COMMUTATIVE MERGE DOES NOT ENSURE ORDER-INVARIANT AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

An associative, commutative, and idempotent merge guarantees order-independent combination of fixed artifacts. In language-model memory, however, the artifacts are generated from batches of observations, and arrival order can change those batches. We test the gap between these two operations. In a controlled 4B study, all 300 world–arm cells produce multiple states under free arrival schedules; the same 300 cells produce a single state when extracted sub-batches are held fixed and only their processing order changes. The control localizes variability before fixed-artifact merging; the relative contributions of changed context and generation remain unresolved. A deterministic counterexample explains why merge algebra alone is insufficient, and canonical replacement gives agreement under fixed extraction outputs without requiring event-wise decomposition. We then measure what agreement establishes about utility. Historical design effects vary across scales and substrates. With atoms fixed, grouping and latest-value cues have positive interactions of 12.87 and 11.11 percentage points under two instructions, both surviving six-test Holm correction under the stated sign-flip reference assumption. In a 12-world online comparison, per-event extraction agrees across the evaluated arrival orders at 23.86% of canonical trie's summed request time, while trie's question-level gains and losses are mixed. Order-invariant merging leaves the production of its inputs outside its guarantee; answer quality and construction cost remain separate empirical outcomes.

1 INTRODUCTION

An agent may receive the same observations in different orders after a delayed tool call, a retry, or recovery from a checkpoint. We ask whether its final memory agrees across these schedules when observation identities, contents, and event timestamps are held fixed. Associative, commutative, and idempotent (ACI) merge guarantees agreement for a fixed collection of objects. In a language-model memory pipeline, those objects are themselves produced from batches of observations, and arrival order can change which observations share a batch. *Does the fixed-input guarantee extend through extraction?*

ACI merge is insufficient for the whole pipeline. Figure 1 gives a deterministic counterexample: extracting the same observations separately or together can yield different artifacts before the same set-union merge. The experiment tests the corresponding boundary. Free arrival schedules produce multiple states in every extraction-based world–arm cell of the controlled 4B comparison; fixing extracted sub-batches and permuting only their processing order produces one state in every cell (Section 2). This control isolates merging of fixed operands, while leaving the relative contributions of context changes and generation variability unresolved.

Agreement without event-wise decomposition. Canonical replacement takes a second route: make each active extraction input a function of the observed set and replace obsolete outputs. With fixed extraction outputs and a deterministic query projection, equal observed sets then give equal visible snapshots (Section 3). Broader replay checks this condition across scales and a second model family using shared artifacts; the online experiment separately generates fresh outputs. The formal note records the sufficient conditions; the empirical work measures the behavior within their scope.

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.

What remains once agreement is controlled. The main claim is that merge algebra alone is insufficient for order-invariant extracted memory. The remaining experiments ask what this property establishes about useful answers. The historical scale–substrate matrix tests whether design rankings transfer between settings. A matched-atom factorial isolates grouping–latest complementarity in exact-output success. A fresh online run holds readout common across construction policies and measures state agreement, accuracy, and request time. These are separate interventions on the pipeline; they do not establish a generally superior memory system. Together they distinguish a reproducibility property from the utility and cost of the memory it produces.

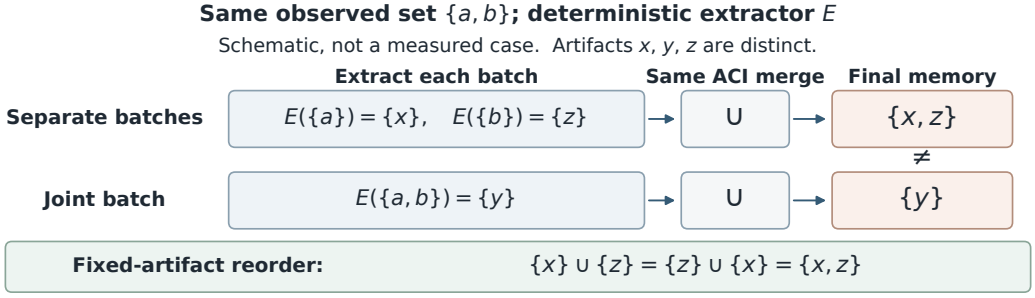


Figure 1: A schematic counterexample, not a measured case. For a fixed deterministic extractor E and distinct artifacts x, y, z, separate and joint extraction of the same events supply different operands to identical set-union merges. Reordering fixed artifacts preserves their union (bottom). This illustrates the distinction tested by the replay control.

2 Tracing arrival-order sensitivity through the pipeline

Historical setting. We reanalyze eight memory designs on 40 synthetic worlds and 20 worlds constructed from Wikipedia infobox revisions, using Qwen3-1.7B, 4B and 8B (Yang et al., 2025). Raw memory preserves the arrival log; sequential and batched memories summarize observations. Five extraction-based designs share atoms but differ in chronological rendering, slot grouping, and lexical or semantic normalization. A further 19-world probe is reconstructed from the conflict slice of MemoryAgentBench (Hu et al., 2025). We report the full 30-cell design comparison below; Appendix B describes the paired world-bootstrap analysis; the supplement provides the complete intervals and inference family.

The intervention: fix extraction outputs and permute their order. Free arrival order changes which observations are jointly extracted. At 4B, all 300 world–arm cells yield multiple states under 20 arrival paths. The five extraction-based arms average 20.0 distinct states on synthetic worlds and 19.2 on revision-derived worlds. Holding the extracted sub-batches fixed and varying only their processing order yields one state in all 300 world–arm cells over five processing orders. The three non-extraction controls retain an average of 5.0 distinct states on synthetic worlds and 4.9 on revision-derived worlds under those five orders. The merge behaves consistently once its inputs are controlled. This intervention localizes the discrepancy upstream of deterministic merging; because it also fixes extraction outputs, it does not separately identify the contributions of changed batch context and fresh-generation variability.

Canonical final-world partitions provide a broader replay check. All 1,290 world–arm cells yield one final state across their paths (Table 1). The calculation fixes both the final-world partition and the extracted artifacts. It therefore tests replay of these artifacts; the online experiment in Section 5 separately regenerates them. This separation is crucial because deterministic merging and independently repeatable extraction are different measurements.

The 900 Qwen cells cover all three scales on both substrates. A Phi-4-mini control contributes 200 further cells: 20 synthetic and 20 revision-derived worlds, five extraction-based arms, and 20 paths per cell. This broadens the conditional replay observation beyond one model family. It keeps the same shared-artifact boundary; second-family answer quality and fresh generation require their own measurements.

Table 1: Historical replay with canonical final-world partitions and shared extraction artifacts. A cell is a world–arm pair within a model run; five extraction-based arms are included. Paths are repeated measurements of the same world.

Setting	World–arm cells	Paths per cell	Single-state cells
Three Qwen scales, synthetic + revisions	900	80	900
Phi-4-mini, synthetic + revisions	200	20	200
Two Qwen scales, benchmark-derived	190	40	190
Total	1,290	—	1,290

Why this points to extraction. Write $E(C)$ for the artifact extracted from a batch C. In Figure 1, $E(\{a\}) = \{x\}$, $E(\{b\}) = \{z\}$, and $E(\{a, b\}) = \{y\}$. The two partitions therefore yield $\{x, z\}$ and $\{y\}$ under the same union, even with deterministic extraction. A separate failure occurs when an updated context produces a new artifact but the old one remains active. Canonical replacement in Section 3 handles both concerns by fixing current extraction contexts as functions of the observed set and replacing obsolete artifacts.

2.1 Agreement does not determine accuracy

Measured accuracy after replay. The corresponding accuracy comparison pairs canonical final-partition replay with arrival-dependent partitioning in 20 contrasts at 4B and 8B. Within the declared 70-test family, one repair contrast survives Benjamini–Hochberg: the 8B synthetic flat arm gains 2.91 points $[1.07, 4.60]$. None survives Bonferroni. These accuracy measurements complement the exact state-agreement counts; their intervals quantify what the finite evaluation resolves. Appendix B gives the uncertainty range, and Section 5 measures actual online construction cost.

Thirty contrasts at three scales. We compare five pairs of designs at each model scale and substrate (Table 2). Fact-wise versus slot-wise rendering contrasts SEMA-FLAT with SEMA; the temporal bundle contrasts SEMA with the chronological CANONICAL renderer. Normalization variants modify the grouped store, and the final contrast compares extracted flat atoms with the raw log. These are design contrasts, including changes in grouping, temporal cues, and retrieval units; they are not five independently randomized components.

Table 2: All 30 historical accuracy contrasts, in percentage points. Positive values favor the first design named. Normalization contrasts use the grouped base; the temporal bundle uses chronological atoms. Each cell averages paired world differences: 40 synthetic or 20 revision-derived worlds. Markers use the retrospective 70-test family: ** Bonferroni; * Benjamini–Hochberg at $q = .05$ only; $^{\circ}$ pointwise bootstrap interval excludes zero only. Full intervals are in the supplement; Appendix B defines the inference.

Contrast	Substrate	1.7B	4B	8B
Fact-wise vs. slot-wise	Synthetic	$+16.51^{**}$	$+4.16^{*}$	$+7.50^{**}$
	Revisions	$+3.49$	$+0.61$	$+0.80$
Temporal bundle	Synthetic	$+0.95$	-2.84	-5.36^{*}
	Revisions	$+10.80^{**}$	$+4.24^{\circ}$	$+3.58^{\circ}$
Semantic normalization	Synthetic	$+6.23^{**}$	$+10.57^{**}$	$+14.66^{**}$
	Revisions	-9.23^{**}	-24.27^{**}	-14.47^{**}
Lexical normalization	Synthetic	$+1.14^{*}$	$+0.68^{*}$	$+0.77^{*}$
	Revisions	$+1.50$	$+0.22$	-0.33
Flat atoms vs. raw log	Synthetic	$+22.70^{**}$	$+9.58^{**}$	$+14.51^{**}$
	Revisions	$+14.27^{**}$	-5.21°	$+1.75$

A synthetic improvement becomes a revision-data loss. Semantic normalization improves accuracy by 6.23, 10.57 and 14.66 points on synthetic worlds at 1.7B, 4B and 8B, respectively. On revision-derived worlds, the corresponding effects are -9.23, -24.27 and -14.47 points. All six directions survive Bonferroni correction in the declared family. Thus a decision to adopt normal-

ization based only on these synthetic results would select a component that lowers accuracy on the revision-derived task at every tested scale. Reconstruction shows that the implementation changes mainly attribute names, and can alter which slot histories fit into the reader's query window (Appendix C.1). The result concerns this implemented normalization and retrieval pipeline, rather than semantic resolution as a universal operation.

Design rankings vary across the recorded checkpoints. Across the ten contrast-by-substrate sequences, four point-estimate sequences are monotone, six are non-monotone, and three cross zero; these are descriptive counts over three checkpoints. None of the five design contrasts has an adjusted positive advantage in all six scale–substrate conditions. The clearest tested reversal is flat atoms relative to the raw log on revision-derived worlds: $+14.27$, -5.21 and $+1.75$ points across the three scales. The paired scale changes are -19.48 points $[-25.64, -13.67]$ and $+6.96$ points $[2.69, 11.27]$. Both survive Benjamini–Hochberg in the 70-test family; the first also survives Bonferroni. The 8B level contrast itself remains unresolved. The full matrix retains adverse and unresolved outcomes alongside positive ones. These retrospective checkpoint comparisons describe the archived runs; they do not isolate parameter-count effects with complete execution provenance. The observed design ranking does not transfer consistently across these runs. Raw memory ranks first of eight on revision-derived worlds at 4B, and fifth and fourth at 1.7B and 8B, respectively. It has the largest median stored-state token count at every scale; this storage measure is distinct from the online request-time comparison.

The temporal bundle motivates a controlled readout experiment. On revision-derived worlds, the temporal bundle has positive point estimates at all three scales, but only the 1.7B contrast survives Bonferroni. On synthetic worlds its point estimate is negative at 4B and 8B. The benchmark-derived supersession probe shows much larger bundled gains: 55.13 points at 4B and 46.89 at 8B, positive in every one of the 19 analysis worlds. Grouping, latest cues, and retrieval units change together in that probe. Reconstructed prompts contain literal latest-value witnesses for nearly all audited questions under both renderers, including many flat-renderer errors (Appendix C). This leads to the controlled question in Section 4: which presentation choices help when every atom is fixed and visible?

3 Formal note: extending fixed-input agreement

The fixed-input merge guarantee remains intact. For chunk-and-join extraction to remain invariant over every partition, the extractor must additionally satisfy $E(A \cup B) = E(A) \vee E(B)$ for disjoint observation sets (Proposition 1, Appendix A). Contextual extraction can violate this condition: interpreting observations together can change the atoms they produce. Figure 1 illustrates this possibility with a deterministic extractor; the replay control tests merging once the generated operands are held fixed.

Canonical replacement. With stable observation identities, make each bucket's serialized extraction input a function of the current observation set; replace its old artifact when that input changes; and fix the full query projection, including rendering and retrieval. Conditional on a fixed extractor, or on a shared immutable table of extraction outputs, equal observed sets then give equal query-visible snapshots (Proposition 2). This accommodates contextual extraction without requiring it to decompose over individual events. An asynchronous implementation must reject outputs for superseded bucket versions.

Two constructions, with costs measured separately. Per-event extraction uses singleton buckets. A capacity-limited canonical trie retains multi-event context while bounding logical update work to at most two extraction evaluations and $k + 1$ event presentations per insertion, for leaf capacity k (Proposition 3). These logical bounds do not determine request time. Section 5 compares the actual costs; Appendix A gives the statements, short proofs, and implementation conditions. The formal note explains the empirical boundary rather than asserting a new merge algebra.

4 GROUPING AND LATEST CUES ARE COMPLEMENTARY

Matched-atom design. We fix the extracted atom multiset and vary its presentation and the answer instruction. Qwen3-4B processes 19 benchmark-derived worlds, each with nine supersession and three clean questions. Four renderers cross slot grouping with latest cues: C has neither, G groups slots, L marks latest values, and GL combines both. Each renderer is evaluated under the original answer instruction and an explicit recency instruction. All renderers retain the same atom fields and conflict flags, expose every atom, and use no retrieval or clipping. Grouping changes slot order and blank-line separation, so ordering is part of the treatment. A separate raw-log condition uses explicit recency and changes the input content.

The worlds come from a MemoryAgentBench reconstruction that selects gold-compatible mappings: 201 of 205 mapped conflict queries are retained and four violating the builder's latest-wins rule are excluded. The factorial then uses the fixed nine-plus-three query template per world. Its task and distribution differ from the original benchmark protocol.

Estimands and inference. The frozen local analysis plan specifies supersession as the primary stratum and gives worlds equal weight. The six primary contrasts are $G - C$, $L - C$, and the interaction $I = GL - G - L + C$ under each instruction. We enumerate all 2^{19} world-sign flips and apply Holm correction across these six tests. The reference null requires the joint vector of world differences to be invariant to flipping any subset of its signs; a zero mean alone does not imply this property. The test is exact under this assumption, which the design does not ensure. Paired world-bootstrap intervals use 100,000 resamples and are pointwise. Sources were previously examined and reused across worlds, so this is an exploratory intervention with a locally frozen plan, rather than external preregistration or independent-source confirmation.

Table 3: Matched-atom supersession exact-output results at 4B, averaging 19 worlds equally. Accuracies are percentages; contrasts are percentage points; intervals are pointwise 95% world-bootstrap intervals. G changes grouping/order, L adds latest cues, and I is their difference-in-differences. Holm correction covers all six rows in the lower panel. Raw is a separate content-changing control.

Instruction	C	G	L	GL	Raw
Original	11.11	23.39	14.62	39.77	—
Explicit recency	9.94	18.13	9.94	29.24	1.17

Instruction	Contrast	Estimate [95% CI]	Raw p	Holm p
Original	$G - C$	12.28 $[4.68, 19.88]$	.00891	.03564
	$L - C$	3.51 $[-0.58, 8.19]$	.25586	.51172
	I	12.87 $[7.02, 19.30]$	.00156	.00934
Explicit recency	$G - C$	8.19 $[2.34, 14.04]$	.02246	.06738
	$L - C$	0.00 $[-3.51, 4.09]$	1.00000	1.00000
	I	11.11 $[5.85, 16.96]$	.00195	.00977

Grouping increases the benefit of latest cues. The grouping–latest-cue interaction is 12.87 points $[7.02, 19.30]$ under the original instruction and 11.11 points $[5.85, 16.96]$ under explicit recency; both survive the six-test Holm adjustment. Descriptively, the combined renderer exceeds the plain renderer by 28.65 and 19.30 points, respectively. Grouping alone adds 12.28 points under the original instruction (Holm $p = .03564$). Its 8.19-point effect under explicit recency has a positive pointwise interval but falls short of the corrected threshold (Holm $p = .06738$). The latest cue alone has no statistically resolved gain under either instruction.

These results identify complementarity in exact-output success for the tested reader and instruction formats: the benefit of marking the latest value is larger when values are grouped. Because all atoms remain visible, the effect arises without adding evidence or improving retrieval coverage. Conflict flags remain fixed, and grouping includes ordering; the factorial therefore targets this presentation combination rather than conflict typing or grouping alone in every possible layout.

The primary contrasts persist under a larger output budget. The execution contains all 2,052 designated answers and 2,644 requests. A separate paired diagnostic regenerates every main prompt

at both 64 and 256 output tokens. Length terminations fall from 74 to seven, while all exact-score correctness indicators remain unchanged (729/2,052 at each budget). Every primary contrast is therefore unchanged under this budget extension. Appendix D reports the full accounting; the result concerns the frozen first-line exact scorer on these prompts.

5 CONTEXTUAL REPLAY COSTS MORE THAN THE AGREEMENT BASELINE

Result in the evaluated execution. Each of the three policies produces identical final atoms and raw answers across its two arrival paths in all 12 worlds. Per-event extraction uses 0.617 request-hours; canonical trie incurs 4.19 times the summed request time and generates 4.36 times as many output tokens, for a 2.74-point gain in world-equal exact accuracy. Per-event time is 23.86% of trie's. This run separates the inexpensive agreement baseline from the additional cost of contextual reconstruction.

Controlled execution. We evaluate per-event extraction, sorted rebatching, and canonical trie on 12 previously studied worlds: four synthetic, four revision-derived, and four benchmark-derived. Each policy receives the same complete events in published order and one fixed shuffle, using arrival chunks of eight, capacity 12, the same Qwen3-4B revision, and canonical flat readout with explicit recency. This holds readout common across construction policies; it does not jointly test the factorial's grouped-plus-latest renderer with canonical trie. The family-balanced scope follows a deterministic ID ordering and was fixed after observing an earlier run's cost but before the new answer scores. This is an exploratory comparison on reused sources.

All 72 policy–world–path cells complete, with 4,926 fresh requests in 5.36 hours on an RTX 5090/WSL setup. Each method and path generates its own extraction outputs. Offline reconstruction reproduces every update and final state, and independent scoring reproduces all 1,050 main-answer scores from the first raw responses. Repetition probes are included in costs and excluded from accuracy (Appendix F).

Table 4: Complete 12-world exploratory run. Accuracy averages worlds equally after averaging their two paths; counts pool the 350 main answers per policy. Request time includes extraction, answering and repetition probes, excluding model loading and host bookkeeping.

Policy	Accuracy (%)	Correct/350	Extract calls	Request hours
Per-event	36.15	128	1,926	0.617
Sorted rebatch	38.44	136	724	2.126
Canonical trie	38.89	138	1,082	2.584

Fewer calls, more tokens, greater summed request time. For each policy, both paths yield identical final atoms and raw answers. Trie exceeds per-event extraction by 2.74 world-equal accuracy points and rebatching by 0.45, while taking 4.19 and 1.22 times their summed request time. It generates 556,680 tokens, compared with 127,598 for per-event extraction and 461,024 for rebatching. The cost ordering persists after removing repetition-probe requests. Thus fewer extraction calls can coexist with more generated tokens and greater summed request time.

Per-event extraction supplies the lowest measured request-time baseline in this run; trie retains within-leaf context and bounds logical updates. The measured ratios describe these implementations and requests, rather than an intrinsic cost of obtaining order invariance.

The aggregate gain contains both improvements and regressions. The two identical paths repeat 175 distinct questions. Against per-event extraction, trie gains 19 and loses 14; against rebatching, it gains 12 and loses 11. World-level win/tie/loss counts are 5/3/4 and 6/1/5. Supersession gains and losses against per-event extraction cancel at eight each. These paired outcomes show where the aggregate accuracy difference comes from, including its unevenness across questions.

We inspect all 33 disagreements against per-event extraction, retaining source events, full states, prompts, and raw answers. Six involve surface-compatible answers under ordinary reading, including list order, date notation, and extra labels (four gains, two losses). Four show a specific source-linked

relation or field change (two gains, two losses); 23 involve mixed context or readout differences. These are post-hoc AI-assisted annotations with original exact scores retained, rather than blinded human adjudications or causal attributions. Table 5 illustrates the range; Appendix G and the supplement contain the complete ledger.

Table 5: Illustrative disagreements from the complete 33-question audit. E denotes per-event extraction and T canonical trie. Source-linked observations are compared within each question; other context can differ. Original exact scores are unchanged.

Case	Observation	Interpretation boundary
Identical target fact static-02/q04	Both states retain lead officer = Bergstrom. E answers Petrov; T answers Bergstrom.	The gain occurs without adding this target fact; it does not isolate the cause of readout.
Relation direction real-02/q014	E reverses the successor relation; T preserves its direction and answers correctly.	A source-linked extraction difference is visible, but the effect of cross-event context is not isolated.
Missing field real-04/q006	T omits Communication studies from both the education atom and answer; E retains it.	Contextual extraction can lose answer-relevant detail.
Answer label real-02/q010	T adds "predecessor:" before the correct name; E returns the name alone.	Exact-output failure need not imply a wrong named answer.

All policies score zero on the 18 distinct tie-conflict questions under the exact-output rule. One trie response contains both required values but omits the required prefix. This concrete case, together with the unchanged budget-check scores in the separate readout experiment, illustrates why exact protocol success and semantic answer compatibility deserve separate reporting.

6 Related work

Replicated state and deterministic interpretation. ACI merge and deterministic interpretation are established tools (Almeida et al., 2014; Kleppmann et al., 2018). Proposition 1 records an elementary join-preservation condition for the chunk-and-join extractor. It states the applicability of the fixed-input merge argument using classical algebra, rather than introducing a new CRDT theorem. StateFuse scopes LLM-backed replayability to fixed validated responses (Volkov et al., 2026); cache merging combines already-generated latent fragments (Baquero & Brito, 2026). Our empirical contribution concerns the preceding boundary: replay separates fixed-artifact merge behavior from upstream variability, and the readout and online studies measure outcomes that state agreement alone leaves open. The canonical construction applies standard radix partitioning and dynamic hashing (Fagin et al., 1979) to this boundary.

Temporal and conflict-aware memory. Zep uses a temporal knowledge graph (Rasmussen et al., 2025); LatticeMind combines explicit status, symbolic checks, and selective model reconciliation, and reports an extraction bottleneck (Zhou et al., 2026). Building on this recognition of extraction and temporal structure, we separate schedule agreement from the contribution of readout: the factorial estimates grouping–latest interactions with atoms fixed. MemoryAgentBench evaluates incremental memory competencies (Hu et al., 2025); our reconstructed worlds support controlled arrival and readout interventions while changing task packaging.

Native agent histories. LongMemEval-V2 evaluates accuracy and latency on web and enterprise trajectories (Wu et al., 2026). Our native records provide a partial raw-retrieval baseline and construction diagnostics, summarized in Appendix E. They expose the additional demands of full-history memory evaluation, including publication of complete memory, missing reader scores, and unequal note capacities.

7 Discussion and limitations

What survives. ACI merge remains sufficient for fixed inputs, and canonical replacement gives conditional agreement when active extraction inputs are functions of the observed set. **What fails.** That fixed-input argument does not extend to a pipeline whose arrival schedule changes extraction contexts. The controlled replay result localizes variability before merge without separating contextual and stochastic extraction effects; the scale–substrate matrix also leaves accuracy dependent on the implemented design and evaluation setting. **What it costs.** Per-event extraction provides the inexpensive agreement baseline in the online run. Contextual replay incurs greater summed request time and produces both improvements and regressions, with supersession gains and losses cancelling. Thus restoring observed agreement and obtaining useful answers are separate achievements.

Scope of the evidence. Our findings concern structured factual worlds and a small set of model configurations. We compare matched constructions to isolate extraction schedules, while comparative utility against external memory systems remains untested. The historical replay counts reuse extraction artifacts; the new factorial uses one scale and one ingestion path, with conflict flags fixed. The online study uses 12 reused worlds and supports descriptive quality–cost comparisons. Benchmark-derived data contain 1,140 event occurrences but 461 unique texts, with 107 shared with calibration. World-level resampling preserves the repeated-measure structure but supplies no independent source holdout. Historical confidence intervals are pointwise; multiplicity and the conditional sign-flip assumption are reported with their corresponding analyses. The factorial's 64/256-token check establishes robustness of its exact scores on the same prompts, while the online prompts have their own output-format and length failures.

Construction and deployment. The replay guarantee assumes fixed extraction outputs or a shared immutable artifact table, and a fully specified query projection. Independent deployments require direct repeatability checks: model/runtime controls, including batch-invariance settings, have supported-configuration limits (vLLM contributors, 2026). Canonical hash neighbors may be semantically unrelated; splitting leaves can change extracted content, and token lengths can exceed an event-count bound. Native contextual-memory utility remains unmeasured: six full-history builds stopped before benchmark reading after 4,030 writer requests, in addition to earlier pilot and validation failures. The partial raw-RAG baseline has 149/294 rule-scored answers correct, 128 text judgments pending, and 29 unsupported image questions; its two histories and unequal memory capacities limit comparisons. Full diagnostics are preserved in the supplement rather than folded into the complete online study.

A fixed-input merge guarantee ends at its operands. A memory pipeline must also account for how those operands are produced and replaced. Once that boundary is controlled, answer quality and construction cost remain separate empirical questions.

Reproducibility statement

The accompanying CPU package reconstructs the historical comparisons, matched-atom factorial, paired output-budget diagnostic, and online results from recorded inputs and outputs. It also contains the complete 33-question ledger, linked prompts and responses, annotation rules, and a record verifier. A supplementary technical note preserves historical implementation corrections and native execution diagnostics, including a corrected bootstrap procedure that retains repeated-world multiplicities. The full 30-cell matrix and its 70-test family are recomputed from the original world-level records and included as an additional CPU analysis. The factorial has a frozen local analysis plan and immutable model revision; the complete online run retains all 4,926 request records and 72 cells. Full legacy caches and native raw records are retained separately and are not all bundled. These materials enable reconstruction of recorded analyses; independent fresh-inference replication remains a separate experiment.

Ethics statement

Revision-derived labels represent archived field strings, including erroneous edits, rather than independently verified facts. Synthetic and benchmark-altered assertions about named entities are task data, not factual claims about those entities. Reproducible memory can preserve false information; the present results do not establish deployment safety or factual reliability.

AI USE STATEMENT

Language models performed extraction and answering. Generative AI assistance supported synthetic data, code, data processing, statistics, theoretical claims and proofs, hypotheses, experiment design, literature research, result interpretation, qualitative case annotation, and drafting. The qualitative annotations are AI-assisted post-hoc observations, not independent human adjudications. Human authors remain responsible for the final text, claims, proofs, data, and artifacts.

REFERENCES

Paulo Sérgio Almeida, Ali Shoker, and Carlos Baquero. Efficient state-based CRDTs by delta-mutation. *arXiv preprint arXiv:1410.2803*, 2014. URL https://arxiv.org/abs/1410.2803.

Carlos Baquero and Luís Brito. Cache merging as a convergent replicated state for multi-agent latent reasoning. *arXiv preprint arXiv:2607.01308*, 2026. URL https://arxiv.org/abs/2607.01308.

Ronald Fagin, Jürg Nievergelt, Nicholas Pippenger, and H. Raymond Strong. Extendible hashing—a fast access method for dynamic files. *ACM Transactions on Database Systems*, 4(3):315–344, 1979. doi: 10.1145/320083.320092. URL https://research.ibm.com/publications/extendible-hashinga-fast-access-method-for-dynamic-files.

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

Martin Kleppmann, Victor B. F. Gomes, Dominic P. Mulligan, and Alastair R. Beresford. OpSets: Sequential specifications for replicated datatypes (extended version). *arXiv preprint arXiv:1805.04263*, 2018. URL https://arxiv.org/abs/1805.04263.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025. URL https://arxiv.org/abs/2501.13956.

vLLM contributors. Batch invariance. Official documentation, accessed September 7, 2026, 2026. URL https://docs.vllm.ai/en/latest/features/batch_invariance/.

Sergey Volkov, Yang Li, and Ye Luo. StateFuse: Deterministic conflict-preserving memory for multi-agent systems. *arXiv preprint arXiv:2607.05844*, 2026. URL https://arxiv.org/abs/2607.05844.

Di Wu, Zixiang Ji, Asmi Kawatkar, Bryan Kwan, Jia-Chen Gu, Nanyun Peng, and Kai-Wei Chang. LongMemEval-V2: Evaluating long-term agent memory toward experienced colleagues. *arXiv preprint arXiv:2605.12493*, 2026. URL https://arxiv.org/abs/2605.12493.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Heng Zhou, Lian Zhang, Yutao Fan, Tiancheng He, Siki Chen, Hejia Geng, Philip Torr, and Zhenfei Yin. LatticeMind: A conflict-aware memory primitive for multi-agent systems. *arXiv preprint arXiv:2608.08236*, 2026. URL https://arxiv.org/abs/2608.08236.

A PROOF DETAILS AND OPERATIONAL BOUNDARIES

The replay evidence suggests two routes to invariant memory: require extraction to distribute over batches, or give contextual extraction a canonical input and replace its obsolete outputs. We state the distinction for a finite observation set S with stable identifiers. Identical deliveries are deduplicated; conflicting contents under one identifier are rejected. Let $(L, \vee, \bot)$ be a join-semilattice and $E : 2^{\mathcal{U}}_{\mathrm{fin}} \to L$ a fixed extractor with $E(\varnothing) = \bot$.

Proposition 1 (The scope of the join argument)**.** *For $F_{\mathcal{P}}(S) = \bigvee_{C \in \mathcal{P}} E(C)$, equality across every partition $\mathcal{P}$ of every finite S holds if and only if $E(A \cup B) = E(A) \vee E(B)$ for all disjoint finite A, B.*

For contextual extraction, this condition can fail. Canonical replacement instead keeps one current artifact per canonical bucket. Fix a bucket map $h : \mathcal{U} \to \{1, \ldots, B\}$ before ingestion, a deterministic serialization σ, and a fixed extractor E_σ on serialized inputs. For $S_b = \{e \in S : h(e) = b\}$, maintain $A_b = E_\sigma(\sigma(S_b))$ only for nonempty buckets. Each update replaces the previous A_b; the visible memory for query q is $M_q(S) = \rho_q(\{(b, A_b) : S_b \neq \varnothing\})$, where the deterministic projection ρ_q fixes resolution, ordering, rendering, retrieval, and truncation.

Proposition 2 (Canonical snapshot agreement)**.** *Under this contract, any two completed ingestion prefixes containing the same observation set have identical visible memories for the same query. No decomposability of E_σ is required.*

Each bucket, serialized input, and current artifact is a function of S. Applying the fixed projection gives the result. For stochastic extraction, the same statement is conditional on a shared immutable artifact table keyed by complete request bytes and model/parser/policy versions. It concerns completed snapshots at equal observed sets; asynchronous workers must reject results for superseded bucket versions. Proofs and counterexamples follow.

A bounded online implementation. Fixed buckets can grow indefinitely. Instead, recursively split a fixed full-digest binary trie until each nonempty leaf holds at most $k \geq 1$ observations. Serialize each leaf canonically, extract its contents, and replace artifacts for changed leaves. Reject full-digest collisions. An insertion updates one leaf; at overflow, its $k + 1$ members split into two nonempty descendants, each within capacity. Unary prefixes require no extraction.

Proposition 3 (Bounded replay work)**.** *With distinct digests, inserting one unique observation changes at most one existing leaf. That leaf is updated, or replaced by two nonempty descendant leaves. Consequently canonical trie replay requires at most two extraction calls and at most $k + 1$ observation presentations per insertion, while preserving snapshot agreement without knowing the final stream length.*

The bounds give at most $2N$ logical evaluations and $(k + 1)N$ observation presentations over N unique insertions. A fragmented evaluation may use several model requests; token lengths, generated outputs, and request time are measured separately. This construction preserves contextual extraction within leaves while making its active input sets reproducible. Section 5 evaluates its actual quality and inference cost against per-event extraction and sorted rebatching.

Proof of Proposition 1. Assume partition invariance. For disjoint nonempty A, B, compare the two-block partition of $A \cup B$ with its one-block partition. This gives $E(A) \vee E(B) = E(A \cup B)$. Empty blocks are handled by $E(\varnothing) = \bot$. Conversely, induction on the number of blocks gives $\bigvee_{C \in \mathcal{P}} E(C) = E(S)$ for every partition. Singleton refinement also gives $E(S) = \bigvee_{e \in S} E(\{e\})$. This characterization concerns the specified chunk-and-join architecture. A different architecture can be path-invariant without satisfying it; for example, collecting a set and recomputing any deterministic contextual function of that set is invariant.

Proof of Proposition 2. Induct on successful insertions. Initially the observation and artifact maps are empty; only nonempty buckets have active artifacts. A duplicate identifier with identical contents changes neither. For a new observation, only its bucket changes. Canonical serialization and fixed extraction determine the new artifact, and atomic replacement removes the obsolete artifact from the visible projection. Every other current bucket artifact is unchanged and remains correct for its current set. Thus after every completed update the artifact map equals the map obtained by computing all current buckets from scratch. Equal observation sets give equal maps and equal projections. The same proof applies to a set-dependent canonical partition when all changed and removed cells are detected and their artifacts recomputed or retracted. Fixed B is sufficient for local updates, not necessary for every possible invariant design.

Serialization and coupling counterexamples. A partition can remain fixed while its member order changes. An extractor that returns the first serialized event produces different outputs from

$[a, b]$ and $[b, a]$ although both represent the same set. A conformance check that converts members to unordered sets cannot detect this failure. Even identical serialized inputs do not imply identical independently sampled outputs: an extractor returning 0 or 1 with equal probability disagrees between two independent executions with probability $1/2$. A cache shared across paths conditions on one draw; it does not establish independent generation reproducibility. Equal states likewise do not imply equal stochastically generated answers; unequal states do not imply unequal answers.

Exact work for fixed buckets. The jth unique arrival in bucket b presents all j current observations. Summing gives $W_b = n_b(n_b + 1)/2$ and N calls overall. Write $\sum_b n_b^2 = N + 2\sum_{i<j} \mathbf{1}\{h(e_i) = h(e_j)\}$. Uniform independent bucket assignments make each indicator's expectation $1/B$, giving $\mathbb{E}W = N + N(N-1)/(2B)$. In the worst case one bucket receives every observation and $W = N(N+1)/2$. For arrival chunks, the corresponding exact count is one call for each bucket whose membership grows at that step, and the work is the sum of its complete current occupancy at those steps. Duplicate-only delivery steps need no extraction. These are logical counts, not a GPU latency or FLOP model.

A black-box lower bound, and its limited meaning. Suppose an algorithm must expose exactly $E(S_b)$ after each changed bucket version, where E is an arbitrary deterministic oracle on finite sets. If the current set S_b has not previously been queried, an exact algorithm must query it: otherwise two oracles agreeing on every previous query but differing on S_b are indistinguishable to the algorithm. With a full-prompt interface charging one presentation per supplied observation, this entails the replay work above for the fixed-bucket target. The lower bound is relative to exact agreement with that target and this oracle interface. It does not apply to incremental sufficient statistics, extractors with a proved update rule, approximate views, delayed freshness, or alternative memory targets. In particular, per-event extraction is an inexpensive invariant baseline and cannot be omitted on the strength of this lower bound.

Proof of Proposition 3. Define the partition by recursively visiting binary prefixes, stopping at every nonempty prefix whose occupancy is at most k. Empty prefixes have no artifact. All existing internal prefixes have occupancy greater than k. Insertion changes only the descent path for the inserted digest. If it reaches an empty branch, it creates one singleton leaf. Otherwise it reaches one existing leaf. Occupancy at most k after insertion requires one updated artifact. Overflow means occupancy exactly $k+1$; extend the prefix until its members first separate into two nonempty children. Both child sizes are between 1 and k, so no further populated split is needed. Two new artifacts replace one old artifact, and the total presented membership is $k + 1$. Distinct finite digests guarantee separation by the last bit; a collision is rejected before mutation. Induction yields at most $2N$ calls and $(k + 1)N$ observation presentations for N successful unique insertions. Canonical serialization and exact replacement give the canonical-view proof at every prefix. These are bounds on extraction work, not on CPU partition maintenance or total memory. All raw observations remain retained.

Token and semantic boundaries. If every normalized observation has at most L input tokens and each call has at most H fixed instruction tokens, the crude bound is $(k + 1)NL + 2NH$, subject to the tokenizer and serialization contract. Event count alone cannot bound tokens for arbitrarily long observations. Context windows can change the extractor's interpretation; splitting a leaf can therefore change semantic content. The theorem promises reproducible current views and bounded replay work, not preservation of all previously extracted facts, accuracy, replica convergence, or deletion support.

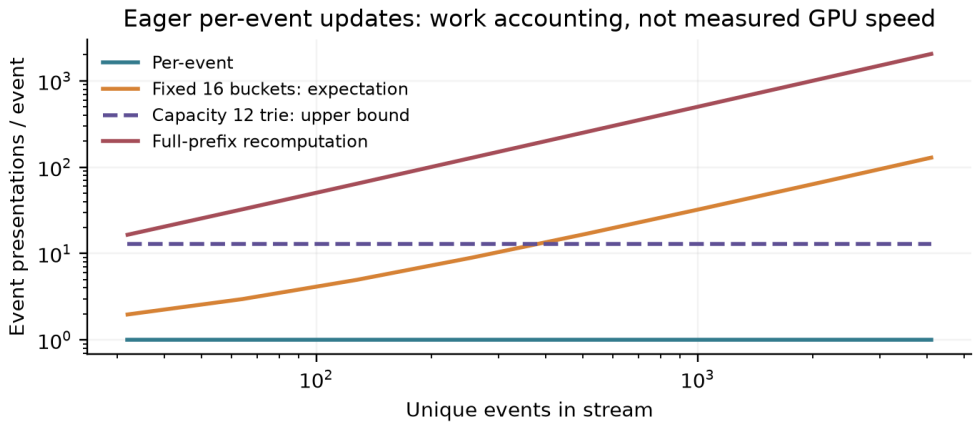


Figure 2: Mathematical work accounting for individual-event arrivals. The fixed-bucket line is an expectation under uniform independent hashing; the trie line is a worst-case upper bound. These are event presentations, not measured tokens, latency, or GPU throughput.

B THREE-SCALE COMPARISONS: ESTIMATES AND ANALYSIS

Design and estimands. The retrospective comparison includes eight memory arms: the arrival log, sequential rewriting, batched summarization, chronological atom rendering, grouped temporal rendering, flat temporal rendering, and lexical and semantic normalization. The scale comparison uses Qwen3-1.7B, 4B and 8B on 40 synthetic worlds and 20 Wikipedia-revision-derived worlds. The benchmark-derived comparison has 19 analysis worlds. Differences average path outcomes within each world and then give worlds equal weight. Scale changes are paired on the same worlds. Storage summaries use the median stored-state token count over recorded query/path rows.

Estimates and full interval record. The complete 30-cell matrix appears in Table 2; its pointwise 95% intervals, all 20 paired scale changes, and all 20 repair contrasts are provided in the supplement's scale reanalysis, with machine-readable estimates and generated tables. The two benchmark-derived representation-bundle contrasts are 55.13 points $[49.74, 60.39]$ at 4B and 46.89 points $[42.06, 51.70]$ at 8B. They are positive in all 19 worlds at each scale. Leaving out any one world keeps the mean within $[54.26, 56.42]$ points at 4B and $[45.51, 48.35]$ at 8B. These benchmark contrasts are separate from the historical 70-test family, and their source texts are reused.

Repair uncertainty. Of the 20 repair contrasts, one passes Benjamini–Hochberg in the 70-test family and none passes Bonferroni. The smallest pointwise lower endpoint across all 20 is -12.49 points, for semantic normalization on 8B revision-derived worlds. Across the 12 contrasts restricted to the chronological, grouped, and flat arms, the smallest lower endpoint is -5.57 points. These ranges leave room for accuracy losses; state agreement and accuracy preservation are distinct claims.

Complete inference family. Repeated paths are averaged within worlds before inference. The historical tests are two-sided one-sample Student tests of paired world differences against zero. Bonferroni and Benjamini–Hochberg adjustments use all 70 tests: 30 level contrasts, 20 paired scale changes, and 20 partition-repair contrasts. Pointwise bootstrap intervals are computed separately with 100,000 family-stratified resamples and seed 20260907, retaining sampled-world multiplicities. Positive point estimates, interval exclusion, and adjusted test decisions are separate properties.

Replay interpretation. Arrival-dependent positional batching changes extraction inputs. The fixed-artifact replay intervention instead reuses extracted sub-batches and changes their processing order, yielding one state for each deterministic extraction-based arm. The 1,290 repaired world-by-arm cells use canonical final-world partitions and shared extraction artifacts. These are conditional final-state replay observations; the implementation can access final-world membership and does not represent a causal online ingestion procedure.

Corrected uncertainty and provenance. All reported intervals are recomputed from world-level rows with 100,000 paired, family-stratified bootstrap resamples and seed 20260907. The released legacy analyzer could collapse repeated sampled-world identifiers, losing their bootstrap multiplicities. The corrected analyzer retains every sampled occurrence; the manuscript and supplementary tables use this corrected calculation. Plan files carry local hashes and the reanalysis is retrospective. Source-run identities, archive checks, cached-output discrepancies, and the unchanged original audit narrative are provided in the supplementary technical report. The records support reconstruction but do not provide complete per-request provenance for the historical runs.

C Recovered reader inputs and normalization diagnostics

We reconstruct the 4B and 8B content-reextract runs; the benchmark contrasts in Appendix B instead report the positional runs. Per scale, all 95 extraction prompts match recovered cache keys and diagnostic atom counts. For CANONICAL, SEMA, and SEMA-FLAT, all 684 world/question/arm groups match stored state hashes, state and visible token counts, and raw answer-prompt cache keys. Each group has 40 historical path observations; the audit covers 54,720 existing rows across both scales, with no persistent-state truncation. The reconstruction uses recorded outputs.

Only after reconstruction do we join the builder's event references. A literal witness requires the referenced value and source to occur in the same displayed date unit at the referenced time. It does not validate the entity/relation or establish sufficient semantic support. The design-latest event is selected by date among the referenced events; historical coverage checks all references. Table 6 keeps these distinct. The original encoded entity/attribute fields contain brittle text splits, so unmatched witnesses remain unresolved rather than being scored as absent semantic evidence.

Table 6: Retrospective supersession readout audit on the content-reextract branch. Visible literal witnesses are counted over 171 questions per scale (19 worlds). Accuracy averages the original path outcomes within worlds; no new significance test is attached.

Scale	Renderer	Latest visible	All referenced visible	Accuracy
4B	Canonical	167/171	160/171	.2627
4B	Sema	166/171	159/171	.8168
4B	Sema-flat	166/171	159/171	.7563
8B	Canonical	170/171	166/171	.2596
8B	Sema	169/171	165/171	.7135
8B	Sema-flat	170/171	166/171	.5810

The literal latest-witness deficit does not track the large accuracy gap. This observation constrains a simple missing-latest-value account but does not isolate cues, grouping, ordering, distractors, or a reader failure mechanism. A stricter displayed entity/attribute match is much less frequent (25–33 of 171 questions across arms/scales), reinforcing the need to distinguish lexical checks from semantic adjudication.

Finally, 17 of 1,368 reconstructed answer-prompt groups contain multiple stored normalized predictions across paths. Replacing all observations by the final cache value would disagree with 171 recorded predictions and change 77 correctness indicators. We retain the original path outcomes. The cache does not identify the execution cause of these historical discrepancies.

C.1 Semantic normalization changes attribute addressing

The sign reversal in Table 2 motivates an error-level audit of the six positional main runs. We reconstruct all 54,420 SEMA/SEMA-SEM world/path/query pairs from recorded extraction and normalization outputs. The 108,840 historical rows match reconstructed state hashes, state and visible token counts, and answer-prompt cache membership. No persistent state is truncated. The analysis retains the original scores and uses 20 paths per world, compared with 40 in the content-reextract analysis above.

The intervention is predominantly attribute normalization. The resolver groups sorted entity strings and sorted attribute strings separately. An entity guard rejects groups covering more than 34%

of the current surface-form list or lacking a shared nongeneric token. Consequently, any two-item group is rejected when the list has at most five forms. Attributes have no such guard. Entity strings change in only 6 of 3,600 world/path/scale cells, whereas attribute strings change in 3,553. Reusing the same cached decisions but applying only the attribute map produces byte-identical full SEMA-SEM states in 3,594 cells. This CPU transformation identifies attribute normalization as the main implemented operation in the nominal "entity-resolution" arm. It reuses cached maps and does not regenerate reader answers or validate semantic equivalence.

Persistent presence and query access differ. For this diagnostic only, an annotated witness is a post-deduplication atom whose normalized value and exact event time/source match a builder-referenced event. A witness is counted as visible if its rendered line is selected. This lineage predicate does not check the entity/relation, prove entailment, or restore source/value binding lost by set-valued rendering. We report the fraction of queries having at least one such witness, averaged within worlds and then equally across worlds (Figure 3). Synthetic coverage rises, whereas revision-derived coverage falls at all scales. The old full-state value-presence flag changes in only 52 of 54,420 pairs, so it cannot stand in for this readout measurement.

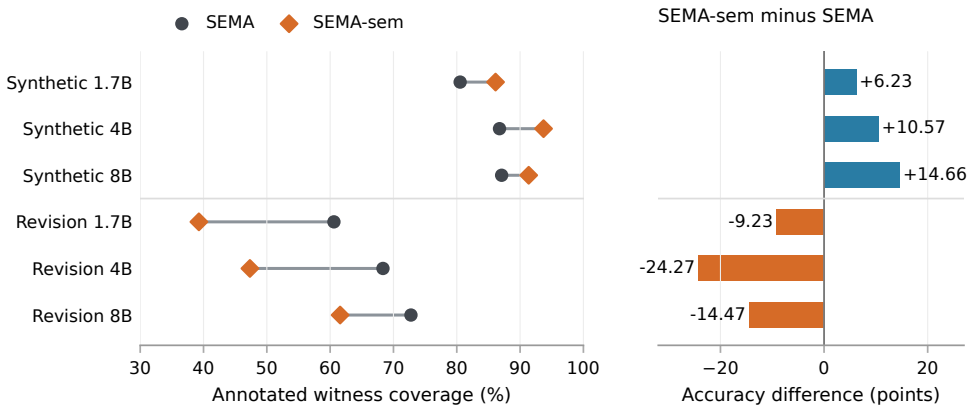


Figure 3: Retrospective semantic-normalization diagnostics. Left: world-weighted query coverage by at least one annotated value/time/source witness in the selected 512-token window. Right: original world-weighted accuracy differences in percentage points, not scores from regenerated cache answers. Paths and questions remain repeated historical observations. The two columns show associations, not a causal mediation estimate.

The whole-line retriever skips a line when its token cost plus one separator exceeds 512, even if the persistent state fits its 6,144-token budget. Attribute normalization can combine histories into a longer slot line while shortening the total state. In the 4B revision-derived run, 1,206 of 5,340 pairs lose all previously visible annotated witnesses. These pairs contain 868 of the run's 1,461 right-to-wrong transitions; 585 of those 868 have a referenced witness on a SEMA-SEM line whose cost, including its separator, exceeds 512. The full 1,206-pair witness-loss stratum contributes -15.82 percentage points to the world-weighted accuracy difference, out of -24.27 overall. This is an additive descriptive partition, not the fraction of errors caused by retrieval: 505 right-to-wrong transitions retain a visible witness, and normalization also changes relation names, temporal grouping and conflict displays.

Reconstruction scope. The extant harness hash matches only one of the six run manifests. Four failed extraction records in two other runs report nonzero atom counts; independently parseable atom-object fragments from their raw responses recover those counts and the resulting stored states without changing text or filling missing values. This does not recover the missing parser source. Final cached answers agree with both stored normalized predictions and raw prefixes in 53,668 of 54,420 pairs; the other 752 pairs remain in the original score totals and are listed separately. Restricting to the agreeing pairs retains all six effect directions, but this selected subset is only a sensitivity check. Cache snapshots do not identify the execution cause of answer discrepancies. These diagnostics link

the normalization intervention to changed attribute addresses and query access. Controlled reader interventions would be needed to estimate mediation or validate the normalized groups semantically.

D MATCHED-ATOM EXECUTION AND READOUT DIAGNOSTICS

Coverage and first-answer scoring. All 19 jobs, 2,052 designated main answers and 2,644 recorded requests are present. Every first designated answer is retained. At the 64-token budget, 74 main answers terminate by length, all under explicit recency. For supersession the counts are $18, 7, 7, 0$ in C, G, L, GL and 39 for raw; three additional raw clean answers terminate by length. All 912 original-instruction answers avoid length termination. Only two worlds meet the local plan's conservative criterion of no main-answer length termination, but the primary analysis includes every world under the frozen exact scoring rule. Some stop-terminated outputs also contain explanatory prose and fail exact scoring.

The raw-log supersession accuracy is 1.17%. Relative to raw, explicit-instruction C and GL differ by 8.77 points $[4.68, 12.28]$ and 28.07 points $[21.05, 35.09]$. Raw changes the input content and is a descriptive control outside the six-test factorial family. Selected fresh probes agree in all 38 extraction and 38 answering comparisons; extraction atom sets also agree in all 38 pairs. The probes sample individual requests rather than complete independent pipelines.

Paired output-budget sensitivity. For each of the 2,052 original main prompts, a paired diagnostic generates fresh answers at 64 and 256 tokens, in a frozen interleaved order. Atom sets, model revision and scoring rule remain fixed. All 4,104 requests complete. Independent rescoring of the recorded first-line outputs reproduces every exact score; no original answer is replaced.

Table 7: Paired answer-budget sensitivity. Counts pool the fixed prompts and conditions; they are not independent source replications.

	64 tokens	256 tokens
Fresh answers	2,052	2,052
Exact correct	729	729
Length terminated	74	7
Generated tokens	15,590	20,877
Summed request seconds	564.862	655.012

All 729 correct and 1,323 incorrect predictions retain their correctness across budgets; the new 64-token scores also match the originals item by item. Every world-level factorial contrast is therefore unchanged. All 74 length-terminated strings change, with 67 now ending normally; 15 normalized first lines change, but none becomes exact-correct. The remaining 1,978 raw strings are identical. The longer allowance resolves most length stops while preserving the measured contrasts under the frozen parser. It does not test semantic equivalence under other scoring rules.

Each budget processes 5,905,781 input tokens. The diagnostic takes 1,292.292 seconds including startup and bookkeeping; summed request time is 1,219.875 seconds. The longer allowance adds 5,287 output tokens and 90.150 seconds of summed request time on this RTX 5090/WSL execution.

Model, execution and inference. The factorial uses Qwen3-4B revision `1cfa9a720891`, with the complete snapshot identifier in the model lock: bf16, temperature zero, seed zero, eager execution, a 16,384-token context, one request at a time and prefix caching disabled. Thinking is disabled where supported by the tokenizer template. The recorded runtime uses vLLM 0.28.0, Torch 2.13.0+cu132, WSL2 and an RTX 5090 reporting 32,607 MiB. All recorded input lengths plus output reserves fit the context.

The main session records 6,207,621 input and 124,801 output tokens over 2,316.143 seconds, including 29.831 seconds of model loading. Summed generation times are 1,571.260 seconds for extraction, 502.074 for answering and 174.127 for repeatability probes. There are no failed jobs or budget stops. These are measurements of this run, rather than a native-history runtime estimate.

Independent reanalysis reproduces all six contrasts, bootstrap intervals and leave-one-world-out ranges. A second sign-flip implementation uses integer probability-mass recursion instead of enumer-

ation and obtains the same raw and Holm-adjusted probabilities. Sign-flip exactness is relative to its reference assumption; source reuse remains relevant to generalization. The locally frozen plan is not public preregistration.

E NATIVE TRAJECTORY EVIDENCE AND REMAINING SCOPE

Raw-state retrieval baseline. The complete text-evidence baseline uses both LongMemEval-V2 Small histories, with 200 trajectories and 5,095 states. It embeds raw state slices with Qwen3-Embedding-8B and reads retrieved source contexts with Qwen3.5-9B at pinned revisions. Embedding uses a 4,096-token prefix; the reader receives up to six source entries within a 200,000-token memory budget, a 262,144-token context declaration and a 20,000-token output allowance. The run records 622 embedding calls and 422 first reader responses. The fixed official parser and rules score 149 of 294 rule questions correctly; 128 model-judged text questions are unscored and 29 image questions are unsupported.

Table 8: Native raw-RAG rule-subset results on two complete Small histories. Pending text judgments and unsupported image questions retain their fixed denominators. The pooled rule fraction is descriptive, not the complete benchmark score.

History	Rule correct/total	Accuracy (%)	Pending judges	Image questions
Web	68/153	44.44	72	15
Enterprise	81/141	57.45	56	14
Pooled	149/294	50.68	128	29

Of 31 length-terminated responses, seven match a rule-scored answer through the official parser's fallback to reasoning text; they have no completed final-content answer. Requiring final content yields 142/294 (48.30%) as a separate scoring sensitivity. Full text accuracy remains bounded by $149/422$ to $277/422$ (35.31–65.64%) until pending judgments are supplied; these are missingness bounds. The two shared histories provide no independent-history significance test. Embedding and reader stages take 17.662 recorded host hours, excluding startup, downloads and packaging.

Contextual-memory constructions. The initial segmented pilot commits two insertions before citation validation stops both histories; it makes eight writer requests and no benchmark answers. Later full-history attempts cover six domain–policy configurations. An unconstrained run records 510 writer responses and six schema/reference failures. A subsequent constrained run records 4,030 memory-writer responses; all six histories stop at one incomplete JSON output each at the 2,048-token reserve. No configuration publishes a complete history or invokes benchmark reading. Its 1,266 supported policy–question rows remain missing and 87 policy–question rows require images. These attempts supply construction diagnostics, not memory-policy accuracy scores. Original failures, separate component-gate successes and execution details remain in the supplementary technical report.

Construction capacities and input work. CPU accounting on both complete histories preserves all 200 trajectories and 5,095 source states. The policies have unequal compression capacities: singleton, rebatch and trie end with 200, 26 and 38 leaves and note ceilings of 1,600, 208 and 304. Conditional published-order map-input workloads are 76.81, 2,011.07 and 457.35 million tokens, respectively. These are serialized-input counts assuming successful updates, excluding reducer work, generated outputs and failures. Sharing a leaf need not expose two sources in the same raw map: rebatch has 684 co-leaf source pairs but only 75 co-map pairs; trie has 469 and 70. Consequently, a common reader cap alone does not make these policies a controlled test of cross-source context. The supplement preserves full accounting, frontend settings and recorded failure costs.

F ONLINE EXECUTION AND PAIRED DIAGNOSTICS

Fixed scope and unit of analysis. The online comparison uses the first four analysis worlds per dataset in a fixed family-balanced ordering: synthetic static-02, temporal-02, static-03, temporal-03; revision-derived real-01 through real-04; and benchmark-derived bench-03 through bench-06. Each

method uses all source events and all original questions, published arrival order and one shuffle (recorded arrival-path seed 20270917), chunk size eight, and leaf/subbatch capacity 12. This path seed is distinct from the bootstrap seed 20260907 used for uncertainty estimation. Per-event, sorted-rebatch and trie each have 24 completed cells. The generator is Qwen3-4B at the same immutable revision as the matched-atom experiment, with temperature zero, seed zero, prefix caching disabled and one sequence per call. Extraction reserves 4,096 tokens; answering reserves 64. No generated artifact is shared across methods or paths. The canonical flat renderer and explicit recency instruction are fixed across policies; readout stays fixed while the extraction policy varies.

The 12-world scope was selected for runtime cost after an interrupted larger execution, using the fixed world ordering before the new outputs were generated. Sources had been studied previously. The comparison is exploratory; paths are repeated measurements, and world-equal accuracy differences are descriptive.

Coverage and raw-evidence reconstruction. All 72 cells complete with 3,732 extraction and 1,194 answer requests. There are 1,050 main answers (350 per policy), 144 fresh extraction probes, and 144 fresh answer probes. Every request record binds its identity, exact prompt bytes, first output, telemetry and recorded model tokens. An offline reconstruction invokes the frozen policy and rendering logic using those recorded outputs, reproducing every update, final atom set and result row. A separate normalization-and-exact-match implementation agrees on all 1,050 main scores. Reconstruction uses recorded outputs. Provider token counts are cross-checked against telemetry and run totals; full independent retokenization and hardware attestation are outside its scope.

All three policies have one final atom-set hash per world across the two arrival paths (12/12 for each). All 144 extraction probes reproduce both text and atom hashes, and all 144 answer probes reproduce text. These measurements establish agreement for the tested requests; the general contract still assumes fixed extraction outputs.

Table 9: Dataset breakdown of the same complete small execution. Each row averages four worlds equally. Counts pool both paths and are shown to make weighting explicit.

Dataset	Policy	World-equal accuracy (%)	Correct/answers
Synthetic	Per-event	34.38	44/128
	Sorted rebatch	34.38	44/128
	Canonical trie	39.06	50/128
Revision-derived	Per-event	42.81	54/126
	Sorted rebatch	47.60	60/126
	Canonical trie	44.27	56/126
Benchmark-derived	Per-event	31.25	30/96
	Sorted rebatch	33.33	32/96
	Canonical trie	33.33	32/96

Pooled main exact accuracy is 128/350 (36.57%), 136/350 (38.86%) and 138/350 (39.43%), respectively. These pooled rates differ from Table 4's world-equal rates because query counts differ. Main answers hit the length cap 10, 10 and six times; the frozen first-line score is retained, not replaced by a semantic judge or a selective rerun. The matched-atom budget diagnostic uses different prompts and therefore does not establish budget robustness for this comparison.

Observed cost. Per-event, rebatch and trie record 2,324, 1,122 and 1,480 total requests, including 398 answers each. Their input-token totals are 2,017,232, 2,017,935 and 2,173,056; generated-token totals are 127,598, 461,024 and 556,680. Summed request times are 2,219.840, 7,653.800 and 9,302.088 seconds. All repetition probes are charged, with no attempt to assign shared startup to one method. Model load takes 23.167 seconds; the request session including bookkeeping takes 19,256.651 seconds; the outer worker duration is 19,286.877 seconds (5.357 hours). Peak VRAM is not established by these records. The cost ordering persists when repetition probes are excluded, as detailed below.

Separate incomplete execution. The interrupted larger execution records 13,058 requests, 173 completed cells, 24 extraction-length failures and one request in an unfinished cell. Its 2,768

main answers are reconstructed separately and excluded from the complete 12-world comparison. Completion follows execution order and stop time: only eight synthetic worlds have completed or failed cells, with no completed revision-derived or benchmark-derived cells. Across four overlapping synthetic worlds, 2,180 matched occurrences have byte-identical prompts and outputs in distinct executions. This provides a limited repeatability observation; it does not constitute a second complete run or independent-source replication.

Post-hoc paired error and scoring audit. All 525 question–policy pairs have byte-identical raw answers across the two arrival paths. For descriptive pairing we therefore collapse to 175 distinct questions, nested within the same 12 worlds; this does not create 175 independent experimental units. Table 10 reports every existing stratum, rather than selecting favorable subgroups. Against per-event extraction, trie improves 19 questions and worsens 14. Its supersession gains and losses exactly cancel (eight each); improvements are not a uniform benefit on temporal questions. Against rebatching the corresponding totals are 12 and 11.

Table 10: Post-hoc breakdown of all 175 distinct questions. Correct counts collapse the two identical paths. "Gain/loss" counts questions where trie alone or the comparator alone is correct. No new model outputs or subgroup significance tests are introduced.

Stratum	Questions	Per-event correct	Rebatch correct	Trie correct	Trie gain/loss vs. Per-event	Trie gain/loss vs. Rebatch
As-of	33	15	18	16	2/1	1/3
Clean	35	20	24	22	6/4	1/3
Paraphrase	14	7	7	9	3/1	3/1
Supersession	75	22	19	22	8/8	7/4
Tie conflict	18	0	0	0	0/0	0/0
Total	175	64	68	69	19/14	12/11

Conflict accuracy is measured by the frozen first-line exact-output protocol. A deliberately narrow sensitivity check compares the whole single-line, pipe-separated value list while allowing omission of the `CONFLICT:` prefix; it permits no extra prose, fuzzy matching or truncation repair. It identifies one additional trie match and none for the comparators. The response is "dormant — under review" to synthetic static-02/q11, for which both required five-field tuples are present. The primary score stays unchanged. One conflict answer per path hits the length limit for both per-event and trie; rebatch has none. Longer-answer robustness and semantic conflict reasoning remain separate evaluation questions.

Evidence presence has a narrower meaning than semantic support. We independently reconstruct the recorded five-field tuple checks from the final atoms and source events. Only the 64 distinct synthetic questions explicitly declare all-required evidence semantics; the others are excluded from this diagnostic. For per-event, rebatch and trie, respectively, all required tuples match after lexical normalization on six, one and three questions; three, zero and one of those questions are answered incorrectly. Conversely, 19, 21 and 23 correct answers lack an all-tuple match. Attribute paraphrases and exact provenance bindings can break this lexical test. Absence therefore cannot be interpreted as lost semantic evidence, nor can these cross-sectional counts identify an extraction-versus-reader causal decomposition.

Repetition probes do not explain the cost ordering. Excluding the 96 probe requests per policy leaves 2,228, 1,026 and 1,384 primary requests, taking 2,149.081, 7,191.926 and 8,855.083 seconds for per-event, rebatch and trie. Trie remains 4.12 times per-event and 1.23 times rebatch request time. This is a sensitivity calculation on the same execution, not a measured rerun without probes; Table 4 retains its original all-request accounting.

G Post-hoc audit of all online answer disagreements

The audit includes every distinct question on which canonical trie and per-event extraction have different original exact scores: 19 gains and 14 losses. Both arrival paths give byte-identical answers

within each policy, so each question is inspected once. No case is removed after inspection. The analysis uses the original query and gold, all source events, full final atoms, and the recorded first answer prompt and output. For all 66 policy–question records, prompt hashes and raw answers match the corresponding primary request. The supplement provides these bindings and the complete ledger.

Annotation protocol and limits. The categories were developed after observing the disagreements and were assigned with AI assistance, with method identities and outcomes visible. They are not blinded, independently human-validated semantic scores. "Surface-compatible" identifies clear cases where both responses express the same requested value through a different date format, list order, or answer wrapper. "Extraction-change" records a specific changed relation argument or omitted/added field in source-linked atoms; it does not attribute the answer difference to that change. Remaining cases are marked "mixed-or-readout." Surface compatibility takes priority; ambiguous entity aliases are not treated as equivalence. Thus Microsoft versus Microsoft Corporation remains mixed, and an answer containing the gold plus an additional award is not classified as formatting alone. Counts are descriptive over selected disagreements, not population error rates.

Table 11: Exhaustive descriptive annotation of the 33 exact-score disagreements. No original answer, parser, or score is replaced.

Category	Trie gain	Trie loss	Total
Surface-compatible answers	4	2	6
Specific source-linked extraction change	2	2	4
Mixed context or readout difference	13	10	23

The six surface cases are benchmark-derived bench-06/q010 (the broadcaster appears on the second line), revision-derived real-01/q003 (award-list order), real-02/q010 (an extra answer label), real-02/q016 (ISO versus written date), real-03/q001 (university-list order), and synthetic static-03/q02 (a sentence containing the correct city). The four extraction-change cases are bench-04/q007 (a lost institution argument), real-02/q014 (successor direction), real-04/q006 (omitted study field), and temporal-02/q05 (a spurious value suffix in one paraphrase). Source identifiers and exact strings are preserved in the ledger.

Implication for the quality claim. The cases support a narrower interpretation than uniform improvement in contextual understanding. Some gains coexist with the same target fact in both states; a gain can also arise through compliance with a literal output format. Conversely, a contextual extraction can remove a field, reverse a relation, or leave the reader choosing an older value. Other atoms, attribute names, ordering, and total prompt context can differ simultaneously. None of these observations isolates a causal benefit of cross-event extraction, and no semantic-adjusted accuracy or new significance test is reported. The original online comparison and its quality–cost limitation remain unchanged.