral point inside a three-capacity structural/evidence sweep.

## G UNBOUNDED DELIVERY AND RETRY STRESS

Each complete history follows chronological order and five deterministic session permutations. A second regime retries a trace-specific 10% of user rounds once, giving 12,000 contexts with identical logical histories.

At 10%, 6,121/6,671 changed-context pairs change the answer; equal-context pairs cannot differ by protocol. Flat answers disagree on 81.61% of schedule pairs and correctness flips on 102 roots. The 102 roots split almost evenly around COVER's fixed answer, explaining the non-significant mean-accuracy difference. Exact-ID removes every correctness flip. The 10% retry rate is a controlled fault-injection setting; production frequencies remain unmeasured.

## H MECHANISM AND COST DIAGNOSTICS

Removing transport-bound identity reduces accuracy from 99.3% to 97.4%; removing the one-shot pointer repair reduces it to 98.5%. Removing typed entity-reference normalization collapses multi-

Table 15: Duplicate-free bounded-state sensitivity. Recall is retrieval of any official answer session; worst requires retrieval under every legal schedule.

| $C$ | FIFO $D_{ctx}$ | COVER $D_{ctx}$ | FIFO mean | COVER mean | FIFO worst | COVER worst |
|---|---|---|---|---|---|---|
| 128 | 99.93 | 0 | 66.43 | 69.80 | 20.40 | 69.80 |
| 192 | 97.44 | 0 | 81.77 | 83.60 | 54.00 | 83.60 |
| 224 | 74.80 | 0 | 87.27 | 88.80 | 74.00 | 88.80 |

Table 16: Complete-history stress at 10% target retry. $D$ is within-root pairwise disagreement; accuracy uses the matched local judge.

| Delivery | Materialization | Acc. | $D_{ans}$ | Answer-fork roots | Correctness-flip roots | $D_{ctx}$ | $D_{state}$ |
|---|---|---|---|---|---|---|---|
| Reorder | Flat append | 47.60 | 0.24 | 3/500 | 0/500 | 0.36 | 100 |
| Reorder | COVER | 47.60 | 0 | 0/500 | 0/500 | 0 | 0 |
| Reorder + retry | Flat append | 47.90 | 81.61 | 491/500 | 102/500 | 88.95 | 100 |
| Reorder + retry | Exact-ID dedup | 47.60 | 0.24 | 3/500 | 0/500 | 0.36 | 100 |
| Reorder + retry | COVER | 47.60 | 0 | 0/500 | 0/500 | 0 | 0 |

hop accuracy to 0.1%. Dropping opaque records or allowing unsafe passthrough raises accuracy to 100% by removing seven fail-closed abstentions; neither produces an observed unsafe-wrong answer. Arrival-order replacement leaves accuracy unchanged. These results do not establish a general safety benefit from opaque retention. Terminal legal schedules do not identify child-first activation. A separate prefix probe does: full causal activation recovers 42/42 delayed children, whereas a mechanical no-reactivation counterfactual recovers 0/42.

Over 1,000 roots, compilation uses 5,000 initial extractions and 25 selective catalog lookups, an average of 1.005 model calls per event. Generation runs at 3.98 calls/s on one 80GB A100 GPU (4.00 calls/s on the 173-root run). Different internal call accounting prevents direct throughput comparisons with Mem0 and A-MEM. Mean selected-record payload is approximately 7.0KB per synthetic root; full database and index storage were not measured.

A source-only calculation over all 6,000 main-set deliveries quantifies canonical replay after late prefix insertion. The checkpointed algorithm that reexecutes every affected suffix uses a mean of 8.50 writer calls per five-event schedule (median 9, range 5–10), compared with 5.025 observed calls for COVER, a $1.69\times$ ratio. Rebuilding every prefix uses 15 calls; waiting for terminal closure uses 5 calls but buffers the entire root and forfeits incremental materialization. These measurements characterize that algorithm's call overhead; accuracy is evaluated separately.

**Matched cached canonical control.** We additionally implement canonical insertion with a cache keyed by event identity, exact local-input hash, and fixed compiler identity and configuration. This control reuses the same 5,000 learned records and 25 selective repairs as COVER, with a fresh cache for each of the 6,000 schedules. Its record map matches COVER at all 30,000 nonempty prefixes and matches all 6,000 frozen terminal maps. Including selective repairs, both paths require 5.025 first-use compilation lookups per schedule. This post-hoc replay uses frozen outputs, without new model calls or private labels; runtime speed is not compared. The result locates the replay-saving property in the event-local compiler interface; canonical sorting can use that same property.

# I STATISTICAL DETAIL

For iid roots define $H_i$ as any harmful semantic failure and $C_i$ as semantic coverage; all within-root schedules remain dependent. With $k_H = \sum_i H_i$, $k_C = \sum_i C_i$, and a fixed operating point, one-sided Clopper–Pearson bounds at $\delta/2$ each give simultaneous coverage at least $1 - \delta$ by a union bound. Post-hoc threshold, model, evaluator, or subgroup selection invalidates this statement unless paid for through multiplicity or evaluated on new independent roots.

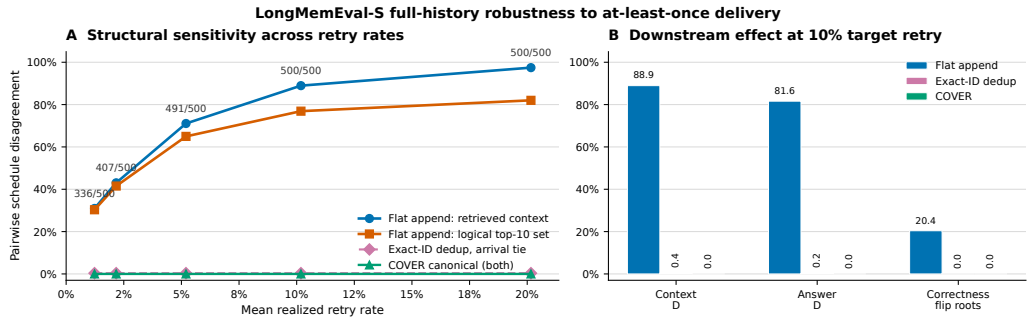


Figure 3: Retry sweep. Exact-ID removes retry amplification; COVER additionally removes residual tie-order and state dependence.

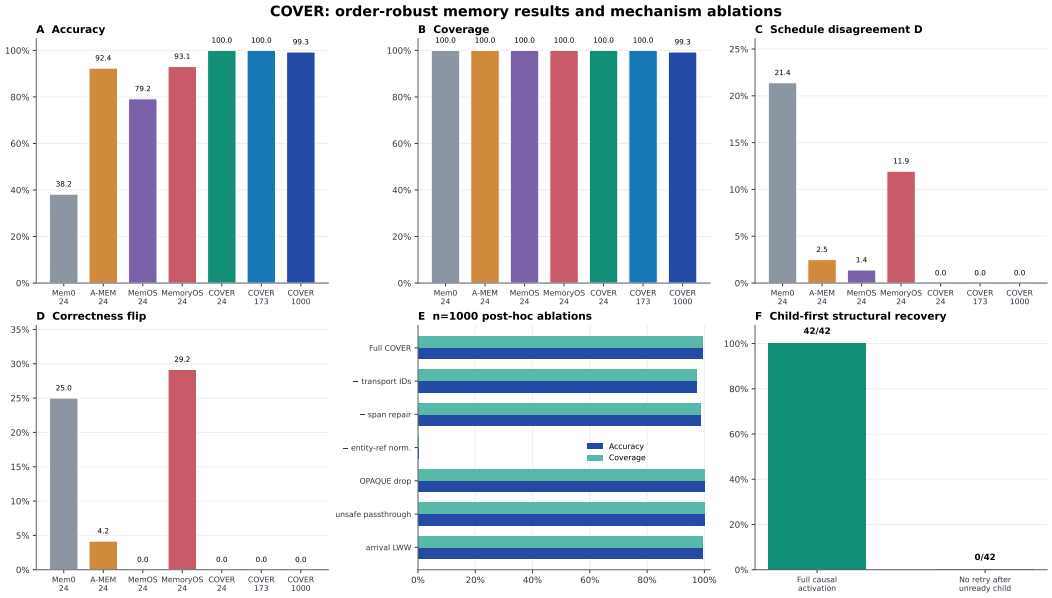


Figure 4: Accuracy, coverage, order disagreement, correctness flips, mechanism ablations, and child-first structural recovery. The 24-root screens and 173/1,000-root COVER studies are separate samples, not paired method rankings. Native coverage denotes parseable non-null answers; COVER uses registered semantic coverage. Ablations reuse frozen outputs, and no-reactivation is a mechanical counterfactual. The absence of unsafe-wrong outcomes does not establish general safety.

## J ADDITIONAL PROTOCOL DETAILS

**Natural dialogue.** The Mixed80 reader uses a fixed 20/80 development/evaluation split. Inputs, predictions, and scores from the 20-root slice determine one instruction that is fixed before private scoring; no subsequent prompt selection is performed. An original-reader aggregate over all 100 roots predates this replay, so the 80-root result is reported as a fixed-protocol replication. Fresh83 provides a root-disjoint evaluation with its root-once protocol, query-only router, literal-span validator, and private scorer fixed before label access. On Mixed80, native Mem0 and COVER raw ACI have exact/F1 of 28.54/33.43% and 30.21/34.50%. Paired COVER−Mem0 differences are $+1.67$ accuracy points (95% CI $[-9.38, 12.29]$), $-14.38$ covered-wrong points ($[-25.63, -3.33]$), $-29.17$ disagreement points ($[-35.33, -22.92]$), and $-33.75$ flip points ($[-45.00, -22.50]$).

**Cross-benchmark utility.** The LoCoMo formal population contains all category-1–4 questions from all ten released conversations; category 5 unanswerable probes are excluded before model

Table 17: FactConsolidation by fixed subset. E/F/J are exact, token F1, and matched-judge accuracy (%). All 100 questions per row are retained.

| Subset | Raw append | | | COVER | | | $\Delta J$ |
|---|---|---|---|---|---|---|---|
| | E | F | J | E | F | J | |
| Multi-hop 6k | 6 | 7.50 | 6 | 32 | 32.33 | 29 | +23 |
| Multi-hop 32k | 9 | 9.93 | 9 | 22 | 22.00 | 22 | +13 |
| Multi-hop 64k | 4 | 6.64 | 4 | 17 | 17.40 | 17 | +13 |
| Multi-hop 262k | 4 | 4.50 | 4 | 6 | 7.00 | 6 | +2 |
| Single-hop 6k | 43 | 43.00 | 38 | 89 | 89.00 | 62 | +24 |
| Single-hop 32k | 58 | 59.17 | 50 | 89 | 89.62 | 72 | +22 |
| Single-hop 64k | 54 | 54.73 | 51 | 93 | 93.00 | 73 | +22 |
| Single-hop 262k | 48 | 50.00 | 45 | 86 | 88.19 | 72 | +27 |

calls. The 272 sessions are compiled once each with Qwen2.5-7B; four sessions (1.47%) are totalized as opaque, and all 4,146 reader rows remain in the denominator. The conversation-cluster bootstrap resamples ten conversations. FactConsolidation uses all 100 questions in each of eight fixed length×hop subsets. Its typed parser covers 51,354/51,354 source records; last-source-ordinal materialization removes 19,852 superseded records. Both conditions use the same iterative BM25 top-24 and Qwen2.5-7B reader. The subset-cluster bootstrap resamples the eight predefined subsets.

**Retry propagation.** Context changes on 6,671/7,500 flat retry pairs, and 6,121 (91.76%) change the normalized answer. Duplicate prompts are evaluated once, so equal-context pairs cannot differ. Of 102 correctness-flip roots, 52 are correct under COVER but mixed under flat schedules and 50 have the reverse fixed correctness. These nearly cancel to nine extra correct flat rows out of 3,000.

**Endpoint relay.** The outcome-blind retail census requires at least one supported official gold write, an explicit order ID on every such write, at least three user orders, and error-free direct gold replay. Two instances (Tasks 64 and 105) are excluded before model execution because their traces require unmodeled dialogue-side state. Adding the three airline and three banking roots yields 80 tasks with 132 authorized write calls in total, including one task requiring no write. The five views are exact-ID arrival FIFO, canonical raw replay, typed arrival FIFO, typed ACI authorizations, and all authorizations in reference order. The matched typed pair uses identical record content, relay prompts, models, and token budgets. Both raw views retain 1,800 tokens; 462/480 contexts per model are truncated. Typed contexts are not truncated. COVER and typed reference order agree on 1,434/1,440 action signatures; the remaining six expose executor order sensitivity. Amounts count each affected reservation/order once, or the posted/per-recurrence banking amount.

## K FULL-HISTORY BREAKDOWN AND MODEL TRANSFER

**Equal-reader-call control.** Raw-2 always rereads the identical canonical raw evidence and verifies its first draft, giving it the same two-call reader limit as the facts/raw cascade. This matches reader-call limits, not write, retrieval, and total token costs; the cascade additionally uses learned compilation. Raw-2 does not retrieve new evidence on its second call.

The cascade remains ahead by 13.45 judge points for Qwen ($[9.03, 18.07]$) and 6.09 for Mistral ($[1.89, 10.29]$), but trails Llama by 7.35 ($[-11.55, -3.36]$). This comparison rules out a larger reader-call limit as the explanation for the Qwen and Mistral gains, but does not isolate representation from retrieval or routing. Raw-2 itself underperforms Raw-1 for these two readers. The Llama contrast shows that the cascade does not transfer uniformly across readers.

Facts alone do not transfer monotonically across readers: they lose 3.36 and 4.62 judge points to raw for Llama and Mistral. The same fixed public-status cascade nevertheless finishes above raw

Table 18: Endpoint relay by executor. Counts are per model; max missed value counts each affected entity once in simulator USD.

| Executor | Memory | Schedule success | All-sched. success | Plan flip | Unsafe root | Max missed value |
|---|---|---|---|---|---|---|
| Llama-3.2-3B | Raw arrival FIFO | 123/480 (25.6%) | 2/80 (2.5%) | 78/80 | 69/80 | $8,864 |
| Llama-3.2-3B | Raw canonical replay | 114/480 (23.8%) | 19/80 (23.8%) | 0/80 | 26/80 | $8,046 |
| Llama-3.2-3B | Typed arrival FIFO | 465/480 (96.9%) | 75/80 (93.8%) | 4/80 | 2/80 | $6,173 |
| Llama-3.2-3B | COVER typed ACI | 474/480 (98.8%) | 79/80 (98.8%) | 0/80 | 1/80 | $672 |
| Llama-3.2-3B | Typed reference order | 468/480 (97.5%) | 78/80 (97.5%) | 0/80 | 2/80 | $6,173 |
| Mistral-7B-v0.3 | Raw arrival FIFO | 117/480 (24.4%) | 4/80 (5.0%) | 76/80 | 69/80 | $9,030 |
| Mistral-7B-v0.3 | Raw canonical replay | 108/480 (22.5%) | 18/80 (22.5%) | 0/80 | 33/80 | $9,030 |
| Mistral-7B-v0.3 | Typed arrival FIFO | 465/480 (96.9%) | 77/80 (96.2%) | 1/80 | 1/80 | $6,403 |
| Mistral-7B-v0.3 | COVER typed ACI | 462/480 (96.2%) | 77/80 (96.2%) | 0/80 | 1/80 | $6,403 |
| Mistral-7B-v0.3 | Typed reference order | 462/480 (96.2%) | 77/80 (96.2%) | 0/80 | 1/80 | $6,403 |
| Qwen2.5-7B | Raw arrival FIFO | 127/480 (26.5%) | 4/80 (5.0%) | 76/80 | 1/80 | $8,864 |
| Qwen2.5-7B | Raw canonical replay | 114/480 (23.8%) | 19/80 (23.8%) | 0/80 | 0/80 | $8,046 |
| Qwen2.5-7B | Typed arrival FIFO | 465/480 (96.9%) | 76/80 (95.0%) | 3/80 | 2/80 | $2,608 |
| Qwen2.5-7B | COVER typed ACI | 468/480 (97.5%) | 78/80 (97.5%) | 0/80 | 2/80 | $1,518 |
| Qwen2.5-7B | Typed reference order | 468/480 (97.5%) | 78/80 (97.5%) | 0/80 | 2/80 | $1,518 |

Table 19: **Reader transfer and equal-call controls in one place.** All histories use fixed Qwen-compiled sources and a Qwen-7B judge. A–B use 476 roots for accuracy; C is a separate predefined 60-root scale probe. Parse-failure rates pool three readout conditions over all 500 roots in A and all 60 roots in C. Raw-2 verifies a first answer using the same raw evidence; two-call limits do not equalize total pipeline cost. Entries are percentages or paired 95% root-bootstrap point differences. Bold marks intervals excluding zero in either direction.

**A. Judge accuracy ($n = 476$) and pooled reader parsing**

| Reader | Raw-1 | Facts | Cascade | Raw-2 | Parse fail |
|---|---|---|---|---|---|
| Qwen2.5-7B | 32.98 | 35.92 | 38.87 | 25.42 | 4.80 |
| Llama-3.2-3B | 12.39 | 9.03 | 17.65 | 25.00 | 59.13 |
| Mistral-7B | 28.57 | 23.95 | 30.67 | 24.58 | 9.80 |

**B. Paired effects on the same 476 roots**

| Reader | Cascade − Raw-1 | Cascade − Raw-2 | Raw-2 − Raw-1 |
|---|---|---|---|
| Qwen-7B | **+5.88 [1.68, 10.29]** | **+13.45 [9.03, 18.07]** | **−7.56 [−11.13, −3.99]** |
| Llama-3B | **+5.25 [3.36, 7.35]** | **−7.35 [−11.55, −3.36]** | **+12.61 [8.82, 16.39]** |
| Mistral-7B | +2.10 [−1.89, 5.88] | **+6.09 [1.89, 10.29]** | **−3.99 [−7.77, −0.42]** |

**C. Reader scale on a separate 60-root set**

| Reader | Raw | Facts | Cascade | Cascade − raw | Parse fail |
|---|---|---|---|---|---|
| Qwen2.5-7B | 36.67 | 40.00 | 45.00 | +8.33 [−6.67, 23.33] | 6.11 |
| Qwen2.5-72B | 56.67 | 48.33 | 58.33 | +1.67 [−11.67, 15.00] | 5.56 |

for all three readers. Its gains are $+5.88$ points for Qwen (95% CI $[1.68, 10.29]$), $+5.25$ for Llama ($[3.36, 7.35]$), and $+2.10$ for Mistral ($[-1.89, 5.88]$). Llama's 59.13% parse-failure rate across all 500 roots and three readout conditions makes its absolute accuracy weak; the selector uses raw on 406/476 roots, containing the noisy structured reader.

Table 21 reports flat append under 10% target retry. COVER has zero answer, correctness, context, and state disagreement in every row. Accuracy varies by type, but schedule dependence is not localized to the weakest types.

Table 20: All-500 local judge accuracy by LongMemEval-S question type.

| Type | $n$ | Raw | Facts | Hybrid |
|---|---|---|---|---|
| knowledge-update | 78 | 44.87 | 56.41 | 41.03 |
| multi-session | 133 | 16.54 | 18.05 | 11.28 |
| single-session-assistant | 56 | 53.57 | 58.93 | 50.00 |
| single-session-preference | 30 | 13.33 | 13.33 | 6.67 |
| single-session-user | 70 | 62.86 | 55.71 | 52.86 |
| temporal-reasoning | 133 | 21.80 | 27.82 | 20.30 |

**Table 21:** Flat append under reorder plus controlled retry by official LongMemEval-S type; disagreements are within-root pairwise rates.

| Type | $n$ | Acc. | $D_{ans}$ | Fork | $D_{correct}$ | Flip |
|---|---|---|---|---|---|---|
| Knowledge update | 78 | 69.9 | 79.1 | 77/78 | 7.4 | 14/78 |
| Multi-session | 133 | 29.3 | 83.3 | 130/133 | 9.7 | 29/133 |
| SS assistant | 56 | 60.1 | 79.8 | 54/56 | 5.7 | 7/56 |
| SS preference | 30 | 40.0 | 88.4 | 30/30 | 14.7 | 9/30 |
| SS user | 70 | 75.5 | 76.4 | 68/70 | 5.4 | 8/70 |
| Temporal reasoning | 133 | 35.7 | 83.4 | 132/133 | 12.2 | 35/133 |

Table 22: Compiler-authority transfer on the predefined balanced 60-root complete-history subset with one fixed Qwen reader/judge.

| Compiler | Chunks | Opaque | Raw | Facts | $\Delta$ | Parse fail |
|---|---|---|---|---|---|---|
| Llama-3.2-3B | 4,613 | 25.38 | 28.33 | 15.00 | -13.33 [-28.33,+1.67] | 4.44 |
| Mistral-7B | 4,613 | 10.25 | 26.67 | 41.67 | +15.00 [+1.67,+28.33] | 6.67 |

Compiler performance also varies on the predefined balanced 60-root subset. With the Qwen reader and judge held fixed, Mistral compilation raises facts over raw by 15.00 points ($[1.67, 28.33]$), whereas Llama compilation changes it by $-13.33$ ($[-28.33, 1.67]$). The contrast shows heterogeneous compiler transfer while confirming that the interface can use compilers other than Qwen. The Mistral pilot estimate does not persist on the independently specified remaining 440 roots: facts minus raw is $+0.68$ judge points ($[-4.32, +5.68]$). The public-status cascade nevertheless improves F1 by $+5.32$ points ($[+1.57, +9.08]$), with an unresolved $+4.09$ judge-point change ($[-0.68, +8.86]$); Table 2 reports the full-scale comparison.

Holding the 60 complete histories, Qwen compiler outputs, context budget, prompt, parser, and Qwen2.5-7B judge fixed isolates reader scale in Table 19. The 72B reader substantially improves the raw condition, while facts alone fall below it. The fixed public-status cascade recovers that loss: cascade judge accuracy is 58.33% versus 56.67% raw (paired $+1.67$ points, 95% CI $[-11.67, 15.00]$). The facts-only comparison therefore varies with reader scale, whereas the label-agnostic dual-view policy contains the observed 72B regression. Compiler granularity must still be selected jointly with reader capability and budget.

**Fact-local totalization diagnostic.** Admitting every individually valid fact while retaining malformed source residue reduces opaque-only chunks from 3.42% to 0.25% for Qwen, 25.38% to 3.69% for Llama, and 10.25% to 1.21% for Mistral. This representation-level recovery does not translate into utility. On the remaining 476 Qwen roots, facts judge accuracy changes from 35.92% to 35.29% (paired $-0.63$ points, 95% CI $[-3.15, 1.89]$). On the balanced 60 roots, Llama facts remain at 15.00%, while Mistral changes from 41.67% to 38.33%. Thus Theorem 5's information refinement does not imply that a fixed-budget neural reader will exploit the added state; this is the budgeted-refinement failure mode proved above.

**Storage and failure accounting.** The fact-plus-evidence payload is $1.75\times$ the source size; index compression is unmeasured. All 1,314 opaque chunks, 72 invalid parses, and native failures remain in the analyses. Action-relay intervals use 10,000 root-cluster bootstrap draws over 80 tasks, jointly resampling their executors and schedules.