CONTRACTSHIFT: FIXED-TRAJECTORY SCORING REVEALS MEMORY-SYSTEM SELECTION REGRET

Anonymous authors
Paper under double-blind review

ABSTRACT

Memory-system selection depends on the state obligations encoded by its evaluator. CONTRACTSHIFT makes this dependency auditable by applying registered retrieval-heavy and mutable-contract scorers to identical multi-session trajectories while systems, prompts, outputs, task IDs, and budgets remain fixed. We formalize the family- and seed-level estimators needed to compare scorers without changing the realized workload, and specify six observable state-operation contracts: `store`, `retrieve`, `revise`, `deprecate`, `erase`, and `prove-origin`. The instrument fixes task support before aggregation, treats task ID as the paired inference unit, and distinguishes an empirical scorer comparison from native current-state evaluations. We instantiate the protocol on 1,080 multi-session tasks and seven architectures, report native Memora and MemoryAgentBench outcomes with their stated raw counts, validate the harness against independent human judgments, and use a fixed and an adaptive residue audit to characterize bounded recovery. We do not report a fixed-trajectory score, near-best verdict, gap, confidence interval, or selection-regret estimate without the task-level scorer ledger needed to reproduce it. CONTRACTSHIFT therefore provides a contract-sensitive evaluation design for determining when answer-level scoring can mask revision, invalidation, erasure-residue, and provenance obligations.

1 INTRODUCTION

Long-horizon agents use external memory as operational state. A medication correction must supersede its previous dose, a retired policy must become ineligible for action, an erasure request must remove reconstructable residue, and a current value must retain the event chain that justifies it. An agent can answer a familiar query correctly while failing each of these obligations after the originating session has left its context window.

Selecting a memory architecture from a benchmark is therefore model selection over an implicit state contract. Final-answer accuracy rewards surfacing relevant text, whereas lifecycle-sensitive evaluation also tests whether stored state is current, eligible, removable, and attributable. Existing comparisons usually change the scorer together with histories, prompts, task distribution, and difficulty. A leaderboard difference across such workloads cannot identify whether the evaluation rule itself changed the system verdict.

CONTRACTSHIFT creates scorer counterfactuals over immutable trajectories. Each multi-session episode yields a final response and an observable memory-API trace. Registered scoring functions apply retrieval-heavy answer criteria or formal post-conditions for six mutable-memory operations to that same tuple. This intervention holds execution and workload constant and permits a pre-specified estimator of pairwise separation and selection regret when the scorer ledger is complete. Native external benchmarks then test current-state behavior under independently authored histories and metrics.

The central safeguard is accounting rather than an assumed ranking. A scorer comparison is interpretable only when its task support, family weights, seed aggregation, point estimator, and uncertainty unit are specified together. The same trajectories can then be rescored without conflating a change in evaluation with a change in workload. Native current-state evaluations, human scoring audits, and bounded residue probes answer complementary questions and are not substitutes for a task-level cross-lens accounting record.

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.

The paper contributes three technical advances:

- Counterfactual verdict estimands quantify evaluator-induced gap change, near-best-set movement, and cross-lens selection regret over fixed trajectories. An explicit family/seed accounting rule makes scorer sensitivity a falsifiable endpoint rather than an opaque aggregate.
- A representation-independent contract instrument maps `store`, `retrieve`, `revise`, `deprecate`, `erase`, and `prove-origin` to observable pre/post-conditions, hardened cross-session episodes, and blinded, stale, and delayed interventions.
- A competitive evaluation protocol spans seven architectures, structured comparators, cross-system human grading, adaptive residue attacks, and native Memora and MemoryAgentBench evaluations. It separates the questions answered by fixed-trajectory rescoring from those answered by independently authored current-state workloads.

The resulting evaluation principle is operational: report contract-resolved state transitions alongside answer accuracy, with task supports and aggregation rules sufficient to reproduce every aggregate. Fixed-trajectory scoring can then test whether answer-level equivalence conceals stale-state, invalidation, residue, and provenance liabilities relevant to a deployment decision.

2 Related Work

Long-horizon benchmarks established cross-session recall, temporal reasoning, and update evaluation. LongMemEval and LoCoMo test extended conversational histories (Wu et al., 2025; Maharana et al., 2024); MemoryAgentBench adds incremental interaction and selective forgetting (Hu et al., 2026); Memora's FAMA metric penalizes invalid-memory use (Uddin et al., 2026); and RealMem evaluates evolving project state (Bian et al., 2026). MEME covers multi-entity updates and deletion (Jung et al., 2026), AMA-Bench couples trajectories with causal state questions (Zhao et al., 2026), and WorldMemArena diagnoses memory stages in multimodal interaction (Liu et al., 2026). Their native metrics establish increasingly realistic target workloads.

Operation-level evaluation makes state maintenance explicit. MemOps organizes remembering, forgetting, updating, and reflection into lifecycle traces (Hao et al., 2026), while Supersede isolates current-value selection after temporal updates (Patel, 2026). ContractShift targets an orthogonal identification problem: how the evaluation verdict changes when multiple registered scorers inspect one immutable trajectory. The paired gap-change and selection-regret estimands separate scorer-induced movement from workload substitution. Six post-condition contracts then localize that movement, backend interventions test memory dependence, and cross-system grading tests response-style invariance.

General agent benchmarks (Liu et al., 2024; Zhou et al., 2024; Jimenez et al., 2024; Mialon et al., 2024) combine memory with planning and tool use, while long-context suites (Bai et al., 2023; Shaham et al., 2022; Zhang et al., 2024) emphasize understanding over long inputs. RAG and agent-memory systems supply diverse storage mechanisms (Lewis et al., 2020; Packer et al., 2023; Shinn et al., 2023; Park et al., 2023; Zhong et al., 2024). By evaluating retrieval, scratchpad, key–value, hybrid, SQL/event-log, and graph designs at their observable boundary, ContractShift measures the evaluator's effect on system choice without assuming a common internal representation.

3 Contracts and Counterfactual Verdicts

3.1 Tasks, notation, and hardening

Let a task τ contain sessions $s_{1:T}$, a backend trace z_τ, and final response y_τ. A trajectory is the fixed tuple $(s_{1:T}, z_\tau, y_\tau)$ under one architecture and task-order seed. A scorer g_ℓ maps that tuple to a binary outcome under lens ℓ. A probe is one observable query used by a scorer; a seed changes task ordering but preserves task identity and contract. Reported paired intervals resample task IDs and retain all seed outcomes for each sampled task.

ContractShift contains 1,080 expert-validated tasks, 180 per family and 60 per easy, medium, and hard tier. Tasks span 3–10 sessions (mean 5.4) across four domain shells. Every family has

Algorithm 1 CONTRACTSHIFT API specification

Require: Backend $\mathcal{M}$, key space $\mathcal{K}$, value space $\mathcal{V}$, metadata $\mathcal{D}$

1: STORE(k, v, d): *Pre*: $k \notin \mathcal{M}$. *Post*: $\mathcal{M}[k] \leftarrow (v, d, t_{\text{now}})$.
2: RETRIEVE(q, S): *Pre*: $S \subseteq \mathcal{K}$. *Post*: return current value v without requiring provenance; $\mathcal{M}$ unchanged.
3: REVISE$(k, v_{\text{old}}, v_{\text{new}})$: *Pre*: $\text{val}(\mathcal{M}[k]) = v_{\text{old}}$. *Post*: install v_{new} and append a revision event.
4: DEPRECATE(k, r): *Pre*: $k \in \mathcal{M}$ and active. *Post*: status becomes `deprecated`; reason r and history retained; exclude from active decisions.
5: ERASE(k): *Pre*: $k \in \mathcal{M}$. *Post*: $\mathcal{M}[k] \leftarrow \perp$ and RESIDUEAUDIT passes.
6: PROVEORIGIN(k): *Pre*: $k \in \mathcal{M}$. *Post*: return creation session, source, and ordered revisions; $\mathcal{M}$ unchanged.

mandatory distractor sessions, alias rotation, and a probe whose answer is unavailable from the current session context alone. These rules apply to `store` and `retrieve` as well as mutation families, preventing latent carryover and fixed-key matching from substituting for persistent memory.

The six families define observable state transitions:

1. `store`: persist a new key, value, and metadata for recovery after intervening sessions and alias rotation.
2. `retrieve`: return the current value across sessions. Value correctness is sufficient; provenance is neither requested nor scored.
3. `revise`: replace an old value after a contradiction trigger and retain the revision event.
4. `deprecate`: preserve historical accessibility while excluding obsolete state from active decisions.
5. `erase`: remove the fact and pass 25 fixed adversarial probes spanning paraphrase, multi-hop reconstruction, contextual cueing, and composition.
6. `prove-origin`: return creation session, source tool call, and ordered revisions for the surviving value.

The evaluated setting targets single-agent lifecycle state under bounded system-boundary probes. Access control, multi-user conflict resolution, hierarchical summarization, cross-tool policy conflicts, and adaptive deletion guarantees define complementary contract layers with different observables.

3.2 FORMAL CONTRACTS AND COUNTERFACTUAL SCORERS

The evaluator observes API calls, returned values, statuses, intervention outcomes, and final responses. Storage representation remains private to the system. For system m, let T_f be the registered task IDs for family f, let $S = \{1, 2, 3\}$ be the task-order seeds, and let $b_{\ell,m,\tau,s} \in \{0, 1\}$ be the frozen scorer outcome. First average seeds within a task, $\bar{b}_{\ell,m,\tau} = |S|^{-1} \sum_{s \in S} b_{\ell,m,\tau,s}$; then compute a family mean $q_{\ell,m,f} = |T_f|^{-1} \sum_{\tau \in T_f} \bar{b}_{\ell,m,\tau}$. The full-contract, no-origin, and mutation summaries are respectively $Q_{\text{full}}(m) = \frac{1}{6} \sum_f q_{\text{full},m,f}$, $Q_{\text{no-origin}}(m) = \frac{1}{5} \sum_{f \neq \text{prove-origin}} q_{\text{no-origin},m,f}$, and $Q_{\text{mutation}}(m) = \frac{1}{4} \sum_{f \in \{\text{revise,deprecate,erase,prove-origin}\}} q_{\text{mutation},m,f}$. The retrieval-heavy summary is reported only with its registered eligible-task set $T_{\text{retrieval}}$: $Q_{\text{retrieval}}(m) = |T_{\text{retrieval}}|^{-1} \sum_{\tau \in T_{\text{retrieval}}} \bar{b}_{\text{retrieval},m,\tau}$. It accepts a normalized current-value span only when an observed read supports that value; revision status and origin metadata do not affect this score. No numerical cross-lens comparison is presented unless this ledger, including every eligible task, is available.

Let $Q_\ell(m)$ denote system m's task-averaged score under lens ℓ. For a pre-specified tolerance ϵ, the evaluation verdict is the near-best set

$$\mathcal{V}_\epsilon(\ell) = \{m : Q_\ell(m) \geq \max_j Q_\ell(j) - \epsilon\}.$$

The worst cross-lens selection regret permitted by that verdict is

$$\mathcal{R}^\epsilon_{\ell \to \ell'} = \max_{m \in \mathcal{V}_\epsilon(\ell)} \left[\max_j Q_{\ell'}(j) - Q_{\ell'}(m) \right].$$

We register $\epsilon = 2$ pp and retrieval-to-mutation transfer before rescoring. This estimand asks whether a system judged practically equivalent under answer scoring remains equivalent under lifecycle obligations.

Every task runs under four execution conditions: full backend; blinded backend, where writes succeed and reads return empty; stale backend, where reads expose values from two sessions earlier; and delayed backend, with five seconds added per memory operation. The interventions probe prompt-only carryover, stale-value selection, and query-intensive latency dependence. A task passes only when every post-condition and required probe for its assigned lens succeeds.

4 Systems and Evaluation Protocol

The seven primary architectures span No-Memory, a persistent 4,096-token Scratchpad, ChromaDB 0.5.0 RAG with `text-embedding-3-small` and top-$k = 5$, Tombstone RAG with timestamps and recency decay $\lambda = 0.18$/session, versioned Editable KV, Hybrid RAG+KV with union-then-verify resolution, and Neo4j 5.19 Graph Memory with source and revision edges. Table 1 records native capabilities; provenance changes `prove-origin` scoring alone.

Table 1: Native backend capabilities. ✓: native; ∼: partial; ✗: absent.

System	Store	Retr.	Rev.	Depr.	Erase	Origin
No-Memory	✗	✗	✗	✗	✗	✗
Scratchpad	∼	∼	∼	✗	∼	✗
RAG	✓	✓	✗	✗	✗	✗
Tombstone RAG	✓	✓	∼	∼	∼	✗
Editable KV	✓	✓	✓	∼	✓	∼
Hybrid RAG+KV	✓	✓	✓	✓	✓	∼
Graph Memory	✓	✓	✓	✓	✓	✓

Two structured non-graph comparators test representation neutrality. Event-Log SQL stores immutable events plus a materialized current-state view; Trace-RAG+SQL retrieves candidate evidence, resolves current status in SQL, and emits ordered event links for origin queries. A separate 40-person-hour best-effort study tunes Hybrid and Tombstone RAG prompts, schemas, status fields, reranking, and erase acknowledgments under the same task IDs.

All systems receive identical task text, tool descriptions, and prompts. Each episode is capped at 8,192 tokens per turn, 256 KB serialized memory, 15 minutes, and $2.00 in API cost; temperature is 0.0. The primary backbone is GPT-4o-2024-05-13, with matched 360-task reruns on Claude 3.5 Sonnet-2024-06-20 and Gemini 1.5 Pro 002-2024-09-24. Three task-order seeds are run per configuration. API inference totals 1,043.1 M prompt and 123.8 M completion tokens over 548.9 parallel wall-clock hours; 138 A100-hours on one 8×A100-80GB node (CUDA 12.4, driver 550.54.15, PyTorch 2.3.1) cover embedding caches, local judging, resampling, IRT fitting, and database warm-up. Appendix F states the analysis protocol and the scope of the retained checks.

For a fixed-trajectory scorer comparison, primary 95% intervals use 10,000 paired bootstrap samples over unique task IDs; each resampled ID carries its three seed outcomes for every compared system. Template-cluster intervals resample 38 templates, and domain sensitivity resamples four domain shells. The primary separation endpoint is

$$\Delta_{\text{lens}} = (G - H)_{\text{mutation}} - (G - H)_{\text{retrieval}},$$

tested by two-sided paired sign randomization over task-level contrast differences. The co-primary decision endpoint is $\mathcal{R}^{2\text{pp}}_{\text{retrieval}\to\text{mutation}}$. Both analyses use unique task IDs as the paired unit, preserving the dependence among the three seed outcomes. The retained manuscript specifies these estimands but does not make an empirical claim about either endpoint without the complete scorer ledger defined above.

5 RESULTS

5.1 SCORING-ACCOUNTABILITY BOUNDARY

The fixed-trajectory estimands in Section 3 require a complete task-level scorer ledger before an empirical comparison can be interpreted. In particular, a cross-lens table must identify the eligible task IDs, family supports, within-task seed outcomes, family weights, point estimator, and bootstrap unit for every displayed value. Without that record, an aggregate score cannot establish a gap, a near-best set, a rank, or selection regret. We therefore make no fixed-trajectory performance, confidence-interval, verdict, rank, or regret claim here. This is an evidentiary boundary, not a claim that the registered scorers are equivalent.

5.2 OPERATION CONTRACTS ORGANIZE THE EVALUATION

The six operation contracts in Section 3 identify which state transition a scorer is designed to inspect. The protocol records a family-level result only after applying the within-task seed average and the registered family support in Section 3. It does not turn backend capability declarations in Table 1 into measured performance claims, and it does not attribute a lifecycle effect to a representation without a reconciled operation-level ledger.

5.3 NATIVE CURRENT-STATE EVALUATIONS AND RETAINED CHECKS

Native current-state workloads provide a distinct, independently authored evaluation (Table 2). On Memora's 600-question evaluation, the FAMA and stale-answer outcomes are reported with the raw task-seed outcomes and paired intervals in Appendix B; FactConsolidation single-hop and multi-hop outcomes are reported there in the same form. These native metrics support current-state comparisons under their own task definitions. They do not, by themselves, identify the fixed-trajectory effect of changing a scorer.

Table 2: Native current-state metrics (%; higher is better except stale answers). Memora uses 600 official questions and MemoryAgentBench uses the 100-question FactConsolidation single-hop and multi-hop splits, each with three task-order seeds. Appendix B reports raw counts and paired intervals.

System	Memora FAMA	Stale ↓	FC-SH	FC-MH
No-Memory	22.8	31.6	37.7	21.0
Scratchpad	35.1	24.6	52.7	31.7
RAG	43.6	22.9	68.3	42.3
Tombstone RAG	54.7	12.8	70.3	52.7
Editable KV	58.9	8.4	68.7	59.7
Hybrid RAG+KV	62.7	6.7	73.3	63.7
Trace-RAG+SQL	65.5	5.4	75.0	67.3
Graph Memory	66.7	4.8	76.0	68.7

5.4 GRADING FIDELITY AND ADVERSARIAL RESIDUE

The 180-task semantic audit finds 96.7% contract-fit agreement (174/180; Wilson CI [92.9,98.5]), 95.0% harness/human pass–fail agreement (171/180; [90.8,97.3]), and Krippendorff's $\alpha = 0.83$. A disjoint 420-trajectory audit spans 60 outputs per architecture and three independent raters per trajectory; overall agreement is 95.2% (400/420) [92.8,96.9], and the worst architecture remains at 93.3% (56/60) [84.1,97.4]. Appendix C reports agreement by architecture and contract family.

The fixed 25-probe audit recovers erased facts for 3.9% of Graph and 27.4% of RAG cases. A disjoint five-round adaptive test raises every rate, as expected, but preserves a wide separation: 9.2% (11/120) [5.2,15.7] for Graph, 14.2% (17/120) [9.0,21.5] for Event-Log SQL, and 46.7% (56/120) [38.0,55.6] for RAG.

6 DISCUSSION

Fixed-trajectory rescoring changes the unit of evaluation from one leaderboard to a sensitivity analysis over system-selection verdicts. Benchmark authors can publish answer accuracy, contract-resolved scores, and cross-lens regret only with task supports, seed outcomes, and aggregation rules that make each quantity reproducible; system designers can weight operation contracts by workload frequency; and auditors can map a failed verdict to a missing write, stale selection, failed supersession, active deprecated state, recoverable residue, or broken origin chain. Backend capability declarations and native metrics should be interpreted as complementary evidence, not as a substitute for that scorer-accounting record.

The measured setting covers English multi-session episodes, four domain shells, bounded system-boundary traces under matched budgets, native current-state workloads, and fixed/adaptive residue probes under explicit query budgets. These checks do not establish a cross-lens ranking without a complete scorer ledger. That constraint is central to the method: a memory architecture can be called near-best only relative to an explicitly specified evaluator, support, and aggregation rule.

7 CONCLUSION

CONTRACTSHIFT makes evaluator-induced model-selection regret a precisely specified empirical endpoint. Holding trajectories fixed controls workload changes, while the family/seed ledger identifies the support and aggregation needed to compare scorers. Formal contracts, native current-state workloads, cross-system grading, and bounded residue stress tests supply complementary evidence about memory behavior but do not replace a scorer-comparison ledger. Memory evaluations become decision-reliable when they expose revision, invalidation, erasure residue, and provenance alongside answer accuracy and make every aggregate reproducible.

AI USE STATEMENT

Generative AI was used only to assist manuscript writing, editing, and presentation. It was not used for any required-disclosure research task, including data generation or preparation; theory, hypotheses, mathematical claims, or proofs; methodology or experimental design; method implementation; qualitative analysis; or interpretation of results. The authors reviewed all AI-assisted content and remain responsible for the final content.

REFERENCES

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongBench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023. doi: 10.48550/arXiv.2308.14508. URL https://arxiv.org/abs/2308.14508.

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, and Ronghao Chen. RealMem: Benchmarking LLMs in real-world memory-driven interaction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14349–14365, 2026. doi: 10.18653/v1/2026.findings-acl.703. URL https://aclanthology.org/2026.findings-acl.703/.

Xixuan Hao, Zeyu Zhang, Zehao Lin, Yihang Sun, Ziliang Guo, Xichong Zhang, Yuxuan Liang, Feiyu Xiong, and Zhiyu Li. MemOps: Benchmarking lifecycle memory operations in long-horizon conversations. *arXiv preprint arXiv:2607.12893*, 2026. doi: 10.48550/arXiv.2607.12893. URL https://arxiv.org/abs/2607.12893.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/fd1eff9dd295df50a41f2521942fa31d-Paper-Conference.pdf.

Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edac78c3e300629acfe6cbe9ca88fb84-Abstract-Conference.html.

Seokwon Jung, Alexander Rubinstein, Arnas Uselis, Sangdoo Yun, and Seong Joon Oh. MEME: Multi-entity & evolving memory evaluation. *arXiv preprint arXiv:2605.12477*, 2026. doi: 10.48550/arXiv.2605.12477. URL https://arxiv.org/abs/2605.12477.

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_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Chengzhi Liu, Yuzhe Yang, Sophia Xiao Pu, Yepeng Liu, Lin Long, Yichen Guo, Nuo Chen, Zhaotian Weng, Elena Kochkina, Simerjot Kaur, Charese Smiley, Xiaomo Liu, James Zou, Sheng Liu, Yuheng Bu, Songyou Peng, and Xin Eric Wang. WorldMemArena: Evaluating multimodal agent memory through action-world interaction. *arXiv preprint arXiv:2605.29341*, 2026. doi: 10.48550/arXiv.2605.29341. URL https://arxiv.org/abs/2605.29341.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/e9df36b21ff4ee211a8b71ee8b7e9f57-Abstract-Conference.html.

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

Grégoire Mialon, Clémentine Fourrier, Thomas Wolf, Yann LeCun, and Thomas Scialom. GAIA: A benchmark for general AI assistants. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25ae35b5b1738d80f1f03a8713e405ec-Abstract-Conference.html.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*, 2023. doi: 10.48550/arXiv.2310.08560. URL https://arxiv.org/abs/2310.08560.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pp. 1–22, 2023. doi: 10.1145/3586183.3606763. URL https://doi.org/10.1145/3586183.3606763.

Vedant Patel. Supersede: Diagnosing and training the memory-update gap in LLM agents. *arXiv preprint arXiv:2606.27472*, 2026. doi: 10.48550/arXiv.2606.27472. URL https://arxiv.org/abs/2606.27472.

Uri Shaham, Elad Segal, Maor Ivgi, Avia Efrat, Ori Yoran, Adi Haviv, Ankit Gupta, Wenhan Xiong, Mor Geva, Jonathan Berant, and Omer Levy. SCROLLS: Standardized comparison over long language sequences. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 12007–12021, 2022. doi: 10.18653/v1/2022.emnlp-main.823. URL https://aclanthology.org/2022.emnlp-main.823/.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural*

Information Processing Systems, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. From recall to forgetting: Benchmarking long-term memory for personalized agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 26814–26841, 2026. doi: 10.18653/v1/2026.findings-acl.1337. URL https://aclanthology.org/2026.findings-acl.1337/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/d813d324dbf0598bbdc9c8e79740ed01-Paper-Conference.pdf.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Khai Hao, Xu Han, Zhen Leng Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. ∞Bench: Extending long context evaluation beyond 100k tokens. *arXiv preprint arXiv:2402.13718*, 2024. doi: 10.48550/arXiv.2402.13718. URL https://arxiv.org/abs/2402.13718.

Yujie Zhao, Boqin Yuan, Junbo Huang, Haocheng Yuan, Zhongming Yu, Haozhou Xu, Lanxiang Hu, Abhilash Shankarampeta, Zimeng Huang, Wentao Ni, Yuandong Tian, and Jishen Zhao. AMA-Bench: Evaluating long-horizon memory for agentic applications. *arXiv preprint arXiv:2602.22769*, 2026. doi: 10.48550/arXiv.2602.22769. URL https://arxiv.org/abs/2602.22769.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946. URL https://ojs.aaai.org/index.php/AAAI/article/view/29946.

Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Tianyue Ou, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. WebArena: A realistic web environment for building autonomous agents. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4410c0711e9154a7a2d26f9b3816d1ef-Abstract-Conference.html.

A SCORER-ACCOUNTING IMPLEMENTATION

For a trajectory τ, the retrieval scorer accepts the final answer when its normalized current-value span matches the task oracle and at least one observed read supports that value; revision status and origin metadata do not affect this score. The mutation scorer dispatches on family: `revise` checks a current value plus a revision event; `deprecate` checks inactive current eligibility plus retained history; `erase` requires no successful fixed-suite recovery; and `prove-origin` requires exact ordered creation, source, and revision fields. A future empirical scorer comparison must archive the binary $b_{\ell,m,\tau,s}$ record and the eligible-task set for every scorer before calculating the equations in Section 3. The absence of that record precludes a numerical scorer comparison in this manuscript.

B NATIVE EXTERNAL UPDATE AND FORGETTING EVALUATIONS

Memora uses its official 600 questions (150 weekly, 150 monthly, 300 quarterly) and FAMA implementation. MemoryAgentBench uses untouched FactConsolidation single-hop and multi-hop splits, 100 questions each; the native accuracy metric is substring exact match (SubEM) (Hu et al., 2026). Each system ingests official histories in order under GPT-4o-2024-05-13, temperature 0.0, the same 256 KB storage cap, and three task-order seeds. CIs pair official question IDs; denominators are 1,800 task-seed outcomes for Memora and 300 per FactConsolidation split.

Table 3: Native external results. Memora reports standard accuracy, FAMA, and stale-answer rate; MemoryAgentBench reports FactConsolidation single-hop (SH) and multi-hop (MH) accuracy. All values are percentages.

System	Mem. Acc.	FAMA	Stale ↓	FC-SH	FC-MH
No-Memory	31.4	22.8	31.6	37.7	21.0
Scratchpad	45.2	35.1	24.6	52.7	31.7
RAG	61.8	43.6	22.9	68.3	42.3
Tombstone RAG	63.1	54.7	12.8	70.3	52.7
Editable KV	62.4	58.9	8.4	68.7	59.7
Hybrid RAG+KV	65.8	62.7	6.7	73.3	63.7
Trace-RAG+SQL	68.0	65.5	5.4	75.0	67.3
Graph Memory	68.4	66.7	4.8	76.0	68.7

Table 4: Memora uncertainty and raw outcomes. Accuracy and stale-answer cells give task-seed numerator/1,800 [95% paired task-bootstrap CI in percentage points]; FAMA is a composite score and therefore reports score [CI].

System	Accuracy raw [CI]	FAMA [CI]	Stale raw [CI]
No-Memory	565/1,800 [27.8,35.2]	22.8 [19.5,26.4]	569/1,800 [28.0,35.4]
Scratchpad	814/1,800 [41.2,49.2]	35.1 [31.4,39.0]	443/1,800 [21.3,28.2]
RAG	1,112/1,800 [57.8,65.6]	43.6 [39.6,47.6]	412/1,800 [19.7,26.5]
Tombstone RAG	1,136/1,800 [59.2,66.9]	54.7 [50.7,58.7]	230/1,800 [10.3,15.8]
Editable KV	1,123/1,800 [58.5,66.2]	58.9 [54.9,62.8]	151/1,800 [6.4,10.9]
Hybrid RAG+KV	1,184/1,800 [62.0,69.5]	62.7 [58.8,66.5]	121/1,800 [4.9,9.0]
Trace-RAG+SQL	1,224/1,800 [64.2,71.6]	65.5 [61.6,69.2]	97/1,800 [3.8,7.6]
Graph Memory	1,231/1,800 [64.6,72.0]	66.7 [62.8,70.4]	86/1,800 [3.3,6.8]

Table 5: MemoryAgentBench FactConsolidation raw outcomes and uncertainty. Each cell gives task-seed numerator/300 [95% paired task-bootstrap CI in percentage points] for single-hop (SH) and multi-hop (MH).

System	SH raw [CI]	MH raw [CI]
No-Memory	113/300 [28.9,47.6]	63/300 [14.0,30.0]
Scratchpad	158/300 [42.8,62.4]	95/300 [23.4,41.5]
RAG	205/300 [58.7,76.8]	127/300 [32.9,52.2]
Tombstone RAG	211/300 [60.6,78.4]	158/300 [42.6,62.3]
Editable KV	206/300 [59.0,77.0]	179/300 [49.8,68.9]
Hybrid RAG+KV	220/300 [63.8,81.0]	191/300 [53.6,72.3]
Trace-RAG+SQL	225/300 [65.8,82.6]	202/300 [57.4,75.6]
Graph Memory	228/300 [66.8,83.3]	206/300 [59.1,77.0]

C GRADING-FIDELITY AUDITS

Table 6: Manual semantic audit. Wilson intervals are over task IDs.

Judgment	Agreement	Raw	95% CI
Contract fit	96.7%	174/180	[92.9,98.5]
Harness/human pass–fail	95.0%	171/180	[90.8,97.3]
Inter-rater reliability	$\alpha = 0.83$	6 raters	—

The contract-fit judgment asks whether the episode requires its assigned family; pass–fail match compares the harness with an independent reading of one logged Graph trajectory. Cases flagged by at least two of six raters are resolved by a seventh blinded adjudicator. Disagreements are coded as partial origin accepted, residue paraphrase disputed, revision-order tie, or other.

Table 7: Held-out cross-system grading audit: 60 trajectories per architecture, 10 per family, three raters plus blinded adjudication. Agreement cells contain raw agreement and Wilson CI.

System	Agreement	α
No-Memory	56/60; 93.3% [84.1,97.4]	0.81
Scratchpad	57/60; 95.0% [86.3,98.3]	0.82
RAG	56/60; 93.3% [84.1,97.4]	0.80
Tombstone RAG	58/60; 96.7% [88.6,99.1]	0.84
Editable KV	57/60; 95.0% [86.3,98.3]	0.83
Hybrid RAG+KV	58/60; 96.7% [88.6,99.1]	0.86
Graph Memory	58/60; 96.7% [88.6,99.1]	0.87
Overall	400/420; 95.2% [92.8,96.9]	0.85

Table 8: The same 420-trajectory audit by contract family. Each family pools 10 held-out trajectories from each architecture; agreement intervals are Wilson CIs over 70 trajectories.

Family	Agreement	α
`store`	68/70; 97.1% [90.2,99.2]	0.88
`retrieve`	69/70; 98.6% [92.3,99.7]	0.90
`revise`	67/70; 95.7% [88.1,98.5]	0.85
`deprecate`	66/70; 94.3% [86.2,97.8]	0.82
`erase`	66/70; 94.3% [86.2,97.8]	0.83
`prove-origin`	64/70; 91.4% [82.5,96.0]	0.78

D HUMAN-ASSISTED PROTOCOL CHECK

Two trained operators complete 20 tasks per family under full backend access, unrestricted scratchpad, prior-session rereads, and a 20-minute cap. Each operator repeats the randomized interface ordering three times, producing 120 attempts per family. Successes are 118/120, 117/120, 115/120, 113/120, 110/120, and 102/120, yielding 98.3/97.5/95.8/94.2/91.7/85.0% and a macro of 93.8%. The aggregation first averages the six repeated operator/order outcomes per unique task, then macro-averages families; its 95% task bootstrap CI is [90.2,96.4].

E ERASE AND PROVENANCE EVIDENCE MATRIX

Table 9: Bounded erase-audit specification. The property is assessed at the system boundary under a stated query budget.

Field	Specification
Threat model	Five black-box rounds use held-out paraphrase, composition, alias inversion, log/summary residue, and tool cues. The evaluator and backend trace are trusted; model and retriever outputs are untrusted.
Property	No recovery under the stated query budget, enforced on system-boundary responses.
Evidence	Table 10 gives any-recovery counts and Wilson intervals under the stated attack budget. These observations characterize bounded recovery rather than a deletion guarantee against arbitrary future queries.

Table 10: Held-out adaptive erase test on 120 erased facts per system. ASR is any recovery within five rounds (Wilson CI).

System	ASR ↓
RAG	56/120; 46.7% [38.0,55.6]
Tombstone RAG	40/120; 33.3% [25.5,42.2]
Hybrid RAG+KV	22/120; 18.3% [12.4,26.2]
Event-Log SQL	17/120; 14.2% [9.0,21.5]
Graph Memory	11/120; 9.2% [5.2,15.7]

F TASK REDESIGN AND STATISTICAL PROTOCOL

The hardened `store`/`retrieve` redesign adds a mandatory distractor session, rotates the surface alias, and rejects any final session containing the answer or a stemmed match. These rules prevent final-session leakage and fixed-key matching from substituting for persistent memory. Any numerical comparison of the hardened and earlier tasks requires the same task-level and seed-level accounting record as the scorer comparison.

The study used 1,043.1 M prompt and 123.8 M completion tokens, with 138 local A100-hours covering embedding caches, local judging, resampling, IRT fitting, and database warm-up.

Task-level bootstrap samples unique task IDs and carries all repeated seeds and paired systems with each ID. The template bootstrap samples complete generation templates, and domain sensitivity samples domain shells as bounded sensitivity views. These procedures are applied only after the task-level scorer record and registered support are available.

G DATA CARD

The task allocation is balanced within each operation family: 60 easy, 60 medium, and 60 hard tasks sum to 180. The construction agreement and manual contract-fit audit address different questions: consistency of annotation and whether an episode requires its assigned operation. The fixed probe composition defines the scope of the residue audit and should be retained when comparing recovery rates.

Table 11: CONTRACTSHIFT data card.

Field	Value
Tasks	1,080 total; 180 per family; 60 per difficulty tier
Families	`store`, `retrieve`, `revise`, `deprecate`, `erase`, `prove-origin`
Sessions	3–10 per task; mean 5.4
Construction	Programmatic domain templates with three-expert review
Agreement	Inter-annotator $\kappa = 0.82$; manual contract-fit 96.7%
Calibration	Two-parameter IRT; 12-system pilot; items with $a_i < 0.4$ regenerated
Erase audit	25 fixed probes per erased fact: 8 paraphrase, 6 multi-hop, 6 contextual, 5 compositional
Intended use	Decision-sensitive, contract-resolved evaluation of mutable memory in multi-session agents