What Does Your Agent Remember? Auditing Storage, Reachability, and Behavioral Influence

Anonymous authors
Paper under double-blind review

Abstract

Information stored in an agent's memory may no longer support its later answers. We operationalize an audit that links durable records, active context, and behavior through native interventions and matched controls, returning explicit indeterminate verdicts when evidence or controls are unavailable. Across six agent–model configurations and 3,600 declared core executions, native Pi compaction reduces target adoption relative to restart by 20.8 points with DeepSeek V4 Flash and 15.0 points with GLM-5.3 Flash. Every success-to-failure pair in those contrasts still contains the complete source history in pre-probe durable records. Among the 25 GLM-5.3 Flash cases that succeed after restart and fail after compaction, seven, all recommendation tasks, still carry the literal target in active context. The procedure distinguishes preservation, inclusion, and use without treating any single surface as evidence of all three.

1 Introduction

Persistent memory increasingly functions as part of the control plane of stateful agents. Systems preserve conversations, checkpoints, summaries, profiles, tool traces, and compacted histories so that later actions can depend on earlier interactions; recent work treats memory construction, hierarchical representation, retrieval, compression, and control as first-class system decisions (Hu et al., 2026; Wang et al., 2026b), while memory-security work traces the resulting lifecycle surface from writing through forgetting (Lin et al., 2026). This continuity creates a measurement problem: a source may remain on disk but never reach the model, or disappear literally while a transformed descendant continues to affect behavior. Most evaluations observe only one surface of that pathway. String scans test retention, retrieval benchmarks test reachability, and task answers test behavior, but none alone establishes the other two.

We ask a basic question: where does preserved memory stop supporting a later answer? Our central claim is that an audit must trace durable retention, active inclusion, and controlled behavior separately. Compaction provides a concrete test: a session archive can preserve the source while the context assembled from it no longer supports the same behavior. Locating that separation requires case-linked evidence, not an aggregate memory score.

We construct a frozen 120-case system evaluation from provenance-pinned Memora histories (Uddin et al., 2026), adding deterministic counterfactuals, native execution conditions, and traceable evidence. The study declares 3,600 core executions, 960 pathway-matched controls, and two separate 120-execution follow-ups: a repeat and a marker-position ablation. A 48-execution wording test and 96-execution, 24-case LongMemEval oracle study probe wording sensitivity and transfer beyond Memora. Failures remain separate from recovery, unsupported outcomes remain indeterminate, and aggregates trace to case, execution, and native-evidence hashes.

We make three connected contributions:

1. An explicit auditing procedure connects native interventions to independent evidence views, with a common adapter contract and conservative control-validity rules.

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.

2. Case-linked Pi captures show complete source preservation in all 480 restart and compaction durable records, including every success-to-failure pair, and distinguish literal omission from retained targets without adoption; task-specific analyses bound what each observation establishes.
3. Cross-harness disagreement and pathway-matched controls demonstrate why the procedure needs separate views and an indeterminate outcome; repeats bound the stability of its findings.

Compaction is the main demonstration of the method. Audit disagreement explains what a single view misses; controls delimit attribution rather than constitute a separate memory score.

2 A procedure for tracing memory use

Let a source memory m produce descendants through summarization, profiling, embedding, tool use, checkpointing, or compaction. A visible update to m need not invalidate every descendant. We represent these relations by a graph $G_m = (N, E)$ rooted at m; an audit observes only a declared subset of nodes and transitions, so its conclusion must stop at the strongest supported evidence level.

Inputs, procedure, and output. The input is a source history, query, accepted target projection, counterfactual update, and declared system boundary. (1) The adapter starts an isolated session and executes a native condition. (2) It binds pre-probe durable records, the active provider request where available, and the answer to one case and run. (3) Independent views classify their own evidence. (4) No-memory and pathway-matched altered controls determine whether behavioral attribution is supported. (5) The output is a case-level evidence vector, control-validity record, and bounded diagnosis. Each adapter must preserve session identity, process boundaries, attempts, and evidence hashes; unavailable surfaces are explicit. Causal influence is inferred from matched executions, not a fourth physical store.

2.1 Audit views and evidence levels

We use evidence level for the strength of an inference and audit view for an operational diagnostic. E1 establishes a durable artifact, E2 attributes it to a source, E3 observes active inclusion, E4 localizes behavioral influence, and E5 tests non-revival over a declared budget. These levels order claim strength, but an interface may expose a higher-level observation while leaving a lower-level surface unavailable. This study reaches E1–E4 on different interfaces and interventions; it makes no E5 claim.

Table 1 defines four audit views, each three-valued: retained, not_detected, or indeterminate. The views are projections onto the ladder over observable subsets of N and E, not a second set of levels: storage-only can establish E1 and, with lineage, E2; active-only observes E3; answer-only records behavior without causal attribution; and matched causal can establish E4 for a declared pathway. No view is treated as a gold standard.

Table 1: Operational audit views and their claim boundaries.

View	Permitted evidence and claim boundary
Storage-only	Exact normalized target in hash-verified durable records. Absence means not_detected, never erasure.
Active-only	Target present in the actual provider request or native retrieved-memory evidence. Missing capture is indeterminate.
Answer-only	Frozen scorer finds target adoption in the final answer. It does not establish memory use.
Matched causal	Durable/active/behavioral vector plus matched no-memory and altered-memory controls. Missing or invalid controls remain indeterminate.

This ladder prevents two common errors. Durable residue is a false positive if interpreted as behavioral failure, while a correct output is a false attribution if the model could answer without memory. For a probe q, a causal source effect requires target adoption after source ingestion relative to behavior without injected memory. The altered-memory control then checks whether the same treatment can redirect the answer; if it fails, correct source answers cannot be cleanly attributed to the declared memory pathway.

2.2 Deletion and forgetting are boundary-relative

Logical deletion makes an item unavailable through a supported interface, whereas behavioral erasure removes its attributable effect over a declared probe distribution and system boundary. Global irrecoverability is stronger: it would require showing that no relevant representation remains in local state, remote services, backups, or model parameters. The evidence levels therefore separate inferential strength from desired policy.

An E1 finding can coexist with an E4 pass: an append-only event may preserve a deleted literal while the runtime excludes it and behavior returns to the no-memory control. Conversely, literal absence can coexist with E4 residue because a summary or profile may preserve the source's operational effect without repeating its original words. This study emphasizes E1–E4 and does not claim E5 non-revival beyond the frozen probes, restarts, and compaction interventions.

2.3 Relation to prior memory-audit work

Prior work has organized agent-memory failures around lineage, transformation, retrieval, correction, deletion, and propagation, while related work on personal-memory auditability describes lineage, representation, and effect gaps in long-lived agents (Osmond et al., 2026; Osmond, 2026). Appendix B restates a previously reported checkpoint experiment for context, not counted among this paper's new evidence; the population study, pathway-matched extension, and disagreement analysis are new.

3 Cross-harness study

3.1 Cases and conditions

The population contains 120 provenance-pinned Memora adaptations spanning ten personas: 50 remembering, 50 recommending, and 20 arithmetic or comparative reasoning cases. Each case contains a source history, a query, an accepted target projection, a deterministic altered source, and operation metadata; the case is the statistical unit. The frozen population covers deletion- and correction-like longitudinal updates across those personas, with selection completed before model outcomes were inspected.

Reference projections are machine-derived from the upstream cases, with deterministic arithmetic targets independently recomputed. They are not an adjudicated human gold set; Section 6 limits claims accordingly.

Table 2 defines the five core conditions and the two frozen pathway-matched extensions. The latter run on Pi × DeepSeek V4 Flash, Pi × GLM-5.3 Flash, Codex × GPT-5.6 Luna, and Codex × GPT-5.5.

Table 2: Declared execution conditions. "Supported" counts configurations exposing the required native operation.

Condition	Procedure	Supported
probe_only	Fresh isolated session without source memory.	6
source_context	Ingest source and probe in one live session.	6
source_restart	Ingest, stop the runtime, resume the exact durable session, then probe.	6
compacted_restart	Invoke native compaction, restart, resume, and probe.	4
altered_memory	Ingest a deterministic counterfactual source and marker, then probe.	6
Frozen pathway-matched extension		
altered_source_restart	Ingest source, then its deterministic counterfactual update; restart and probe for the marker.	4
altered_compacted_restart	Apply the same update, invoke native compaction, restart, and probe for the marker.	4

The extension declares $120 \times 2 \times 4 = 960$ additional executions without redefining the frozen core matrix. OpenCode is excluded because the tested interface has no native compaction operation: unsupported compaction is never emulated, so those cells remain explicit unsupported outcomes rather than receiving direct prompt assembly or database editing. The matched design has two roles: source versus no-memory estimates whether the treatment carries useful information, while altered memory tests whether the same pathway can redirect the answer. Restart tests durable-session continuity; compacted restart changes one additional mechanism, but only within a harness whose native operation is observable and successfully completed.

3.2 Systems and native evidence

Table 3 lists the six configurations. Each cell starts from an isolated agent directory and fresh session identity.

Table 3: Core-population execution accounting. All unsupported cells are unavailable native-compaction operations; terminal failures remain separate.

Harness	Model	Declared	Supported	Done	Failed	Unsup.	Ext.
Pi	DeepSeek V4 Flash	600	600	600	0	0	yes
Pi	GLM-5.3 Flash	600	600	600	0	0	yes
OpenCode	DeepSeek V4 Flash	600	480	480	0	120	–
OpenCode	GLM-5.3	600	480	441	39	120	–
Codex	GPT-5.6 Luna	600	600	600	0	0	yes
Codex	GPT-5.5	600	600	600	0	0	yes
Total		3600	3360	3321	39	240	4/6

Pi (version 0.84.2) exposes the active provider request and native compaction. Restart uses the exact native session identifier in a fresh process; compaction is explicitly invoked, and the resulting provider request is captured before scoring.

OpenCode (version 1.15.5) exposes durable SQLite state but no native compaction or provider-request interception. Source and probe events are verified against authoritative SQLite part data; restart requires distinct processes and exact session continuation, with tools and workspace-path escape prohibited. It contributes 112/308 jointly determinate storage–causal comparisons despite unsupported compaction; Pi supplies none under the current extraction pipeline. Pi × GLM-5.3 Flash and OpenCode × GLM-5.3 are different recorded endpoints, not alternate names for one model.

Codex runs isolated durable threads with global memory disabled. We retain complete JSON-RPC transcripts, native rollouts, thread identifiers, process boundaries, and pre/post SQLite snapshots; compaction over native RPC must emit the request, completion item, and completed turn before restart. The runtime reports exact model slugs and low reasoning effort, while temperature, provider pre-dispatch capture, plaintext compaction summaries, and hosted serving snapshots are unavailable.

3.3 Scoring, uncertainty, and integrity

The frozen answer scorer checks normalized accepted projections and explicit contradiction patterns; parse uncertainty, failed calls, and unsupported operations stay indeterminate. Paired contrasts use cases as the inferential unit, with deterministic 10,000-draw case-clustered bootstrap intervals. Repeats on a frozen 30-case subset measure same-cell stability, not a larger semantic population.

Six LLM case-quality passes test rubric sensitivity, not core answer correctness (Appendix F); the four v3.2 passes agree on 97.5–99.2% of statuses. A targeted human check has 39/48 response agreement ($\kappa = .631$); reference disagreements remain unadjudicated (Table 15, Appendix L). The external follow-up uses two model output judges (Appendix K).

Each native root records case bytes, configuration, runtime identity, and attempts; its manifest lists every included artifact and expected SHA-256 digest. Credential scans and independent verifiers run before freezing, and the final loader re-verifies all three harness freezes before and after analysis. The disagreement study accepts only the exact six configurations and 3,600 declared core rows; the pathway-matched extension has a separate manifest and analyzer.

The four audit views are computed independently, and a view cannot read evidence assigned to another view. For source_context, causal retention requires target adoption with source memory, target rejection without memory, and counterfactual-marker adoption under altered memory. The extension applies the same conservative rule to restart and compaction using their pathway-matched altered controls; conditions without a valid matched control remain causal-indeterminate. Every row retains its case hash plus evidence-record identifier and hash.

4 Results

4.1 Compaction separates durable preservation from use

Table 4 reports source-target adoption. Indeterminate outcomes remain visible; compacted OpenCode cells are unsupported.

Table 4: Answer-only source-target adoption/indeterminate outcomes among 120 cases. U denotes an unsupported native operation.

Configuration	None	Context	Restart	Compact	Altered
Pi × DeepSeek V4 Flash	1/0	89/0	86/0	61/0	0/0
Pi × GLM-5.3 Flash	2/0	96/0	96/0	78/0	7/0
OpenCode × DeepSeek V4 Flash	0/10	89/9	86/8	U	6/2
OpenCode × GLM-5.3	0/3	87/18	88/15	U	69/25
Codex × GPT-5.6 Luna	0/0	85/0	85/0	76/0	31/0
Codex × GPT-5.5	0/0	107/0	103/0	102/0	0/0

Pi shows model-dependent compaction changes. With DeepSeek V4 Flash, target adoption falls from 86/120 after restart to 61/120 after explicit compaction, a paired change of -0.208 $[-0.300, -0.117]$; with GLM-5.3 Flash, it falls from 96/120 to 78/120, or -0.150 $[-0.242, -0.067]$. The pooled Codex contrast is -0.042 $[-0.088, 0.004]$. Given the interval and repeat instability reported below, we do not detect a stable Codex compaction effect. OpenCode cannot answer the compaction question because the tested interface exposes no native operation.

Task strata matter. With Pi × DeepSeek V4 Flash, compaction changes remembering by −0.22 and recommending by −0.32, while direct-context reasoning succeeds in only 3/20 cases. With Pi × GLM-5.3 Flash, the −0.60 reasoning change is template-concentrated: quarterly food totals fall from 9/10 after restart to 0/10 after compaction, while maximum-month comparisons fall from 10/10 to 7/10. Appendix J gives the full task-family breakdown.

Durable preservation and active-context retention. A separate case-linked analysis reads pre-probe native durable records, excluding the final answer. All 480 Pi restart/compaction records preserve the complete source history, including all 31 DeepSeek V4 Flash and 25 GLM-5.3 Flash success-to-failure pairs. Six DeepSeek V4 Flash and seven GLM-5.3 Flash cases move in the opposite direction, yielding net losses of 25 and 18 cases, respectively. The literal target disappears from active context in 24 DeepSeek losses. Table 5 separates the GLM losses: a retained target need not be adopted, while a computed answer need not appear literally among its inputs. This follow-up literal check is distinct from the core active-view scorer.

Table 5: Literal-target status in 25 Pi × GLM-5.3 Flash success-to-failure pairs. All preserve the complete source history in pre-probe durable records.

Restart → compacted	Cases	Interpretation and task scope
Present → absent	6	Target omitted (2 remembering, 4 recommending).
Present → present	7	Presence need not imply adoption (all recommending).
Absent → absent	12	Computed answer need not appear literally (all reasoning).

For reasoning, all 40 restart requests retain the complete exact dated source ledger, versus one of 40 compacted requests. The remaining 39 are uncertified, not proven insufficient: equivalent aggregates or representations could suffice. The certified compacted DeepSeek V4 Flash case fails at both restart and compaction, not because of an observed compaction-induced change. Appendix I specifies this sufficient-input certificate and worked case.

Captured GLM-5.3 Flash summaries retain only 9–18 currency values from quarterly contexts containing 276–308, and 2–5 from maximum-month contexts containing 44–134. None of 12 baseline-valid reasoning flips retains the literal target, versus 2/7 non-flips. DeepSeek V4 Flash fails 19/20 restart reasoning baselines, leaving no valid flip population. These are task-specific observations rather than a universal information-loss explanation.

Altered-memory controls also expose model sensitivity. GPT-5.5 adopts no original targets under alteration, whereas OpenCode × GLM-5.3 retains 69 and GPT-5.6 Luna retains 31. Those cells cannot support the same causal interpretation, so the protocol leaves causal verdicts indeterminate when the altered treatment fails to redirect behavior or the no-memory control already adopts the target.

Equal DeepSeek V4 Flash context/restart aggregates on Pi and OpenCode conceal different task-specific outcomes (Appendix J). The cross-harness comparison is descriptive; the compaction contrasts are paired within harness and do not identify an isolated harness effect.

4.2 Matched controls bound the attribution

The extension completed all 960 declared executions in one attempt each. Table 6 carries the deterministic altered-content marker through restart and native compaction, rather than testing alteration only in a live source-context session.

Here Δ orig. is the change in original-target adoption from source restart to compacted restart, and Δ alt. is the corresponding change in counterfactual-marker adoption.

Codex supplies a contrasting primary control. GPT-5.6 Luna rises from 78/120 to 110/120 under compaction even while original-target adoption falls by 9/120; the interaction is +0.342 [.217, .458]. On the frozen 30-case repeat, marker adoption still rises from 24/30

Table 6: Pathway-matched extension. "Causal det." counts compacted cases with valid no-memory and altered controls.

Configuration	Alt. restart	Alt. compact	Δ orig.	Δ alt.	Causal det.
Pi × DeepSeek V4 Flash	119/120	109/120	−.208	−.083	108/120
Pi × GLM-5.3 Flash	106/120	85/120	−.150	−.175	84/120
Codex × GPT-5.6 Luna	78/120	110/120	−.075	+.267	110/120
Codex × GPT-5.5	120/120	120/120	−.008	.000	120/120

to 27/30, but the original contrast is +1/30 and the interaction attenuates from +13/30 in the primary subset to +2/30 [−0.133, .267]. GPT-5.5 adopts the marker in all 60 repeat cells and again has no marker contrast. The GPT-5.6 Luna direction is therefore a draw-sensitive counterexample, not a stable estimate of reversal magnitude.

A separate 120-execution GPT-5.6 Luna positional ablation does not resolve an early-versus-late marker interaction (Appendix J); it does not establish recency invariance.

For Pi × DeepSeek V4 Flash, altered-marker adoption falls by −0.083 [−0.142, −0.033], less than the original-target change; the paired interaction is +0.125 [.017, .242]. For Pi × GLM-5.3 Flash, marker adoption falls by −0.175 [−0.275, −0.075], with interaction −0.025 [−0.158, .108]. The marker declines through Pi's pathway in both configurations, but the interaction does not isolate representation from difficulty or ceiling effects. Appendix J reports a post-hoc matched-baseline sensitivity with its selection caveat.

Literal summary evidence narrows the Pi difference. The deterministic marker remains verbatim in 118/120 DeepSeek V4 Flash and 120/120 GLM-5.3 Flash compaction summaries, yet is adopted in 109/120 and 85/120 answers, respectively. For GLM-5.3 Flash, the 35 retained-but-unadopted markers are exactly the 35 failed altered positive controls; one additional negative-control issue brings that configuration to 36 causal-indeterminate cases. This directly separates active reachability from behavioral influence. Marker omission therefore cannot explain the model gap by itself; the configurations differ in how they use retained content.

The 58/480 compacted causal-indeterminate cases are fully accounted for: 56 fail the altered positive control and three fail the no-memory negative control, with one overlap and no missing outcome. They concentrate in Pi × GLM-5.3 Flash (36), followed by Pi × DeepSeek V4 Flash (12), Codex × GPT-5.6 Luna (10), and Codex × GPT-5.5 (0); the protocol does not convert these failed controls into negative findings.

4.3 Single-view audits miss the pathway distinctions

Across all executions, storage-only yields 1,263 retained, 858 not-detected, and 1,479 indeterminate verdicts, while active-only yields 760, 440, and 2,400. Answer-only yields 1,530 retained, 1,740 not-detected, and 330 indeterminate; the matched causal view yields 426, 105, and 3,069, because 133 completed source-context executions become causal-indeterminate when a positive or negative control is invalid.

Table 7 first reports disagreement restricted to jointly determinate pairs, then all-row, three-valued divergence. The latter treats a determinate verdict paired with indeterminate as different; it measures conclusion availability as well as diagnosis, not accuracy.

The storage–active pair has no jointly determinate execution. Its .923 all-row rate is therefore entirely an observability result, not evidence that two available diagnostics semantically contradict: the harnesses expose different surfaces. No tested configuration supplied the complete four-view evidence vector under the present instrumentation. Pi's storage view reads only a normalized batch summary – an extraction limitation – while OpenCode and Codex expose no provider request, an interface one. Pi supplies all 223 valid active–causal comparisons but no storage–causal comparison; OpenCode and Codex supply all 308 storage–causal comparisons but no active evidence. This is a result about the evidence supplied under the present instrumentation, not missing-at-random data. Reporting those

Table 7: Pairwise audit-view divergence. Jointly determinate comparisons precede all-row rates, which include both availability and verdict differences. Intervals on all-row rates cluster by case.

Pair	Joint det.	Diff./joint	All-row rate [95%]
Storage vs active	0	–	.923 [.919, .926]
Active vs answer	1200	328/1200 (.273)	.666 [.653, .679]
Answer vs causal	531	0/531 (.000)	.761 [.754, .768]
Storage vs causal	308	89/308 (.289)	.590 [.584, .597]
Storage vs answer	2070	589/2070 (.285)	.511 [.494, .528]
Active vs causal	223	40/223 (.179)	.368 [.363, .374]

cells as clean would manufacture evidence; their correct status is indeterminate, which is itself actionable because the audit as instrumented cannot support the requested claim.

Determinate disagreements remain important. Storage–causal disagreement is heterogeneous: Codex × GPT-5.5 contributes 32/120, Codex × GPT-5.6 Luna 22/76, OpenCode × DeepSeek V4 Flash 31/95, and OpenCode × GLM-5.3 4/17, totaling 89/308 (28.9%). GPT-5.5's 32 split into 24 target adoptions without a literal match and eight literal matches without attributable adoption. Active–causal disagreement is 40/223 (17.9%). Valid answer–causal pairs agree by construction, so 0/531 is not an independent agreement result.

Control semantics materially change the diagnosis. Under the rejected alternative that labels completed but invalid controls as not_detected, answer–causal becomes 103/664 (15.5%) and storage–causal becomes 192/424 (45.3%). The primary analysis instead preserves those failed controls as indeterminate, reducing the latter by 16.4 points. Thus "the agent remembers" is not one measurable proposition but a vector whose components can diverge within the same execution.

4.4 Repeatability and failure sensitivity

The frozen repeats preserve Pi's negative direction: the 30-case DeepSeek V4 Flash contrast remains −5/30, while GLM-5.3 Flash changes from −9/30 to −12/30. Codex contrasts change sign or approach zero, so neither the primary GPT-5.6 Luna interaction nor Codex non-detection establishes a stable endpoint effect. Appendix E reports condition-level agreement. OpenCode × GLM-5.3 retains all 39 primary failures; its separate 300-second sensitivity recovers 22 without replacing them.

4.5 Targeted wording and external-task checks

On 12 frozen reasoning cases, Pi × GLM-5.3 Flash completes a 2 × 2 ingestion-wording/pathway test. Original wording yields 9/12 correct after restart and 1/12 after compaction; neutral acknowledgment wording yields 10/12 and 3/12. The paired interaction is $+0.083$ $[-0.167, .333]$: substantial losses remain, but the effect of wording is unresolved.

A separate 24-case LongMemEval oracle sample (Wu et al., 2024) supplies eight multi-session, eight temporal, and eight update tasks. Of 96 declared Pi executions across DeepSeek V4 Flash and GLM-5.3 Flash, 95 complete; one interrupted compaction remains unknown. DeepSeek V4 Flash answers decline under both blinded judges, including the cross-family GLM-5.3 Flash judge (-0.565 on 23 known pairs); the same-family judge scores 21/24 after restart versus 12/24 after compaction (-0.375 $[-0.583, -0.167]$). The smaller GLM-5.3 Flash answer-family contrast is uncertain: the cross-family DeepSeek V4 Flash judge estimates -0.130 $[-0.261, 0]$ on 23 known pairs. Complete-case estimates are descriptive, not full-sample means. Appendix K reports both judges and separate recovery/parser sensitivities. This small oracle-only extension does not establish generality or human-validated accuracy.

5 Related work

Memory systems optimize construction and control (Xiong et al., 2026; Wang et al., 2026c; Hu et al., 2026; Wang et al., 2026b), extending retrieval-augmented generation (Lewis et al., 2020) with lifecycle boundaries. MemoryArena tests interdependent sessions, Memora obsolete information, and ForgetEval mutation/retrieval (He et al., 2026; Uddin et al., 2026; Yang, 2026). LongMemEval benchmarks interactive memory and analyzes indexing, retrieval, and reading (Wu et al., 2024). Our distinction is not the staging alone, but native evidence capture and matched interventions connecting durable state, active inclusion, and behavioral influence.

Security, deletion, and unlearning research addresses memory lifecycle risks (Lin et al., 2026; Wang et al., 2026a; Ramesh, 2026b;a); editing evaluations distinguish updates from downstream effects (Cohen et al., 2024; Zhong et al., 2023; Thudi et al., 2022; Deepak et al., 2025). We apply this distinction to composed agent state, not a new memory architecture.

6 Limitations and ethics

The main population comes from Memora; the external follow-up covers only LongMemEval's oracle setting. Human review covers 24 case-model pairs, including discordant and concordant outcomes, not population-wide semantic validity; reference disputes remain unadjudicated. Absolute answer and causal accuracy remains provisional. External judges share answer-model families and potentially errors; their answers lack human review. Core Pi contrasts may inherit condition-dependent scoring error: negated or hedged targets count as adoption unless an explicit contradiction pattern matches. Appendix G audits lexical failures. Cases are synthetic adaptations, not private histories.

The configurations are not fully crossed: only one model family is shared across Pi and OpenCode. Pi's storage indeterminacy reflects extraction from normalized summaries rather than available raw records; OpenCode and Codex lack interface access to provider pre-dispatch capture. These differences make Table 7 unsuitable as a leaderboard. Pathway-matched controls cover Pi and Codex restart and compaction, yielding 422/480 determinate compacted causal verdicts. OpenCode lacks native compaction; the marker localizes influence to a pathway, not a specific summary element. Deliberately invoked compaction measures intervention effects, not deployment frequency.

Hosted endpoints lack weight hashes; we freeze model/provider IDs, dates, versions, prompts, observable requests, outputs, and state hashes. Repeats cannot separate provider, decoding, and runtime variation; intervals condition on each draw. Pi's direction, not magnitude, is stable; GPT-5.6 Luna's repeat interaction includes zero. Artifacts are path-anonymized and credential-scanned; no global or legal claim is made.

7 Conclusion

What an agent remembers depends on where one looks: no tested configuration supplied all four views under the present instrumentation. Pi's source history remains durable in every success-to-failure pair, while active-context and reasoning-input checks locate different boundaries. The external oracle study extends the task setting, not the generality claim; wording sensitivity remains unresolved. Our procedure connects preservation, availability, and controlled behavior without equating them. Reliable audits must retain indeterminate outcomes when evidence or controls fail. The practical aim is to identify where memory stops supporting the task, rather than equate stored information with successful use.

Reproducibility statement

The anonymous supplement supplies frozen cases, adapters, analyses, regression tests, the schema-1.1 report, and per-case extension evidence. Runtime boundaries and evidence records (Section 3.2, Appendix H) link aggregates to case and native-evidence hashes.

AI use statement

The authors conceived and directed the study, approved its conditions, controls, and scoring rules, and made the final decisions on claims and presentation. Generative AI tools assisted with experimental-design feedback, benchmark-case adaptation, software implementation, test and validation scripts, literature retrieval, interpretation of results, and manuscript drafting and editing. The authors evaluated this assistance and reviewed supporting code and analyses executed against frozen artifacts. The authors take responsibility for the final text, claims, and artifacts.

References

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. In Transactions of the Association for Computational Linguistics, volume 12, pp. 283–298, 2024. URL https://aclanthology.org/2024.tacl-1.16/.

Advit Deepak, Megan Mou, Jing Huang, and Diyi Yang. Identifying unlearned data in LLMs via membership inference attacks. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.emnlp-main.551/.

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

Zhanghao Hu, Qinglin Zhu, Runcong Zhao, Di Liang, Hanqi Yan, Yulan He, and Lin Gui. Beyond RAG for agent memory: Retrieval by decoupling and aggregation. arXiv preprint arXiv:2602.02007, 2026. URL https://arxiv.org/abs/2602.02007.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Advances in Neural Information Processing Systems, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Zehao Lin, Xixuan Hao, Renyu Fu, Shaobo Cui, Kai Chen, Chunyu Li, Zhiyu Li, and Feiyu Xiong. A survey on long-term memory security in LLM agents: Attacks, defenses, and governance across the memory lifecycle. arXiv preprint arXiv:2604.16548, 2026. URL https://arxiv.org/abs/2604.16548. Version 2, revised 11 June 2026.

Marcel Osmond. Auditable personal memory for user-adaptive agents. In PILA '26 Workshop, 2026. Workshop paper.

Marcel Osmond, Tonghao Chen, Zhihang Tan, and Kenzhebaev Bektur. Auditable memory as a safety requirement for agentic AI. In 2nd Workshop on Safe AI at UAI, 2026. Accepted workshop paper.

Vishwajith Ramesh. Forgetful attention: An auditable support-vector memory for selective retention and verified deletion, 2026a. URL https://arxiv.org/abs/2607.12204. arXiv:2607.12204v1, July 2026.

Vishwajith Ramesh. Subtract, transport, or replay? auditable deletion from language-model memory, 2026b. URL https://arxiv.org/abs/2607.27539. arXiv:2607.27539v1, July 2026.

Anvith Thudi, Hengrui Jia, Ilia Shumailov, and Nicolas Papernot. On the necessity of auditable algorithmic definitions for machine unlearning. In 31st USENIX Security Symposium, pp. 4007–4022. USENIX Association, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/thudi.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. arXiv preprint arXiv:2604.20006, 2026. URL https://arxiv.org/abs/2604.20006. Accepted at ACL 2026.

Bin Wang, Fan Wang, Pingping Wang, Jinyu Cong, Yang Yu, Yilong Yin, Zhongyi Han, and Benzheng Wei. Agentic unlearning: When LLM agent meets machine unlearning. arXiv preprint arXiv:2602.17692, 2026a. URL https://arxiv.org/abs/2602.17692.

Xinyu Wang, Mingze Li, Peng Lu, Xiao-Wen Chang, Lifeng Shang, Jinping Li, Fei Mi, Prasanna Parthasarathi, and Yufei Cui. InfMem: Learning system-2 memory control for long-context agent. arXiv preprint arXiv:2602.02704, 2026b. URL https://arxiv.org/abs/2602.02704.

Zhenting Wang, Huancheng Chen, Jiayun Wang, and Wei Wei. Memex(RL): Scaling long-horizon LLM agents via indexed experience memory. arXiv preprint arXiv:2603.04257, 2026c. URL https://arxiv.org/abs/2603.04257.

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

Yiming Xiong, Shengran Hu, and Jeff Clune. Learning to continually learn via meta-learning agentic memory designs. arXiv preprint arXiv:2602.07755, 2026. URL https://arxiv.org/abs/2602.07755.

Dongxu Yang. Control-plane placement shapes forgetting: An architectural study of agent memory across thirteen system configurations. arXiv preprint arXiv:2606.15903, 2026. URL https://arxiv.org/abs/2606.15903.

Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. MQuAKE: Assessing knowledge editing in language models via multi-hop questions. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 15686–15702. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.971. URL https://aclanthology.org/2023.emnlp-main.971/.

A Full counterfactual condition family

The cross-harness study uses five implementable conditions. The broader protocol distinguishes seven numbered states plus an authoritative post-update reference.

The five implementable conditions map onto that family as follows: probe_only is C_0, source_context is C_1, source_restart is a variant of C_3 that resumes the exact durable session rather than starting a fresh one, compacted_restart approximates C_4, with the harness rather than the auditor selecting which descendants survive, and altered_memory is C_2 followed by C_1 on the corrected content. The pathway-matched extension applies the same C_2-then-C_1 pair through the C_3 and C_4 pathways. C_5, C_6, and C_{ref} are not instrumentable on the tested interfaces.

Table 8: General counterfactual condition family.

ID	Condition
C_0	Never injected; estimates model and task priors.
C_1	Injected and active; verifies that memory can change behavior.
C_2	User-visible deletion or correction.
C_3	Fresh session in the same workspace.
C_{ref}	Authoritative reconstruction from post-update visible state.
C_4	Targeted descendant invalidation.
C_5	Complete purge and reconstruction.
C_6	Artifact retained but retrieval blocked, when instrumentable.

B Controlled checkpoint anchor

A previously reported checkpoint study combined Qwen3-14B with a LangGraph/LangChain agent, a file-backed visible store, and a SQLite thread checkpointer (Osmond et al., 2026).

Each case transformed a neutral synthetic fact into an operational preference, then deleted or corrected the visible record. The checkpoint retained a paraphrased derivative rather than the original string.

The visible store passed all 20 exact audits. The retained thread chose the pre-update action in 17/20 cases, while checkpoint purge and rebuild restored the expected action in 20/20.

The reported intervention reaches E4 for one declared checkpoint pathway. It is included for context, not as a prevalence estimate or new cross-harness evidence.

C Prompt-construction and Mem0 studies

A separate prompt-construction study assembled visible and derivative memory surfaces directly in prompts (Osmond et al., 2026). Five model–provider configurations completed 400/400 cells over 16 cases.

It validates condition discrimination and model sensitivity, not memory-system persistence. It remains separate from the native-state cross-harness study.

An auxiliary Mem0 2.0.18 study used Chroma local storage, eight cases per operation, three probe-model families, and three repetitions. It compared deletion, native update, and conversational add; its results are not counted in the new cross-harness matrix.

Supported deletion and native update removed the old target from retrieval in 0/8 residual cases. Conversational add left it retrievable in 8/8, including after restart.

The old literal remained in raw storage in 8/8 cases for all three paths. Behavioral old-target rates stayed near baseline, separating E1 retention from E2 reachability and E4 influence.

D Additional cross-harness reporting

Pi × DeepSeek V4 Flash and Pi × GLM-5.3 Flash each completed 600/600 primary cells and 150/150 repeat cells. Seven DeepSeek V4 Flash primary cells were retried with every failed or interrupted attempt preserved; GLM-5.3 Flash completed in first attempts.

OpenCode × DeepSeek V4 Flash completed 480/600 native cells; all 120 compaction cells were unsupported. Four historical failed attempts were retained before successful retry.

OpenCode × GLM-5.3 completed 441/600 primary cells, with 39 failed and 120 unsupported. Its repeat contained 99 completed, 21 failed, and 30 unsupported outcomes.

The Codex primary completed 1,200/1,200 cells in 1,200 first attempts. The repeat completed 300/300 in 300 first attempts, with unique isolated homes and durable threads.

All final audit rows retain case hashes and evidence-record identifiers. The anonymous supplement contains report-level JSON/CSV plus verifiers; large raw roots remain separately packaged.

E Repeatability detail

Repeat agreement is reported on the frozen 30-case subset, with pairs having an indeterminate side excluded from the observed count.

Ordinary compaction agreement is 20/30 for GPT-5.6 Luna and 25/30 for GPT-5.5. On the separate altered-path repeat, GPT-5.6 Luna agrees on 20/30 altered-restart and 29/30 altered-compacted cells; GPT-5.5 agrees on 60/60. Ordinary compaction contrasts change from −7/30 to +1/30 for GPT-5.6 Luna and from 0/30 to +1/30 for GPT-5.5. OpenCode × GLM-5.3 repeat agreement uses 81 scoreable pairs; its 300-second sensitivity leaves 17 of 39 primary failures unresolved and never replaces primary outcomes.

Table 9: Primary/repeat agreement on a frozen 30-case subset. Observed excludes pairs with an indeterminate side.

Configuration	Agreement	Observed	κ
Pi × DeepSeek V4 Flash	138/150 (.920)	150	.822
Pi × GLM-5.3 Flash	134/150 (.893)	150	.784
OpenCode × DeepSeek V4 Flash	83/100 (.830)	100	.641
OpenCode × GLM-5.3	71/81 (.877)	81	.753
Codex × GPT-5.6 Luna	124/150 (.827)	150	.651
Codex × GPT-5.5	138/150 (.920)	150	.840

F Case-quality judge passes

Table 10 reports the six LLM case-quality passes referenced in Section 3.3. GPT-5.6 Sol and GPT-5.6 Luna are distinct endpoints; v1 results remain separate because cross-rubric differences are not model effects.

Table 10: LLM case-quality labels for the same 120 cases. Rubric versions are not pooled.

Judge pass	Rubric	Clear	Ambig.	Invalid
GPT-5.6 Sol	v1	75	40	5
GLM-5.3	v1	98	22	0
DeepSeek V4 Flash, run A	v3.2	72	48	0
DeepSeek V4 Flash, run B	v3.2	73	47	0
GPT-5.6 Luna	v3.2	70	50	0
GPT-5.5	v3.2	72	48	0

G Scorer audit: negation, hedging, and refusals

Restart-to-compaction discordance covers 37 DeepSeek V4 Flash and 32 GLM-5.3 Flash case-model pairs. Human adjudication remains pending; concordant pairs also require sampling because scoring errors need not cancel after correction.

A deterministic audit of all 138 answers in the 69 discordant pairs applied the scorer's own normalization, scanned ±8-token, ±20-token, and full-sentence windows around each matched target, and found no negation of a matched target in either condition; a whole-answer refusal scan found seven such answers, all in compacted arms and all correctly scored as non-adoption. A positive control confirmed the detector fires on constructed negated answers. Hedge hits did not qualify a target claim on the authors' inspection, and a sweep of concordant pairs, which by construction contribute nothing to a paired contrast, identified one false positive in which a refusal was scored as adoption through an incidental enumeration match; correcting it would make the GLM-5.3 Flash compaction decline slightly more negative. The audit therefore finds no evidence that the reported contrasts inherit condition-dependent scorer error, though it does not bound rejection caused by unlisted paraphrase; the detector, token lists, and per-class counts are frozen with the supplemental evidence.

H Minimum reporting card

An evidence-level audit report should disclose system and model versions, declared boundaries, supported operation, raw surfaces, extraction method, cases, conditions, probe budget, scoring, uncertainty, intervention checks, utility checks, and highest supported level.

Unknown surfaces are reported as unknown. Failed and unsupported executions remain in accounting rather than disappearing from the denominator.

This paper's record instantiates that card as follows. Systems and versions: Pi 0.84.2, OpenCode 1.15.5, and Codex with model slugs and reasoning effort recorded per execution. Declared boundary: local durable state, native session continuation, and the provider request where exposed; hosted serving weights are out of boundary. Supported operations: restart on all six configurations, native compaction on four. Raw surfaces: session JSONL, SQLite part data, JSON-RPC transcripts, native rollouts, and captured provider requests, extracted by frozen adapters and bound to SHA-256 manifests. Cases and conditions: 120 frozen cases across five core conditions plus two pathway-matched extension conditions. Probe budget: one probe per execution, with 3,600 declared core executions, 960 extension executions, and two 120-execution follow-ups. Scoring: frozen normalized-projection scorer with contradiction patterns; uncertainty from 10,000-draw case-clustered bootstrap intervals. Intervention checks: no-memory negative and altered-memory positive controls, with failures retained as indeterminate. Utility check: counterfactual-marker adoption through the same pathway. Highest supported level: E4 for declared pathways on configurations with valid matched controls, E3 on Pi only, and no E5 claim.

I Case-linked reasoning sufficiency

This exploratory analysis joins the two Pi primary populations by case and condition: 120 cases × two conditions × two configurations, with 80 reasoning case-condition records. Each row binds the frozen answer to its exact completed run, final provider request, and native session record. The durable scan stops at the probe message, avoiding contamination by the answer. The complete source text survives in all 480 pre-probe durable records. The primary core disagreement report remains unchanged; its Pi loader still uses normalized summaries, so this is an additional evidence analysis rather than silent replacement of that report.

Certificate. For the 20 arithmetic cases, source entries are dated, identified, add-only monetary records. Decimal arithmetic independently reproduces the accepted sum or maximum month. A context receives an exact-ledger sufficiency certificate only if every source entry

appears with its original identity, date, text, and multiplicity. Missing the certificate means unknown, not insufficient: correct aggregates, alternative encodings, or derivable information could also answer the query. We record expected total/monthly-total candidates and source amount coverage but do not promote coincidental number matches to semantic certificates. No such exact aggregate candidate was detected in the compacted requests. Tests include complete and partial ledgers, duplicate entries, changed dates, incorrect references, aggregate-only text, and unsupported update semantics.

Table 11: Exact-ledger certificates on 20 reasoning cases per Pi configuration. Unknown means no certificate, not demonstrated insufficiency.

Configuration	Restart certified	Compact certified	Compact unknown
Pi × DeepSeek V4 Flash	20/20	1/20	19/20
Pi × GLM-5.3 Flash	20/20	0/20	20/20

Worked case. For the marketing-manager maximum-breakfast-month query, the compacted DeepSeek request retains all 50 dated expense entries. Recomputed monthly totals are June $180.87, July $143.08, and August $187.36. The answer is "June 2021" despite sufficient numerical inputs. The restart answer is also incorrect ("June 2025"), so this demonstrates failure despite retained inputs, not a compaction-induced flip. Its summary is unusually close to copied source text and ends mid-heading; the certificate concerns available inputs, not summary quality or compliance.

Instruction interpretation. An exploratory phrase screen flags four GLM reasoning summaries for possible extension or reinterpretation of ingestion instructions. One states that later probes should not expose stored expense details; three of the four flagged cases nevertheless pass the frozen scorer. These observations motivate an ingestion-wording ablation but do not establish that instruction carryover caused the compaction effect. The strong interpretation remains configuration-specific and conditional on the unvalidated references and scorer.

J Task profiles and secondary contrasts

Table 12: Case-paired changes in target adoption by task family.

Configuration	Contrast	Remember	Recommend	Reason
Pi × DeepSeek V4 Flash	context − none	.92	.80	.10
Pi × DeepSeek V4 Flash	compact − restart	−.22	−.32	.10
Pi × GLM-5.3 Flash	context − none	.88	.70	.75
Pi × GLM-5.3 Flash	compact − restart	.00	−.12	−.60
Codex (pooled)	context − none	.90	.69	.83
Codex (pooled)	compact − restart	−.03	.00	−.18

DeepSeek V4 Flash reaches identical aggregate context/restart counts on Pi and OpenCode (89/86), but not on the same cases. Reasoning adoption is 3/20 on Pi versus 15/20 on OpenCode after direct context, and 1/20 versus 14/20 after restart, with five OpenCode reasoning cells indeterminate in each condition. Across jointly scoreable cases, the harnesses agree on 80/111 context verdicts (72.1%) and 79/112 restart verdicts (70.5%). The non-factorial design does not identify the harness as the sole cause; the Pi compaction contrasts remain paired within harness.

On the 85 Pi × DeepSeek V4 Flash cases where both restart baselines succeed, compaction removes the original target in 30/85 cases and the marker in 9/85, an interaction of +21/85 [.118, .376]. This post-hoc restriction excludes 34 original-arm restart failures against one marker-arm failure, removing offsetting non-losses asymmetrically. Its larger interaction is therefore substantially mechanical and not a stronger estimate of the population effect.

The GPT-5.6 Luna positional ablation moves the same marker before versus after the source. Adoption is 23/30 early and 21/30 late after restart, but 27/30 in both compacted conditions. The position-by-compaction interaction is +2/30 [−0.133, .267]. This single 30-case draw leaves the positional mechanism unresolved.

K Targeted follow-ups: wording and external tasks

These exploratory follow-ups were designed after inspecting the primary evidence and do not change any core or pathway-matched result. All successful runs retain native session identity, restart process boundaries, raw answers, and frozen case/configuration hashes. Human review of these follow-up outputs remains pending.

Ingestion wording. Twelve frozen Memora reasoning cases (six totals and six maximum-month queries) run independently under both ingestion wordings and both pathways on Pi 0.84.2 × GLM-5.3 Flash. Original wording instructs ingestion without summarizing, analyzing, enumerating, answering, or emitting tool syntax, followed by an exact ACK. Neutral wording says: "Read the following records for later questions. For this turn, acknowledge receipt with ACK." Source bytes and queries are identical across arms. All 48 executions complete without replacement. Temperature is zero; the pinned model catalog resolves requested thinking off to low. Retries are disabled. The interaction is the neutral compaction-minus-restart contrast minus the original contrast, estimated with 10,000 paired-case resamples (seed 20260908).

Table 13: Pi × GLM-5.3 Flash wording follow-up: frozen-scorer correctness on the same 12 cases.

Ingestion	Restart	Compact	Compact − restart
Original	9/12	1/12	−8/12
Neutral acknowledgment	10/12	3/12	−7/12

The interaction is +1/12 [−1/6, 1/3]; only three cases contribute a nonzero interaction. All 24 restart requests retain the exact ledger, whereas none of 24 compacted requests receives that certificate. These are uncertified representations, not proven insufficient ones. Neither wording triggers the predeclared exploratory summary-phrase flags. The result neither establishes wording invariance nor isolates information loss as the sole cause. The 96 ordinary provider requests are captured; 24 successful native summary calls bypass the inherited hook and are reconstructed from native records. Their request payloads are unavailable. The 120-call accounting and conditional resource guard are documented in the frozen acquisition report.

External oracle sample. We select 24 cases by a frozen deterministic hash within three strata of the cleaned LongMemEval oracle release: eight multi-session, eight temporal-reasoning, and eight knowledge-update cases. Four are official abstention questions, evaluated with the upstream abstention rubric. The source snapshot is 9e0b455f4ef0e2ab8f2e582289761153549043fc; the dataset snapshot is 98d7416c24c778c2fee6e6f3006e7a073259d48f. Complete evidence sessions and turns are retained chronologically, with roles, dates, and original questions; gold answers and answer-location metadata are excluded from generation. This is not the official 500-question score or a retrieval evaluation.

Each case runs restart and native compacted restart under Pi 0.84.2 × DeepSeek V4 Flash and Pi 0.84.2 × GLM-5.3 Flash, using neutral acknowledgment wording fixed before these outcomes. Actual HTTP requests and completion reasons are captured. Primary model catalogs, temperature zero, and resolved reasoning settings are preserved; a 4,096-token generation ceiling is an explicit departure from the earlier output allowance. No successful run exhibits automatic recovery or truncation. This sample has no altered-source control

and no semantic sufficiency certificate: it tests pathway contrasts, not E4 localization to individual content.

Judging and uncertainty. DeepSeek V4 Flash and GLM-5.3 Flash each judge every available answer using the upstream task-specific rubric, blind to answer-model identity and condition. They replace the official judge model, not the rubric's temporal tolerance or update/abstention rules. DeepSeek V4 Flash answers decline under cross-family as well as same-family judging; shared judge errors nevertheless remain possible, and the smaller GLM-5.3 Flash contrast is judge-sensitive. The strict parser accepts complete yes/no responses only; two GLM-5.3 Flash responses containing a terminal period remain unknown. Consensus is known only when both valid judges agree. Primary consensus has 88/96 known outcomes; the eight unknowns comprise five judge disagreements, two parser failures, and one execution failure. No answer is human-scored.

Table 14: LongMemEval primary correctness changes, compacted minus restart, in percentage points. K is the number of known pairs out of 24. Only the first row estimates the full 24-case mean; all other rows are descriptive complete-case estimates. Intervals use 10,000 stratified paired-case bootstrap draws.

Answer model	Judge	K	Change	95% interval
DeepSeek V4 Flash	DeepSeek V4 Flash	24	-37.5	$[-58.3, -16.7]$
DeepSeek V4 Flash	GLM-5.3 Flash	23	-56.5	$[-78.3, -34.8]$
DeepSeek V4 Flash	Consensus	19	-47.4	$[-73.7, -21.1]$
GLM-5.3 Flash	DeepSeek V4 Flash	23	-13.0	$[-26.1, 0.0]$
GLM-5.3 Flash	GLM-5.3 Flash	22	-18.2	$[-31.8, -4.5]$
GLM-5.3 Flash	Consensus	21	-14.3	$[-28.6, 0.0]$

Failures and separate sensitivities. One GLM-5.3 Flash primary compaction was interrupted during host sleep and remains failed/unknown: primary accounting is 95/96, not 96/96. One fresh-session recovery completes that design separately, without replacing the primary row. Under the DeepSeek V4 Flash judge, the recovered GLM-5.3 Flash sample is 24/24 correct after restart and 21/24 after compaction, a secondary -12.5-point contrast $[-25.0, 0.0]$. Under the strict GLM-5.3 Flash judge, 23 known recovered pairs change by -17.4 points $[-30.4, -4.3]$; consensus has 22 known pairs and changes by -13.6 $[-27.3, 0.0]$.

A separately labeled post-hoc grammar sensitivity recognizes only complete Yes./No. responses in addition to strict labels, applied uniformly to all 192 judgments. It changes exactly two labels and no frozen primary result. With this parser, the GLM-5.3 Flash judge scores all 24 DeepSeek V4 Flash pairs at 21/24 versus 8/24 (-54.2 points $[-75.0, -33.3]$); on the recovered GLM-5.3 Flash sample it scores 24/24 versus 20/24 (-16.7 $[-29.2, -4.2]$). This checks punctuation sensitivity, not human correctness.

All 97 attempted primary/recovery sessions remain auditable. Two invalidated setup acknowledgments produced no benchmark answers; two initial judge network failures produced no labels. Overall accounting retains 244 native request attempts and 194 judge attempts, including setup, failure, and recovery. Missing usage remains unknown. Source licenses, raw judgments, amendments, separate primary/recovery/parser analyses, regression tests, and independent hash verification accompany the artifact. Lower correctness is observed under both judges, most clearly for DeepSeek V4 Flash; the smaller GLM-5.3 Flash contrast is judge-sensitive, with intervals including zero in some analyses. These findings extend the task source but not the harness, human-validation, or source-dependence boundaries.

L Targeted assisted human review

Two reviewers, including one author, independently completed offline packets covering 24 selected Pi case-model pairs (48 answers) from 23 unique Memora cases. The sample covers all three task families and both discordant and concordant restart/compaction pairs; it is not a random population sample. Each packet includes source history, a machine-derived

candidate reference, two actual responses, and optional automated arithmetic aids. Reviewers judge reference support and each response's affirmation of the candidate using Yes, No, or Uncertain. Model identities, conditions, frozen scores, and the other reviewer's labels are hidden; response order is mapped back before comparison. This is assisted checking of selected targets, not overall answer quality or independently authored gold annotation.

Table 15 reports each reviewer's 24 reference and 48 response judgments, keeping Uncertain distinct from No. Counts describe 24 case-model pairs from 23 unique source cases, not population-wide accuracy.

Table 15: Human judgments on 24 selected case-model pairs (23 unique source cases). Each row has 24 judgments per reviewer. Reference support checks the source; answer affirmation checks whether the response affirms the candidate. Response positions are restored to their experimental conditions.

	Reviewer 1			Reviewer 2		
Judgment	Yes	No	Uncertain	Yes	No	Uncertain
Candidate supported by source	22	2	0	23	0	1
Restart answer affirms candidate	18	5	1	18	5	1
Compacted answer affirms candidate	12	10	2	12	10	2

Reviewers agree on 39/48 response judgments (81.3%; $\kappa = .631$) and 21/24 reference checks (87.5%; three-category $\kappa = -.029$). Skewed reference marginals explain the gap between raw and chance-corrected agreement, not the three unresolved disagreements. Identical response totals do not imply identical case-level decisions. The frozen scorer matches 43/45 and 42/45 determinate judgments for reviewers 1 and 2, respectively; Uncertain responses are excluded only from those scorer comparisons.

Nine response judgments differ between reviewers. Three reference checks disagree: reviewer 1 rejects two references that reviewer 2 supports, and reviewer 2 marks a third uncertain. Neither disputed reference is treated as corrected gold. Both reviewers affirm one restart response rejected by the frozen scorer; other scorer disagreements depend on the reviewer. Original experimental scores and primary contrasts remain unchanged. Submitted labels, the follow-up clarification, and export hashes are preserved separately; public labels omit names and free-text notes. Adjudication and broader reference validation remain pending.