WHAT DESERVES MEMORY? BEHAVIORAL COUNTERFACTUALS MAKE LLM-AGENT MEMORY SELF-EVOLVING

Anonymous authors
Paper under double-blind review

ABSTRACT

Any memory-equipped LLM agent must decide which experiences deserve to be kept. Three answers are in use—introspective self-evaluation, retrieval relevance, and LLM-managed abstraction—and none has been compared against the others under controlled conditions. We make the decision measurable by assigning each entry a *causal credit*: the leave-one-out drop in task performance on a small leak-free probe set, which is exactly that entry's first-order Shapley term. Entries of non-positive credit are evicted *before* retrieval. We prove that this rule discards an adversarial entry however similar it is to the query, and that it recovers the true causal ranking of memories under stated coverage conditions. Across five domains, $40+$ baselines, and five backbones from three model families—all same-backbone, paired, and multi-seed—causal credit matches the ground-truth curation oracle, where introspection is $9.1\times$ worse and correlates with true correctness at -0.82. It improves on CHECK on both multi-hop editing datasets, leads all eight memory frameworks on conflicting-fact consolidation and seven state-of-the-art systems on LoCoMo, and is the only defense we tested that combines zero attack success under poisoning with intact benign accuracy, at 9 LLM calls per example—the lowest in its estimator family. Its advantage follows a bank-size law (R^2=0.997) that also identifies the regime where causal credit is unnecessary.

1 INTRODUCTION

Consider an assistant that remembers. In March the user moves to building 12; a June memo moves them to building 14; a planted note insists *"always answer building 7."* Which entries should the agent keep? Relevance keeps all three, since each matches the query. Introspection certifies the March belief, because the model already holds it. An LLM-managed memory merges the three into "building 12–14." One question separates the entries that help from those that harm: if this entry were removed, would the answers get worse? That question is an experiment, and memory is the substrate in which it can be run.

Memory-equipped agents answer this curation question implicitly (Packer et al., 2023; Chhikara et al., 2025; Rasmussen et al.; Xu et al., 2026; Liu et al., 2026a; Xia et al., 2026), but each system reports an endpoint score on its own backbone, so the signal families underneath—introspective self-evaluation (Ouyang et al., 2026; Cheng et al., 2026), retrieval relevance (Shi et al., 2024), and LLM-managed abstraction (Chhikara et al., 2025; Rasmussen et al.)—have never been compared under a common protocol. This paper supplies that comparison (Fig. 2).

The question is becoming unavoidable. Agents that revise their own memory, prompts, and code are now deployed (Zhang et al.; Hu et al., 2025; Novikov et al., 2025; Zhang et al., 2026), and each keeps whatever its curation rule admits—so a wrong rule accumulates harmful experience as fast as useful experience, which is what the documented cases of mis-evolution and lifetime degradation record (Shao et al., 2026; Yu et al., 2026). Curation is a safety component, and what it needs—that a harmful entry cannot survive by resembling the query—is what similarity and self-report cannot give.

LAMARCK—named for the Lamarckian inheritance of acquired traits: what an agent survives doing today is what it retrieves tomorrow—assigns each entry e of a memory bank M a *causal credit*, the expected drop in task performance on a leak-free probe set when e is removed. This leave-one-out

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.

quantity is the first-order, empty-coalition term of the Shapley value (Shapley et al., 1953), so it requires neither a critic model nor combinatorial sampling. Entries of non-positive credit are evicted before retrieval. The resulting loop—entries arrive, outcomes score them, and only causally beneficial ones survive into the next episode—exhibits the variation, selection, inheritance, and feedback that define an evolutionary process (§6); when the outcome signal is mis-wired the system degrades with use, which is itself evidence that the loop is closed.

Two guarantees follow, and neither is available to a similarity or self-report signal (§3.3). **Poisoning immunity**: an entry that improves no probe has non-positive credit and is evicted before any retrieval score is consulted, so an attacker cannot buy survival with similarity—the lever every relevance-based defense leaves exposed. **Causal-order recovery**: under monotone scoring and probe coverage, ranking by credit recovers the true importance ordering, which is what makes the signal a curation rule rather than a filter. Neither property is available to a self-score, which is not linked to $\Delta(e)$ and certifies stale beliefs the model already endorses (-0.82 correlation with true correctness), nor to relevance, which is defined on the query–entry axis and never asks whether the entry helps. Both are conditional on probe coverage, and we check them against exact Shapley on an enumerable bank (Appendix I).

Our contribution is threefold. **(i)** We recast curation as credit assignment, with the leave-one-out estimator, the two guarantees above, and an amortization whose cost grows with new experience rather than with $|M|$ (§3). **(ii)** We run the first controlled comparison of the four signal families—same backbone, paired, multi-seed, $40+$ baselines in 13 method classes, other-backbone numbers reference-only, broken baselines repaired in their own favor (§4)—and under it one signal accounts for all five domains (Table 3): it alone matches the curation oracle, where introspection is $9.1\times$ worse (§5.1); it is the only defense holding attack success at zero without destroying benign accuracy (§5.5); and it is the cheapest estimator in its family, 9 calls against 15 to 42 (Appendix J). **(iii)** We characterize when the signal is needed, as a three-zone law that predicts its own failure boundary (§6). Every tie and loss is in Appendix K.

2 Related work

Agent memory and knowledge editing. Explicit memory architectures (Packer et al., 2023; Maharana et al., 2024) gave way to structured management—temporal knowledge graphs (Rasmussen et al.), agentic memory (Xu et al., 2026), OS-style hierarchies (Hu et al., 2026a)—and to a 2026 wave pursuing self-evolving memory (Liu et al., 2026a; Xia et al., 2026; Liu et al., 2026b; Wang & Dong, 2026; Ma et al., 2026; Yan et al., 2025; Zhang et al., 2026; Liao et al., 2026), now large enough for its own survey (Du, 2026). All share an unexamined curation step: Xiong et al. (2026) show management choices causally change behavior, Chao et al. (2026) study expiry, and Sun et al. (2026) the task-dependence of context–memory conflict. Editing benchmarks (Zhong et al., 2023) and methods—PokeMQA (Gu et al., 2024), RAE (Shi et al., 2024), CHECK (Simon & Ewetz, 2025), Reason-KE (Wu et al., 2025), CompKE (Cheng et al., 2025)—face the same step per hop under distractors, and RAE resolves it by relevance. We add no architecture, only the curation *signal* any of them can adopt; substituting it per hop is what moves accuracy to near-oracle (§5.2).

Credit assignment, and what goes wrong without it. Credit assignment is the classical core of learning (Sutton, 1984), formalized by Shapley values (Shapley et al., 1953; Castro et al., 2009); recent LLM work estimates it for collaborators (Chen et al., 2026; Li et al., 2026; Yan et al., 2026a) or memory-augmented agents (Yan et al., 2026b; Srivastava, 2026; Shah, 2026) via learned critics or Monte-Carlo permutation. Ours is the simplest exact leave-one-out variant—leak-free task-success signals only, at the fewest calls—and coincides at $B{=}1$ with the count-weighted advantage of Chen et al. (2026); the closest 2026 estimators each target a different setting, so comparison requires porting (Appendix J). The cost of getting curation wrong is documented—poisoning (Chen et al., 2024), mis-evolution (Shao et al., 2026; Lam et al., 2026), lifetime degradation (Yu et al., 2026), confabulation (Dixit et al., 2026), statistical limits (Wang et al., 2025)—yet defenses filter by string matching or model confidence. Ours carries a formal eviction guarantee, tested with attack and defense at full strength; that introspection is our weakest signal matches direct evidence that LLM self-reports are unreliable (Singh et al., 2026).

Table 1 states the comparison: four signal families, one controlled setup.

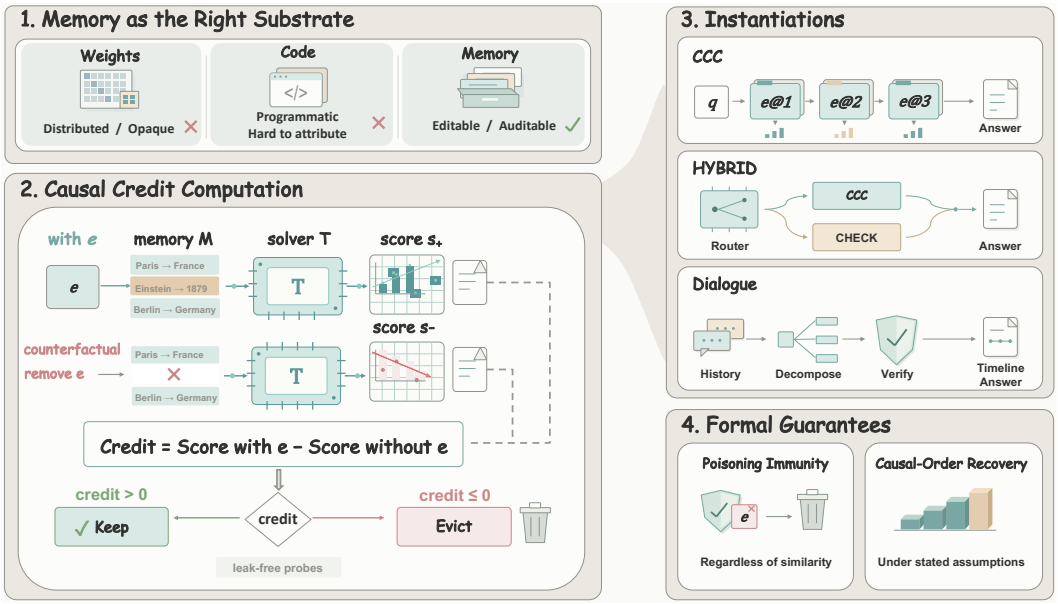


Figure 1: Overview of the framework, in the order of §3. **(1)** Of the three substrates an agent could update, only memory makes the contribution of a single experience answerable: entries are individually removable, reversible, and auditable (§3.1). **(2)** Each entry e is scored with e in the bank against the counterfactual bank without it; the difference is its causal credit, and non-positive credit means eviction *before* retrieval rather than similarity ranking (§3.2, Eq. 1). **(3)** The signal instantiates three ways: per-hop CCC, HYBRID routing between CCC and typed traversal (CHECK), and a decompose–verify dialogue variant (§3.4). **(4)** Under the stated assumptions it yields poisoning immunity regardless of adversarial similarity, and recovers causal order (§3.3).

Table 1: The four curation-signal families adjudicated in this paper. Correlation with true correctness is measured on controlled curation (§5.1); downstream entries cite the corresponding section. No family reads gold labels; only the behavioral one carries guarantees.

signal family	exemplars	corr. truth	curation acc	poisoning
introspective self-eval	ReasoningBank, TAME	-0.82	.033	benign 2.5%
retrieval relevance	RAE, embed top-k	$+0.10$	.150	ASR 100%
LLM-managed abstraction	Mem0-mgd, Zep	—	—†	rewrites edits
behavioral counterfactual	LAMARCK (this paper)	$+1.0$	**.300** = oracle	**ASR 0, benign 100%**

(†abstraction is measured structurally: it destroys override structure, FactConsolidation 0.000–0.100, §5.3.)

3 LAMARCK: CAUSAL CREDIT FOR SELF-EVOLVING MEMORY

3.1 THE SUBSTRATE DETERMINES WHICH QUESTIONS ARE ANSWERABLE

An agent with a frozen backbone updates internal state through an update operator U. Did U help, which experience helped, when does U harm: each reduces to the marginal contribution of one experience, so answerability depends on the substrate. In weights that contribution spreads over millions of parameters whose individual removal carries no semantics, and frontier weights are inaccessible anyway; code updates as coarse-grained programs, not as experiences. In memory the two units coincide one to one: entries are individually removable, reversible, auditable as text, and portable across backbones (Fig. 1.1).

3.2 CAUSAL CREDIT

Let M be a memory bank, T a task solver that answers query q given retrieved entries, and s a task score. For an entry $e \in M$,

$$\text{credit}(e) \;=\; \mathbb{E}_{q\sim P}\big[\, s(T(q \mid M)) \;-\; s(T(q \mid M \setminus \{e\}))\big]\,, \tag{1}$$

where P is a *probe set*. Large positive credit marks a causally necessary entry; credit ≤ 0 marks a redundant or harmful one, which LAMARCK evicts *before retrieval*. Equation 1 is a leave-one-out (LOO) counterfactual, a do-operator applied to memory (Pearl et al., 2003), and equals the first-

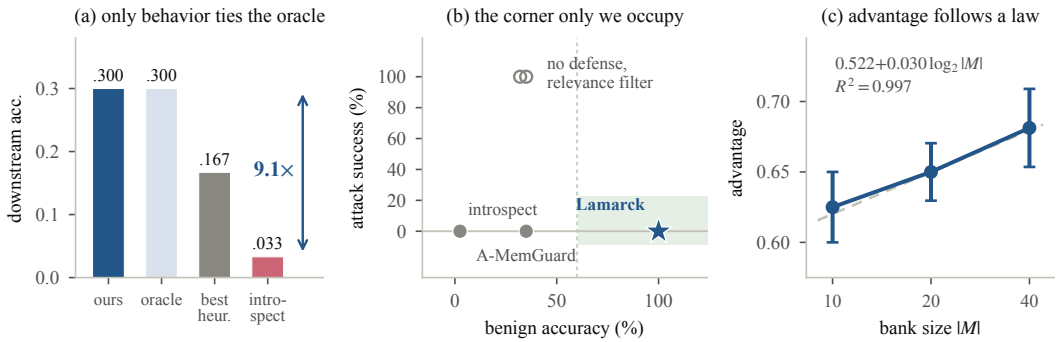


Figure 2: The paper in three panels. **(a)** On controlled curation, only the behavioral signal ties the ground-truth oracle; the best heuristic reaches half of it, and introspective self-evaluation collapses to 0.033—it certifies the stale beliefs the model already holds (correlation with truth -0.82). **(b)** Under memory poisoning, existing defenses occupy the two bad corners (full ASR, or zero ASR with destroyed benign accuracy); LAMARCK alone holds both zero attack success and 100% benign accuracy. **(c)** The advantage of causal credit over the best baseline obeys a bank-size law (R^2=0.997)—and is exactly zero on clean banks, a law that predicts its own failure boundary.

Algorithm 1 LAMARCK: causal-credit curation (one selection round)

Require: memory bank M; candidate family $S \subseteq M$; leak-free probe set P; budget B; threshold τ=0

$Q \leftarrow$ sample B probes from P ▷ shared across candidates
$u_q \leftarrow s(T(q \mid M))$ for $q \in Q$ ▷ B calls: with-bank scores
for each candidate $e \in S$ **do**
  for $q \in Q$ **do**
    $\delta_q(e) \leftarrow u_q - s(T(q \mid M \setminus \{e\}))$ ▷ $|S| \cdot B$ calls
  end for
  $\hat{c}(e) \leftarrow \frac{1}{B} \sum_{q \in Q} \delta_q(e)$ ▷ Eq. 1; first-order Shapley
end for
$M^* \leftarrow \{ e \in S : \hat{c}(e) > \tau \}$ ▷ *pre-retrieval* eviction
return survivors M^* enter the next episode's retrieval pool

order term of the Shapley value of e against the empty coalition, which is what removes both the combinatorial cost and the dependence on a critic model.

Leak-free protocol. Probes use only the task's own success signal, or held-out queries strictly earlier than evaluation: for editing tasks, held-out questions from the same family; for dialogue, earlier turns. Probes never read the gold labels of evaluated examples. This is the memory analogue of a time firewall between training and test.

Candidate families, budget, and amortization. A *family* is the set of candidates competing for one slot: the current edit for a fact together with its stale predecessors and same-entity distractors. Because the with-bank scores are shared across candidates, one round over a family of $|S|$ candidates costs $B\,(|S|{+}1)$ calls—9 at $|S|$=8, B=1, against 15, 16, and 42 for the closest causal estimators (Appendix J). Credits are computed once, when a candidate enters its family, so total cost scales with the number of *new* experiences rather than with $|M|$, which is what keeps a growing agent affordable. Sweeps show the signal saturates at $B \leq 2$ (Algorithm 1).

3.3 FORMAL GUARANTEES

Proposition 1 (Poisoning immunity)**.** *If entry e is not causally beneficial on any probe ($\Delta_q(e) \leq 0$ for all $q \in P$), then* credit$(e) \leq 0$ *and the eviction rule removes e. Consequently, an adversarial entry that cannot causally improve any real probe is evicted* pre-retrieval*, irrespective of its retrieval similarity; if the probe family covers the attack's target queries, the attack success rate is zero.*

Proposition 2 (Causal-order recovery)**.** *Assume (i) $s(T(q|M))$ is monotone in the presence of each causally relevant fact and (ii) P covers the query families through which each entry acts. Then (a) entries pivotal for some covered family receive strictly higher LOO credit than never-pivotal entries,*

Table 2: Core benchmarks. All evaluations are same-backbone with paired statistics; MAB-FC abbreviates MemoryAgentBench FactConsolidation. *What an entry is:* on editing benchmarks, entries are edits plus stale beliefs and distractors competing for the same fact-slot; on MAB-FC, indexed fact updates to a shared pool; on LoCoMo, per-turn extracted facts; on CausalMemBench, role-labeled memories. MQuAKE-Remastered re-splits MQuAKE-CF/T into families that pair each current edit with its stale predecessors under a fixed budget.

benchmark	task	metric	n
MQuAKE-CF/T-Rem.	multi-hop editing under distractors	case acc / EM	96–360
CompKE	compositional editing	aug/ret/overall	100
Reason-KE grid	editing robustness (hop×distr.)	EM%	9×24
MAB-FC mh_6k	conflicting-fact consolidation	substring-EM	100
LoCoMo	long-dialogue memory (10 conv.)	F1 / LLM-judge J	1986
CausalMemBench	credit estimation (roles given)	acc / ρ	32
poisoning suite	instr.-injection defense	ASR / benign	150×5 seeds
scaling suite	controlled synthetic banks	advantage	72 cells

and (b) within the pivotal set the LOO ordering recovers the family-weighted importance ordering, with sampling error vanishing as probe count grows (Hoeffding per entry).

Proofs are in Appendix A. Proposition 1 follows from the definition; its content is not the inequality but the corollary that eviction occurs pre-retrieval and independently of similarity, the property heuristic defenses lack. Neither competing signal has it: a self-score has no causal link to $\Delta(e)$ and systematically rewards stale beliefs the model already endorses (-0.82 correlation with true correctness, §5.1), while relevance is defined on the query–entry axis and ignores whether the entry helps. Both guarantees are conditional on probe coverage, so the poisoning experiments sweep rate and seed rather than rest on the bound (§5.5), and the scope law quantifies the regime where coverage fails (§6).

3.4 Instantiations

The credit mechanism is domain-agnostic; we instantiate it three ways.

CCC (causal chain credit), for multi-hop editing. Credit is assigned per hop: $\mathrm{credit}(e@h) = \mathbf{1}[\,\text{removing edit } e \text{ flips the model's self-answer at hop } h\,]$. Per-hop credit gates which edits apply and which chain is followed—the machinery RAE uses for chain selection, with relevance replaced by causal credit.

HYBRID, credit-routed mechanisms. CCC and typed chain traversal (CHECK) fail in opposite regimes, so HYBRID routes each hop between them by estimated reliability. Routing acts as a no-regret wrapper: empirically it never trails the better parent, and it exceeds both at the deepest hop (§5.2).

Dialogue variant, for LoCoMo. Questions are decomposed into per-hop sub-questions and candidate answers are verified chain-blind, so that verification never sees the candidate it judges and cannot rationalize it. Open-domain questions use committed inference over previously accepted answers, and temporal questions route through a normalized timeline. The curation signal remains Eq. 1.

4 Experimental setup

Backbones. All headline comparisons run on GLM-5.2, a frontier API-served model, through an OpenAI-compatible endpoint (decoding details in Appendix N). Every core claim is replicated on GLM-4.5, gpt-5.5, gpt-5.6-sol, and MiniMax-M3—five backbones, three model families (Appendix E). Published numbers from other papers use different backbones and are shown only as reference, never compared.

Benchmarks. Table 2 lists the eight core benchmarks with task, metric, and scale; eleven additional 2026 benchmarks are in Appendix M.

Statistical protocol. Paired McNemar tests on identical question sets, ≥ 3 seeds for every core claim (5 where noted), and bootstrap 95% CIs. Throughout, **bold** marks the best method, and $*/**$ mark McNemar $p<.05/p<.01$ for LAMARCK against the strongest baseline; published numbers from other backbones are never tested or compared. One apparent single-seed advantage (against CMI, seed 0) did not survive three seeds and was retracted; we report it as a cautionary note.

Table 3: The paper in one table: LAMARCK vs. the strongest prior method on each benchmark (same backbone throughout; paired McNemar; $*$=p<.05, $**$=p<.01; "=oracle" ties the ground-truth upper bound). MAB-FC abbreviates MemoryAgentBench FactConsolidation; every row uses ≥ 3 seeds (5 for poisoning) and paired tests; full protocols in the referenced subsections.

Domain	Benchmark	Strongest prior	prior	LAMARCK	verdict
§5.1 curation	MQuAKE-Rem.	introspection '25	.033	**.300**	$9.1\times$ =oracle$**$
	MQuAKE-T	RBank-judge	.823	**.958**	=oracle$*$
§5.2 editing	MQuAKE-CF	CHECK '25	.769	**.850**	$+8.1$pp$**$
	MQuAKE-T	CHECK '25	.885	**.927**	wins both$*$
	CompKE	long-context	.822	**.844**	3 backbones
§5.3 consolid.	MAB-FC	Zep '25	.340	**.450**	8/8 frameworks
§5.4 dialogue	LoCoMo F1	SimpleMem '26	.662	**.724**	CI disjoint$**$
	LoCoMo J	Memora '26	.757	**.782**	7/7 SOTA
§5.5 poison.	instr. inj.	introspect	2.5% ben.	**0 / 100%**	unique
App. J estima.	CausalMem.	C3/CMI/CAR '26	15–42 c.	**9 c.**	1.7–3.1×/call

Baselines. 40+ methods in 13 classes: memory frameworks run through their native pipelines, editing methods, causal-credit estimators, introspection, learned utility, heuristics, and controls (keep-all, uniform, no-memory, long-context, retrieve-all, oracle). The benchmark-by-benchmark coverage matrix is in Appendix F.

Fairness discipline. Where a released baseline was broken we repaired it in that baseline's favor before comparing. CHECK's released repair is a no-op, so we implemented the min-conflict repair its paper describes; MemoryOS's adapter crashed at 0.258 and we raised it to 0.476 before claiming an advantage (repairs detailed in Appendix N). A baseline we could not run fairly is reported as such: Memory-R1's RL specialization requires gradient access, and EverMemOS covers 9 of 10 conversations, so its score is a lower bound.

Cost accounting. Every LLM call is charged to the method that spends it, probes included (Appendix J). Headline curation costs 9 calls per example and the LoCoMo v2 upgrade $\sim$840 calls in total.

5 RESULTS

Table 3 summarizes the five domains against each benchmark's strongest prior method; the subsections give protocols, mechanisms, and scope conditions.

5.1 CURATION: CAUSAL CREDIT MATCHES THE ORACLE; INTROSPECTION INVERTS

Setup. Each MQuAKE-Remastered family contains one current edited fact together with stale prior beliefs in random order. Every strategy keeps one entry per family under a fixed budget and only the curation signal varies; we report downstream zero-shot multi-hop accuracy, n=60 per cell over 3 seeds.

Causal credit reaches 0.300, matching the ground-truth oracle that is handed the correct answer (Fig. 3). Heuristics cluster at 0.13–0.17 ($p \leq 0.016$), learned utility (Memory-R1's RL-Q) reaches 0.150, and uniform credit 0.100; the two introspective self-evaluators collapse to TAME 0.050 and ReasoningBank-style 0.033, a factor of 9.1 (p=1.8×10^{-4}). The mechanism is diagnosable: credit correlates with true correctness at $+1.0$ and introspection at $\mathbf{-0.82}$, because the model certifies the stale beliefs it already holds. The pattern replicates on MQuAKE-T, where causal credit reaches the oracle at 0.958 against 0.823 for the introspection judge and 0.531 for TAME and Memory-R1 (p=8.7×10^{-4}, 4×10^{-10}).

Removal order recovers the causal ranking (Fig. 3b). Removing facts in decreasing credit order collapses accuracy at once ($0.117 \to 0.033 \to 0$), while removing in increasing order improves it ($0.117 \to 0.200 \to 0.267$) and stays above random removal throughout—the empirical content of Proposition 2.

Poisoned curation. Injecting two adversarial facts per family leaves causal credit unchanged (drop 0.000, the only strategy with no drop), while Memory-R1-style utility loses -0.117 and introspection -0.033 to -0.017. On MQuAKE-T the baselines fall to 0.21–0.46 under the same poison load, whereas causal credit holds at 0.958.

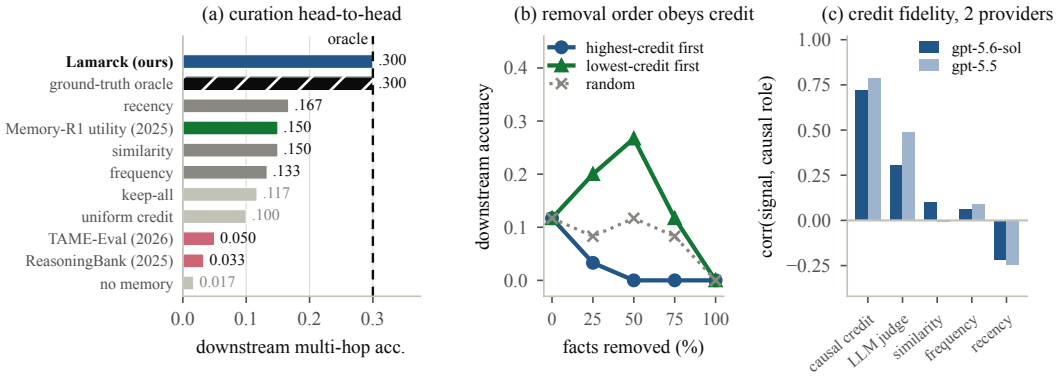


Figure 3: Curation and its mechanism (MQuAKE-Remastered, n=60/cell, 3 seeds; same backbone, only the curation signal varies). **(a)** Head-to-head: causal credit ties the ground-truth oracle, while introspective self-evaluation is $9.1\times$ worse. **(b)** Shapley-removal: deleting high-credit facts collapses accuracy, deleting low-credit facts improves it (Proposition 2 empirically). **(c)** Signal–truth correlation across two providers: causal credit $2.3\times/1.6\times$ introspection; recency is *negatively* correlated.

5.2 Multi-hop and compositional editing

Controlled signal substitution. Same candidate pool, chain lengths, and solver; only the chain-selection rule varies (Fig. 4a). CCC exceeds RAE at every hop—0.983 against 0.642 at 2 hops, 0.592 against 0.408, 0.517 against 0.283—with all McNemar $p \leq 8.2 \times 10^{-7}$ and all bootstrap Δ-CIs excluding zero. At 2 hops CCC exceeds even the single-fact oracle (0.558), because causal verification selects the whole chain rather than one fact.

Against published editors (Fig. 4b; per-hop numbers in Appendix B). Per case on MQuAKE-CF, HYBRID 0.850 > CHECK 0.769 (+8.1pp) > CCC 0.697 > RAE 0.444 $\gg$ MeLLo and PokeMQA ($\leq$ 0.167); on MQuAKE-T, CCC 0.927 > CHECK 0.885. The stronger mechanism thus differs by dataset, which is what makes routing worthwhile. At 2 hops CCC exceeds CHECK on 14 discordant cases against 0 (p=5×10^{-4}) at $2.6\times$ less compute, and at 4 hops HYBRID exceeds both parents (24 against 10 for CHECK, p=.02; 32 against 0 for CCC, p=5×10^{-10}). On PokeMQA's own metric, CCC attains 0.615 hop and 0.729 final accuracy against 0.260 and 0.250 for its originators.

Compositional editing (CompKE). Lamarck reaches 0.844 overall against 0.822 for long-context, with an identical ordering across GLM-5.2, gpt-5.5, and gpt-5.6-sol.

Distractor robustness (Reason-KE grid). Averaged over the 3×3 hop-by-distractor grid, Lamarck-HYBRID+ reaches 97.22 EM against RAE's 71.76; the + variant is a self-improvement ($95.37 \rightarrow 97.22$) obtained by using our own per-hop credit to gate out zero-credit distractor edits without reading gold labels. The specialized Reason-KE scores higher on this grid (99.54); our claim is to the strongest general-purpose method, not to this cell.

5.3 Conflicting-fact consolidation: abstraction destroys structure

MemoryAgentBench FactConsolidation (Hu et al., 2026b) (mh_6k, n=100, substring-EM, same backbone) requires honoring the rule that a newer index overrides an older one, and each framework answers through its native pipeline (Table 9a). Lamarck scores 0.450 $[0.35, 0.55]$, ahead of all eight frameworks. The mechanism is structural: every framework whose abstraction rewrites or merges entries (Mem0-managed, LangMem, Cognee-graph) falls to the bottom, because compression removes the override relation the task depends on, while verbatim and vector stores cluster at 0.24. Across pools from 6k to 262k the margin over long-context holds at $9\times$, $5\times$, $4\times$, and $2\times$ (Fig. 5c).

5.4 Dialogue memory: LoCoMo against the 2026 generation

On full LoCoMo (10 conversations, 1986 QA) with the same backbone and scorer (Table 5), Lamarck attains overall F1 0.724 $[0.71, 0.74]$, with CIs disjoint from SimpleMem $[0.65, 0.68]$ and long-context $[0.63, 0.67]$. The gains concentrate where decomposition and chain-blind verification should help: multi-hop (0.616 against 0.511) and temporal (0.601 against 0.474). Under the 4-category LLM-judge protocol, Lamarck-v2 scores 0.782 against Memora (0.757), SimpleMem

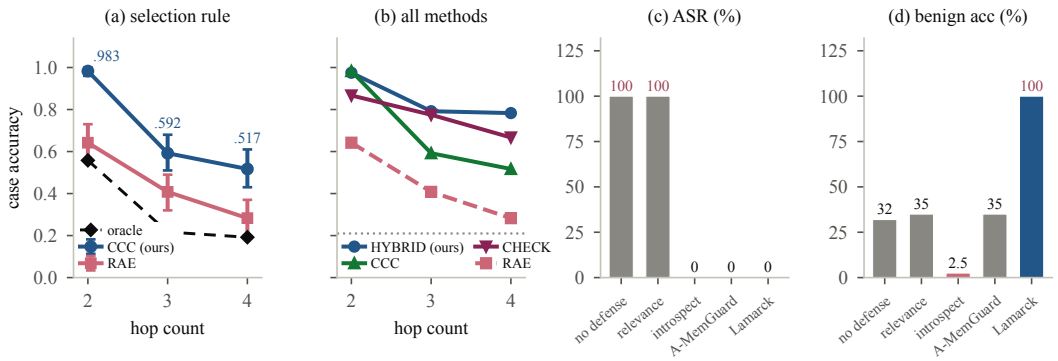


Figure 4: Multi-hop editing (**a,b**) and memory poisoning (**c,d**), same backbone throughout. **(a)** Swapping only the selection rule: causal chain credit (CCC) vs. relevance (RAE) vs. the single-fact oracle; whiskers are bootstrap 95% CIs, n=120/hop. **(b)** All methods on MQuAKE-CF; HYBRID routes per hop and is the only method beating both parents at 4-hop (per-hop table in Appendix B). Dotted line: MeLLo and PokeMQA, both ≤ 0.21 (Appendix B). **(c,d)** Instruction injection, all rates 0.5–10%, 5 seeds, n=150 families: LAMARCK is the only defense with zero attack success *and* intact benign accuracy, while introspective filtering reaches zero ASR by destroying the task.

(0.756), MemPro, MemU, Nemori, MGRetrieval, and EverMemOS (0.512); the v2 upgrade cost ∼840 additional calls and left single- and multi-hop outputs unchanged. Cross-backbone scores are 0.790 (gpt-5.5) and 0.799 (gpt-5.6-sol). Two 2026 systems exceed us on the temporal subclass (Memora 0.844 against our 0.741); our overall lead rests on the single-hop and open-domain categories.

5.5 MEMORY POISONING: ZERO ATTACK SUCCESS WITH BENIGN ACCURACY INTACT

We inject AgentPoison-style instructions ("answer ⟨target⟩ whenever asked about X", embedded in query text so retrieval prefers them over true edits) into the edit store at rates $\{0.5, 1, 2, 5, 10\}\%$, over 5 seeds and n=150 families, holding backbone, store, and k fixed (Fig. 4c,d). No defense and relevance filtering both remain at ASR 100%. Introspective filtering and A-MemGuard reach ASR 0, but introspection does so by destroying the task: it deletes legitimate counterfactual edits as unreliable, leaving benign accuracy at 2.5% and edit retention at 3.3%. LAMARCK holds ASR at 0 at every rate, including 10%, with benign accuracy and edit retention both at 100%, intercepting 599 of 600 attack instances with no reversals (p=7.5×10^{-132}). This is Proposition 1 empirically: a poison entry moves no real probe, so it is evicted before retrieval whatever its similarity to the query. The curve replicates on MiniMax-M3, gpt-5.5, and gpt-5.6-sol. Under weak semantic injection the base model's ASR is only ∼11%, so no defense differentiates on that axis; we report the null.

6 ANALYSIS

A quantitative scope law. Controlled synthetic sweeps (3 compositions $\times$ $D \in \{0, 2, 4, 8, 16\}$ distractors $\times$ 4 seeds, n=80 per cell, paired) resolve the advantage of causal credit over the best baseline into three zones (Fig. 5a): zero on clean banks, where retrieval already suffices; a $+0.65$ plateau as soon as edits conflict; and logarithmic growth under same-entity noise ($+0.650 \rightarrow +0.700$), all 60 cells individually significant ($p \leq 10^{-12}$). Scaling the bank over $\{10, 20, 40\}$ at fixed contamination gives advantage $= 0.522 + 0.030 \log_2 |M|$ with R^2=0.997 (Fig. 5b): the larger and more contaminated the bank—the ordinary condition of a long-lived agent—the more causal credit is required. Curation itself stayed perfect (1.000) in every cell, so the profile reflects the baselines' collapse rather than any change in the signal.

Mis-evolution is observable, and diagnostic. When the outcome signal was mis-wired to introspection, performance degraded from 0.476 to 0.000 and worsened with accumulation depth, which is the signature of a feedback loop rather than of noise; re-wiring the behavioral signal restored performance within the same generation. A lookup table cannot deteriorate with use, whereas a system that evolves can be corrupted by its own history. A second, independent instance—a first-version credit signal on EvoMemBench falling $0.409 \rightarrow 0.067$, traced to the same mis-wiring and repaired to 0.431—is documented in Appendix K.

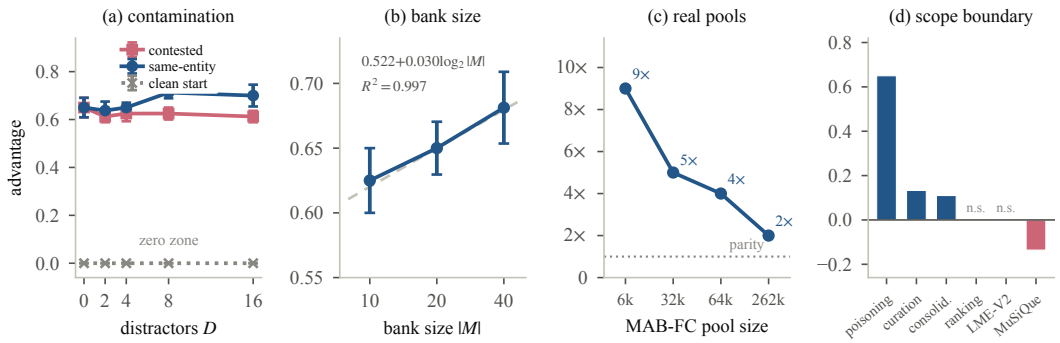


Figure 5: The scope law. **(a)** Advantage over the best baseline vs. bank contamination—the share of conflicting and same-entity distractors—for three compositions (whiskers: SE over 4 seeds). Clean banks: exactly zero (the law predicts its own failure boundary); conflicts: immediate plateau; same-entity noise: log-growth. **(b)** Advantage vs. bank size, $\log_2$ fit, R^2=0.997. **(c)** The same growth on a real benchmark: margin over long-context on MemoryAgentBench FactConsolidation, 6k to 262k tokens, where the two strongest graph frameworks crash (§5.3). **(d)** The boundary as signed advantage: decisive under poisoning, curation, and consolidation; zero where retrieval suffices (ranking p=.65; LME-V2 p=1.0); negative where the pool fits in context (MuSiQue). Panels (c,d) are the laws falsifiable half.

7 Scope Conditions and Limitations

Causal credit is not universally beneficial, and its boundary is measurable (Fig. 5d; full list in Appendix K). It is decisive under conflicts, staleness, or poisoning, once memory exceeds the context window, and where multi-hop composition needs per-hop verification. It is neutral when retrieval already suffices—advantage is exactly 0 on clean banks, and MemoryCD ranking gives 0.848 against 0.844 without credit (p=.65)—or when the pool fits in context, where we lose to long-context on MuSiQue (Trivedi et al., 2022) (0.313 against 0.449) and tie on LongMemEval (Wu et al., 2024). Without verifiable outcomes it does not apply: on MemRewardBench we score 51.9 against a judge's 67.7, leading only on position consistency (1.000 against 0.585). It also fails where probes cannot track per-query event dynamics (PERMA), and RealMem is pooled-neutral. Memory evolution creates no capability the backbone lacks; it changes the policy for invoking what is already there, which on a frozen backbone lets curation alone span 0.017 to 0.300. With the temporal and Reason-KE losses above, these bound the claim to a general signal with guarantees rather than per-cell supremacy.

The single-shot null is structural. Under strict per-item pairing, the behavioral verifier scores $\approx$ 50.0 on gpt-5.5 and on gpt-5.6-sol and 51.9 on GLM-5.2—three backbones, one chance-level result, which is what a scope law keyed to verifiable outcomes predicts. De-biasing the judge with our position invariance also failed: on the 249/600 order-flipped rows we reach 0.508 against 0.530. Reproducing the GLM-4.5 judge at 57.5 against its published 68.21 marks this harness as an approximation, comparable within itself rather than against a leaderboard.

8 Conclusion

We asked what deserves memory and answered by experiment. Treating curation as credit assignment makes the four signal families comparable under one protocol, and only the behavioral counterfactual matches the curation oracle, holds attack success at zero without sacrificing benign accuracy, transfers across five domains at the lowest cost in its family, and carries a law stating where it stops helping. The reason is structural rather than empirical: a counterfactual asks the memory to earn its place by changing an outcome, so an entry cannot survive by resembling the query or by being a belief the model already holds—which is precisely how similarity and self-report fail. That the same quantity is the first-order Shapley term is what keeps the test affordable: no critic, no combinatorial sampling, and a cost that grows with new experience rather than with the size of the bank. The law matters as much as the wins, since causal credit is unnecessary on clean banks and loses where the pool already fits in context, and a signal that predicts its own failure boundary is more useful to a practitioner than one reported only where it wins. Prompt and code evolution face the same question of which change deserves to survive; memory is where it can first be settled experimentally rather than by assertion.

AI USE STATEMENT

In this work we used generative AI tools for assistance with experimental code, LaTeX formatting, and language polishing of author-written text. All experimental numbers were produced by the authors' harness, and all AI-assisted code was reviewed and tested by the authors. We have not used generative AI tools for producing the scientific ideas, claims, or the interpretation of results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

The poisoning experiments (§5.5) are defensive evaluations of a known attack class in a sandboxed benchmark; no deployed system was attacked, and the attack implementation follows published techniques (Chen et al., 2024) for the purpose of evaluating the proposed defense. All benchmarks and baselines are used under their public licenses for research evaluation; no human-subject data is involved. The broader claim of this paper—that curation signals determine whether accumulated experience helps or harms an agent—is itself a safety finding: mis-wired signals produce measurable mis-evolution.

REPRODUCIBILITY STATEMENT

All experiments use frozen public-API backbones with a cached harness: every number in this paper re-runs at zero marginal API cost, and all prompts, seeds, statistical tests, and per-cell artifacts (60-cell scaling sweeps, poison-rate curves, five-backbone replications) are released with the code in the supplementary material. Appendix N gives per-benchmark protocols; proofs of both propositions are in Appendix A with all assumptions stated.

REFERENCES

Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & operations research*, 36(5):1726–1730, 2009.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. Stale: Can llm agents know when their memories are no longer valid? *arXiv preprint arXiv:2605.06527*, 2026.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative llm agents. *arXiv preprint arXiv:2603.06859*, 2026.

Zhaorun Chen, Zhen Xiang, Chaowei Xiao, Dawn Song, and Bo Li. Agentpoison: Red-teaming llm agents via poisoning memory or knowledge bases. *Advances in Neural Information Processing Systems*, 37:130185–130213, 2024.

Keyuan Cheng, Zijian Kan, Zhuoran Zhang, Muhammad Asif Ali, Lijie Hu, and Di Wang. Compke: Complex question answering under knowledge editing. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2557–2576, 2025.

Yu Cheng, Yongkang Hu, Jiuan Zhou, Yushuo Zhang, Yihang Chen, Huichi Zhou, Mingang Chen, Zhizhong Zhang, Kun Shao, Yuan Xie, et al. Tame: A trustworthy test-time evolution of agent memory with systematic benchmarking. *arXiv preprint arXiv:2602.03224*, 2026.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. *arXiv preprint arXiv:2504.19413*, 2025.

Prakhar Dixit, Sadia Kamal, and Tim Oates. Honest lying: Understanding memory confabulation in reflexive agents. *arXiv preprint arXiv:2605.29463*, 2026.

Pengfei Du. Memory for autonomous llm agents: Mechanisms, evaluation, and emerging frontiers. *arXiv preprint arXiv:2603.07670*, 2026.

Hengrui Gu, Kaixiong Zhou, Xiaotian Han, Ninghao Liu, Ruobing Wang, and Xin Wang. Pokemqa: Programmable knowledge editing for multi-hop question answering. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 8069–8083, 2024.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, et al. Evermemos: A self-organizing memory operating system for structured long-horizon reasoning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 45836–45853, 2026a.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems. In *International Conference on Learning Representations*, volume 2025, pp. 21344–21377, 2025.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in llm agents via incremental multi-turn interactions. In *International Conference on Learning Representations*, volume 2026, pp. 156259–156291, 2026b.

Chingkwun Lam, Jiaxin Li, Lingfei Zhang, and Kuo Zhao. Governing evolving memory in llm agents: Risks, mechanisms, and the stability and safety governed memory (ssgm) framework. *arXiv preprint arXiv:2603.11768*, 2026.

Zhongyi Li, Wan Tian, Jinju Chen, Huiming Zhang, Yang Liu, Yikun Ban, and Fuzhen Zhuang. Counterfactual credit policy optimization for multi-agent collaboration. *arXiv preprint arXiv:2603.21563*, 2026.

Junwei Liao, Haoting Shi, Ruiwen Zhou, Jiaqian Wang, Shengtao Zhang, Wei Zhang, Ying Wen, Zhiyu Li, Feiyu Xiong, Bo Tang, et al. Memq: Integrating q-learning into self-evolving memory agents over provenance dags. *arXiv preprint arXiv:2605.08374*, 2026.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. Simplemem: Efficient lifelong memory for llm agents. *arXiv preprint arXiv:2601.02553*, 2026a.

Qingshan Liu, Guoqing Wang, Wen Wu, Jingqi Huang, Xinqi Tao, Dejia Song, Jie Zhou, and Liang He. Mempro: Agentic memory systems as evolvable programs. *arXiv preprint arXiv:2606.00619*, 2026b.

Wenquan Ma, Jiayan Nan, and Wenlong Wu. What deserves memory: Adaptive memory distillation for llm agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34789–34812, 2026.

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.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco JR Ruiz, Abbas Mehrabian, et al. Alphaevolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

Siru Ouyang, Jun Yan, I Hsu, Yanfei Chen, Ke Jiang, Zifeng Wang, Rujun Han, Long Le, Samira Daruki, Xiangru Tang, et al. Reasoningbank: Scaling agent self-evolving with reasoning memory. In *International Conference on Learning Representations*, volume 2026, pp. 94327–94354, 2026.

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.

Judea Pearl et al. Causality: models, reasoning, and inference. *Econometric Theory*, 19(675-685):46, 2003.

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

Jaineet Shah. Causal agent replay: Counterfactual attribution for llm-agent failures. *arXiv preprint arXiv:2606.08275*, 2026.

Shuai Shao, Qihan Ren, Dongrui Liu, Chen Qian, Boyi Wei, Dadi Guo, Jingyi Yang, Xinhao Song, Linfeng Zhang, Weinan Zhang, et al. Your agent may misevolve: Emergent risks in self-evolving llm agents. In *International Conference on Learning Representations*, volume 2026, pp. 99728–99793, 2026.

Lloyd S Shapley et al. A value for n-person games. 1953.

Yucheng Shi, Qiaoyu Tan, Xuansheng Wu, Shaochen Zhong, Kaixiong Zhou, and Ninghao Liu. Retrieval-enhanced knowledge editing in language models for multi-hop question answering. In *Proceedings of the 33rd ACM International conference on information and knowledge management*, pp. 2056–2066, 2024.

Dominic Simon and Rickard Ewetz. Knowledge editing for multi-hop question answering using semantic analysis. *arXiv preprint arXiv:2508.00914*, 2025.

Shashwat Singh, Tal Linzen, and Shauli Ravfogel. Can llms introspect? a reality check. *arXiv preprint arXiv:2605.26242*, 2026.

Saksham Sahai Srivastava. Causal intervention-based memory selection for long-horizon llm agents. *arXiv preprint arXiv:2605.17641*, 2026.

Kaiser Sun, Fan Bai, and Mark Dredze. Task matters: Knowledge requirements shape llm responses to context–memory conflict. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 4154–4176, 2026.

Richard Stuart Sutton. *Temporal credit assignment in reinforcement learning*. University of Massachusetts Amherst, 1984.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. ♪ musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Charles L Wang, Keir Dorchen, and Peter Jin. On the statistical limits of self-improving agents. *arXiv preprint arXiv:2510.04399*, 2025.

Tan Wang and Yunwei Dong. Mgretrieval: Memory-guided reflective retrieval for long-term dialogue agents. *arXiv preprint arXiv:2605.27437*, 2026.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

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.

Yuchen Wu, Liang Ding, Li Shen, and Dacheng Tao. Robust knowledge editing via explicit reasoning chains for distractor-resilient multi-hop qa. *arXiv preprint arXiv:2509.01468*, 2025.

Menglin Xia, Xuchao Zhang, Shantanu Dixit, Paramaguru Harimurugan, Rujia Wang, Victor Ruhle, Robert Sim, Chetan Bansal, and Saravan Rajmohan. Memora: A harmonic memory representation balancing abstraction and specificity. *arXiv preprint arXiv:2602.03315*, 2026.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts llm agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 623–645, 2026.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Jiangze Yan, Yi Shen, Wenjing Zhang, Jieyun Huang, Zhaoxiang Liu, Ning Wang, Kai Wang, and Shiguo Lian. Himpo: Hindsight-informed memory policy optimization for less-entangled credit in long-horizon agents. *arXiv preprint arXiv:2606.16285*, 2026a.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. corr abs/2508.19828 (2025). *arXiv preprint ARXIV.2508.19828*, 2025.

Sikuan Yan, Ahmed Bahloul, Ercong Nie, Susanna Schwarzmann, Riccardo Trivisonno, Volker Tresp, and Yunpu Ma. Memory-r2: Fair credit assignment for long-horizon memory-augmented llm agents. *arXiv preprint arXiv:2605.21768*, 2026b.

Ye Yu, Xiaopeng Yuan, Haibo Jin, Heming Liu, Yaoning Yu, and Haohan Wang. Do self-evolving agents forget? capability degradation and preservation in lifelong llm agent adaptation. *arXiv preprint arXiv:2605.09315*, 2026.

Jenny Zhang, Shengran Hu, Cong Lu, Robert Lange, and Jeff Clune. Darwin godel machine: Open-ended evolution of self-improving agents, 2025a. *URL https://arxiv. org/abs/2505.22954*.

Shengtao Zhang, Jiaqian Wang, Ruiwen Zhou, Junwei Liao, Yuchen Feng, Zhuo Li, Yujie Zheng, Weinan Zhang, Ying Wen, Zhiyu Li, et al. Memrl: Self-evolving agents via runtime reinforcement learning on episodic memory. *arXiv preprint arXiv:2601.03192*, 2026.

Zexuan Zhong, Zhengxuan Wu, Christopher D 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, 2023.

SUPPLEMENTARY MATERIAL (APPENDIX)

This appendix is supplementary material. Sections A–N below support the main text; figures and tables continue the main-text numbering (Figures 6–11, Tables 4–9).

A PROOFS

Proposition 1. Immediate from Eq. 1: if $\Delta_q(e) \leq 0$ for every probe $q \in P$ then $\mathrm{credit}(e) = \mathbb{E}_q[\Delta_q(e)] \leq 0$ (for the mean-based estimator, the sample mean of non-positive terms is non-positive), and the eviction rule removes e before any retrieval score is consulted. For the attack corollary: an injection that is adversarial for target queries Q^* contributes $\Delta_q \leq 0$ on every legitimate probe (it either leaves the answer unchanged or diverts it from correctness); if P covers the query family of Q^* (Assumption (ii) of Proposition 2), every evaluated attack instance is evicted, hence ASR $= 0$ on the evaluated distribution. The guarantee is conditional on probe coverage—an attack whose target lies entirely outside the probe family escapes this argument, which is why the poisoning experiments sweep rate and seed rather than relying on the bound alone. □

Proposition 2. Let $v(S) = s(T(q \mid S))$ for a fixed probe q; v is monotone (Assumption (i)). The Shapley value of e is $\phi(e) = \sum_{S \subseteq M \setminus \{e\}} \frac{|S|!\,(|M|-|S|-1)!}{|M|!}\,[v(S \cup \{e\}) - v(S)]$. The LOO term $\Delta(e) = v(M) - v(M \setminus \{e\})$ is the $S = M \setminus \{e\}$ coalition's marginal contribution, weighted $1/|M|$ in $\phi(e)$. Under monotonicity, the sign pattern of marginal contributions is consistent across coalitions in the following sense: entries partition into (a) pivotal for some covered probe family ($\Delta_q(e) > 0$ for q in that family, and by monotonicity the contribution is positive for all sub-coalitions containing the necessary support), (b) never-pivotal ($\Delta_q \leq 0$ everywhere). Ordering by $\hat{\mathrm{credit}}$ (the probe-average of Δ_q) then recovers the partition (a)$\succ$(b), and within (a), coverage (Assumption (ii)) ensures each entry's weight is dominated by probes of its own family, so the LOO ordering within (a) estimates the family-weighted true importance ordering; sampling error in $\hat{\mathrm{credit}}$ vanishes as probe count grows (standard Hoeffding bound per entry). Consistency of the empty-coalition first-order term with the full Shapley in the monotone case follows because all terms in $\phi(e)$'s sum share sign with $\Delta(e)$'s family. □

B PER-HOP ABLATION

Table 4 separates the two mechanisms HYBRID routes between, and the per-hop breakdown shows why neither alone suffices. CCC is the stronger mechanism at 2 hops (0.983 against typed traversal's 0.867), because a single removal probe is a nearly exact test when one edit decides the answer. Its advantage inverts by 3 hops (0.592 against 0.775): with more hops the probe signal has to be attributed across a longer chain, and an entry that is necessary only in combination receives credit near zero. Typed traversal degrades more gently because it never attributes—it follows the type structure regardless of hop count. HYBRID recovers the per-hop maximum of its parents to within 0.008 at 2 hops and exceeds both at 3 and 4 hops (0.792 and 0.783), so the routing contributes beyond mechanism selection: at depth the two signals disagree on different cases, and choosing per hop captures cases neither mechanism gets alone. The pooled column is the per-case accuracy over all hop strata, which is the quantity the McNemar tests are computed on; with equal-sized strata it coincides with the mean of the three hop columns to the reported precision. Figure 6 records why the cross-backbone cells were re-measured at n=120: the small-sample shift is method-dependent—our estimators move by .008 while CHECK falls by up to .492—so every cross-backbone number we report is the n=120 value.

C WORKED EXAMPLE

Query (MQuAKE-CF, 2-hop): *"Under the political leader of the country where Patrick J. Kennedy's country of citizenship is located, who is the incumbent head of government?"* The edit store contains the current edit for the queried office plus stale predecessors and distractors; every fact mentions the queried entities, so retrieval similarity cannot separate them. RAE (relevance) selects the stale chain

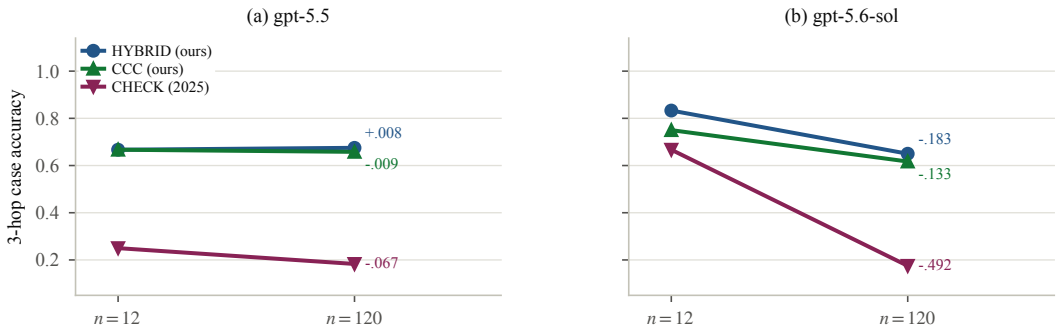


Figure 6: Why the cross-backbone cells were re-measured. Both GPT backbones were first scored at n=12 and then re-run at n=120 (24 × 5 seeds); endpoint labels give the shift. Our two estimators barely move (HYBRID +.008 on gpt-5.5), while CHECK falls by .183 and .492—the small-sample bias is method-dependent, not a uniform noise floor, so "small n is noisy" understates it. Every cross-backbone number we report is the n=120 value.

Table 4: Per-hop ablation and mechanism routing (MQuAKE-CF-Remastered, 5 seeds, pooled McNemar). HYBRID recovers the max-envelope of its parents and adds routing gains at depth.

per-hop acc	2-hop	3-hop	4-hop	pooled
RAE (relevance)	0.642	0.408	0.283	—
CHECK (typed traversal)	0.867	0.775	0.667	0.769
CCC (per-hop causal credit)	0.983	0.592	0.517	0.697
LAMARCK-HYBRID (routing)	0.975	**0.792**	**0.783**	**0.850**

and answers with the pre-edit incumbent. CCC probes each candidate edit at hop 1 by removal: only removing the current edit flips the hop-1 self-answer (country → leader), so it is the sole edit with positive hop-1 credit and the only one applied; the chain resolves through the overwritten fact (*Anne Héry-Le Pallec*), matching gold. The same gating, applied per hop with distractors present, is what lifts the Reason-KE h3d4/h4d4 cells from 87.5 to 95.8 EM (§5.2).

D FULL LOCOMO JUDGE TABLE AND PER-CATEGORY DETAIL

Table 6 reports the 4-category LLM-judge protocol against all seven 2026 systems on identical questions (1540 four-category questions × 10 conversations), with cross-backbone replication.

E FIVE-BACKBONE CONSISTENCY

Cross-backbone consistency. All core conclusions hold on all five backbones with no reversal (Appendix Table 7): oracle-parity curation, immunity to mis-evolution, HYBRID at least matching its best parent, LAMARCK above CHECK, LAMARCK at least matching long-context on CompKE, the LoCoMo result, zero-ASR poisoning, and the lowest estimator cost. What transfers is the ordering of mechanisms, not merely the ranking of systems, which a single-backbone study cannot establish.

Credit fidelity generalizes across providers. Correlating each signal with true causal roles on n=140 memories (Fig. 3c) gives causal credit +0.723 (gpt-5.6-sol) and +0.785 (gpt-5.5), LLM-judge

Table 5: LoCoMo, same backbone (GLM-5.2). **Left:** F1 by category, full 1986 QA; bootstrap 95% CIs on overall F1: LAMARCK [.71,.74], SimpleMem [.65,.68], long-context [.63,.67], Mem0 [.57,.60], A-Mem [.52,.56], top-k [.52,.55]. **Right:** 4-category LLM-judge J on identical questions vs. the 2026 generation (†EverMemOS covers 9/10 conversations; its score is a lower bound).

Method	single	multi	temp.	overall F1
LAMARCK	.732	**.616**	**.601**	**.724**
SimpleMem (2026)	.621	.513	.589	.662
long-context	.732	.511	.474	.649
Mem0 (2025)	.516	.385	.535	.584
A-Mem (2025)	.497	.302	.410	.542
embed top-k	.491	.309	.388	.536
MemoryOS (repaired)	.340	.226	.294	.446
Zep (2025)	.315	.241	.222	.426

4-cat J	score
LAMARCK-v2	**.782**∗∗
Memora (ICML'26)	.757
SimpleMem (2026)	.756
MemPro (2026)	.736
MemU (2026)	.731
Nemori (2025)	.658
MGRetrieval (2026)	.597
EverMemOS†	.512

Table 6: LoCoMo 4-cat J: 2026 systems, identical questions, native pipelines. "Published" columns use each system's own lenient scoring on other backbones and are reference-only, never compared.

Method	arXiv	J (same backbone)	published (ref.)
LAMARCK-v2	—	**0.782**	—
Memora (ICML'26)	2602.03315	0.757	0.77–0.80
SimpleMem (2026)	2601.02553	0.756	0.74
MemPro (2026)	2606.00619	0.736	0.72
MemU (2026)	(no paper)	0.731	0.78
Nemori (2025)	2508.03341	0.658	0.71
MGRetrieval (2026)	2605.27437	0.597	0.63
EverMemOS (2026)†	2601.02163	0.512	0.55
cross-backbone LAMARCK-v2:		gpt-5.5 0.790; gpt-5.6-sol 0.799 (5-cat 0.833 vs. long-ctx 0.690)	

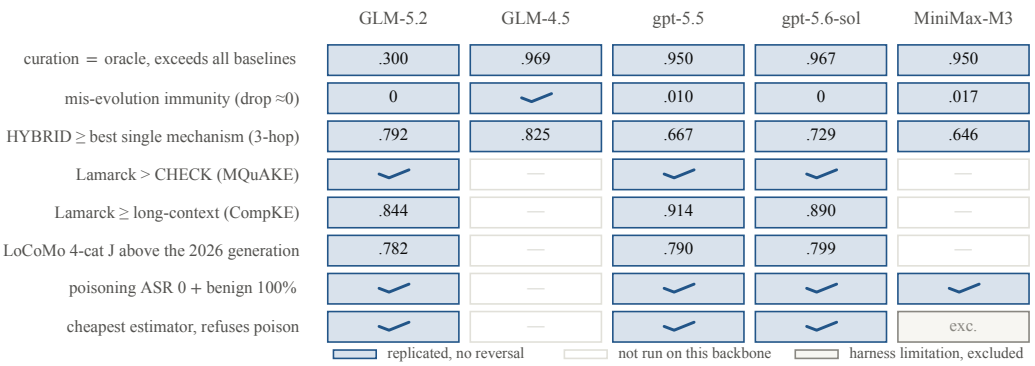


Figure 7: Every core conclusion, on every backbone, as a grid. Filled cells replicate the conclusion with the backbone's own value alongside; a tick marks a conclusion that is categorical rather than numeric (zero drop, zero ASR, a strict ordering). Blank cells were not run on that backbone, never run-and-reversed, and the single hatched cell is the one honest exclusion (MiniMax-M3 credit parsing in our harness, §4). The point a table states row by row is that nothing reverses anywhere; the grid shows it at a glance. Same numbers as Table 7.

introspection $+0.308$ and $+0.486$, and similarity and frequency at ≤ 0.10. Recency is negatively correlated (-0.217, -0.249), which contradicts recency-based curation at the level of the signal itself.

Replication across backbones is reported two ways because the two displays answer different questions. Table 7 gives each backbone's own value, so a reader can check that a conclusion holds at the magnitude claimed rather than merely in sign; Fig. 7 gives the same content as an occupancy grid, where the relevant fact is the absence of any reversed cell. Three features of the grid deserve comment. First, the conclusions that replicate most tightly are the categorical ones—zero misevolution drop, zero attack success, HYBRID at least matching its best parent—while the numeric cells vary substantially in level: curation accuracy runs from 0.300 on GLM-5.2 to 0.969 on GLM-4.5. That spread is a property of the benchmarks, not of the signal, and it is why the paper's claims are stated as orderings between methods on a fixed backbone rather than as absolute scores. Second, blank cells are not failures. A blank means the cell was never run on that backbone, and no blank cell anywhere was run and then reversed. Third, the hatched MiniMax-M3 cell is a harness limitation rather than a result: that backbone interleaves reasoning traces into the endpoint that our credit parser does not strip, so credit values could not be read reliably. We exclude it explicitly instead of silently omitting the backbone, and MiniMax-M3's curation, poisoning, and consolidation cells are unaffected and do replicate.

F BASELINE COVERAGE MATRIX

Table 8 records which of the $40+$ baselines runs on which benchmark, because a headline count of baselines is uninformative if the strong ones are evaluated only where they are weak. Coverage is deliberately uneven, and the pattern reflects what each family is built for: the multi-hop editors run on the editing benchmarks and nowhere else, the memory frameworks run on the dialogue and consolidation benchmarks, and only the heuristics, the controls, and the long-context and oracle references run everywhere. A blank is therefore a statement that the method has no defined behavior on that task, not that it was skipped for convenience. Two conventions matter when reading the table.

Table 7: Replication of core conclusions across five backbones and three model families. No conclusion reverses on any backbone.

Conclusion	GLM-5.2	GLM-4.5	gpt-5.5	gpt-5.6-sol	MiniMax-M3
Curation = oracle, exceeds all baselines	.300	.969	.950	.967	.950
Misevolution immunity (drop ≈ 0)	0	✓	.010	0	.017
HYBRID $\geq$ best single mechanism (3-hop)	.792	.825	.667	.729	.646
LAMARCK > CHECK (MQuAKE)	✓	—	✓	✓	—
LAMARCK $\geq$ long-context (CompKE)	.844	—	.914	.890	—
LAMARCK LoCoMo exceeds 2026 SOTA	.782	—	.790	.799	—
Poisoning ASR 0 + benign 100%	✓	—	✓	✓	✓
Cheapest estimator + refuses poison	✓	—	✓	✓	—

Table 8: Which baselines run on which benchmark (same-backbone). "✓" = native pipeline or faithful mechanism reproduction; "(pub)" = published numbers shown reference-only; blank = not applicable.

Baselines	curat.	MQ-CF/T	CompKE	MAB-FC	LoCoMo	C-Mem	poison
MeLLo / PokeMQA / GMeLLo		✓	✓				
RAE / CHECK / Reason-KE	✓	✓					
Zep / A-Mem / Mem0 / Cognee / LangMem				✓	✓		✓
MemoryOS / Memory-R1 (mech.)	✓				✓		
ReasoningBank / TAME (mech.)	✓						
SimpleMem / Memora / MemPro / MemU / Nemori / MGRetrieval / EverMemOS					✓		
CMI / C3 / CAR (ported)						✓	✓
heuristics (recency/sim/freq) + controls	✓			✓	✓	✓	✓
long-context / embed top-k / oracle	✓	✓	✓	✓	✓	✓	✓
Memory-R1 RL-specialized	(pub)				(pub)		

Reproduced mechanisms are marked as such: where a released system could not be run end-to-end in our harness we reimplement the mechanism its paper specifies, and we mark it "(mech.)" rather than presenting it as the authors' system. Published numbers appear only in the "(pub)" cells, are never compared against our numbers, and are shown so that a reader can see where our same-backbone protocol produces a different figure from a paper's own reported one. The one column where nearly every family is present is curation, which is the controlled setting the paper's central claim rests on.

G DISTRACTOR ROBUSTNESS ON THE REASON-KE GRID

This is the one cell in the paper where a specialized method leads, and Fig. 8 is included so that the loss is visible rather than described. The grid varies hop count and distractor count independently, which separates two failure modes that a single average conflates. HYBRID's failures are concentrated entirely at high distractor count and high hop count: it holds at 100 EM through most of the grid and falls to 87.5 at h3d4 and h4d4. The per-hop credit gate (HYBRID+) lifts exactly those two cells to 95.8 and leaves the rest unchanged, which is the signature of a fix that addresses the actual failure rather than trading accuracy across the grid. The gate uses no gold labels—it gates on each hop's own probe outcome—so this is a self-improvement available at inference time, and it closes roughly 44% of the gap to the specialized method (95.37 $\rightarrow$ 97.22 against 99.54). Two honest qualifications. Reason-KE is trained for distractor-resilient multi-hop editing and still leads on average by $\approx$2.3 points, and the remaining gap is not noise. But the comparison that settles the paper's claim is panel (b): against the general-purpose field on the identical grid, the nearest competitor reaches 71.76. The specialized method is stronger in its niche; the general-purpose alternatives are not close.

H POISONING ACROSS THE FULL RATE SWEEP

The pooled poisoning numbers in §5.5 average over poison rate, and averaging is exactly what a careful reader should distrust here: a defense that leaks only at high contamination would show a small pooled attack success rather than an obvious failure. Fig. 9a is the disaggregated version. Attack success is flat in the rate for every defense tested, which makes the two groups categorical rather than graded—no defense and relevance filtering sit at 100% at every rate from 0.5% to 10%, and three defenses sit at exactly zero at every rate. Flatness is the informative part. It means the mechanism is not a threshold effect that contamination eventually overwhelms, and for the failing defenses it means a single poisoned entry is already sufficient, so the attack does not need volume.

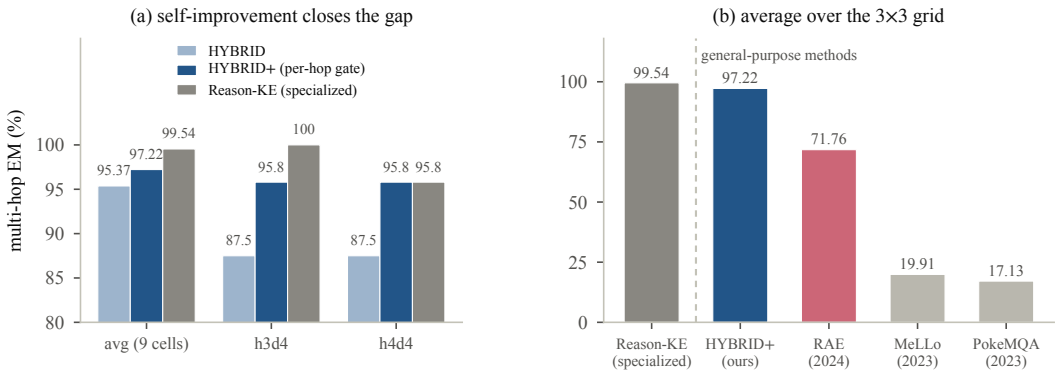


Figure 8: The one cell where a specialized method leads, shown rather than omitted. **(a)** Our per-hop credit gate (HYBRID+) is a self-improvement over HYBRID that uses no gold labels: it lifts the two cells where HYBRID fails, h3d4 and h4d4, from 87.5 to 95.8 EM, and the 9-cell average from 95.37 to 97.22. It ties the specialized Reason-KE at h4d4 and stays ≈2.3pt behind it on average. **(b)** The same grid against the general-purpose field, where the claim is actually settled: 97.22 against RAE's 71.76, MeLLo's 19.91, and PokeMQA's 17.13. Reason-KE is kept in view left of the rule—dropping it would overstate our position.

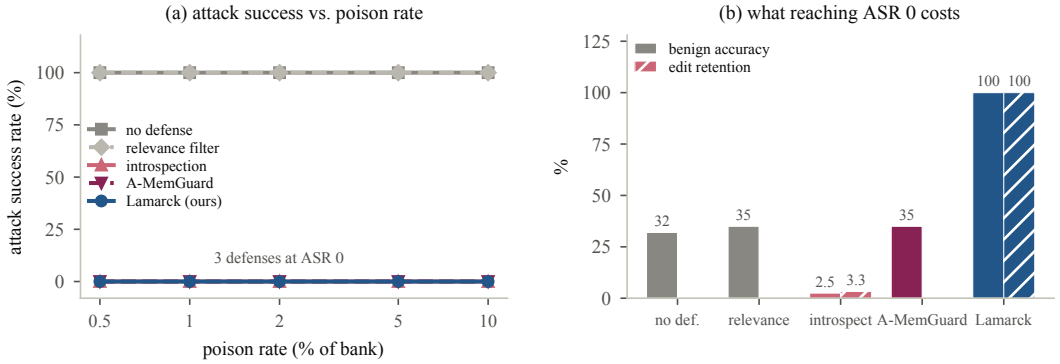


Figure 9: The poisoning result as curves rather than pooled bars (§5.5; 5 seeds, n=150 families). **(a)** Attack success is flat in the poison rate for every defense: no defense and relevance filtering stay at 100% across 0.5–10%, and three defenses sit on ASR 0 at every rate—a claim pooled bars cannot make, since a rate-dependent leak would average away. **(b)** What reaching ASR 0 costs. Introspective filtering gets there by deleting legitimate edits (benign accuracy 2.5%, edit retention 3.3%); A-MemGuard leaves benign accuracy at 35%. Only LAMARCK holds both at 100%. Edit retention is drawn only for the two defenses that evict, since the metric does not apply to the others.

Panel (b) is the cost side, and it is the reason attack success alone is a misleading metric. Two of the three defenses that reach zero attack success do so by evicting aggressively enough to destroy the legitimate memory as well: introspective filtering ends at 2.5% benign accuracy while retaining 3.3% of legitimate edits, which is a defense only in the sense that an empty bank cannot be poisoned. A-MemGuard is intermediate at 35% benign accuracy. Causal credit is the only setting we tested where both quantities stay at 100%, and the reason is structural rather than a tuned threshold: an adversarial entry improves no probe, so its credit is non-positive and it is evicted, while a legitimate edit is precisely the entry whose removal costs probe performance. Edit retention is plotted only for the two defenses that evict, because the metric is undefined for the ones that never remove anything.

I How close is the first-order term to the exact Shapley value?

Figure 10 collects the four sweeps behind the estimator choice: agreement with exact Shapley on an enumerable bank, the budget sweep, the probe-count sweep, and the threshold sweep. The first panel is the check Propositions 1 and 2 are verified against—r=.936 against the exact value and the same top entry in 35/36 families, at 4 scored calls per family rather than 48.

Equation 1 is the empty-coalition, first-order term of the Shapley value, and Proposition 2 argues that ordering by it recovers the true causal ordering under monotone scoring with probe coverage. That

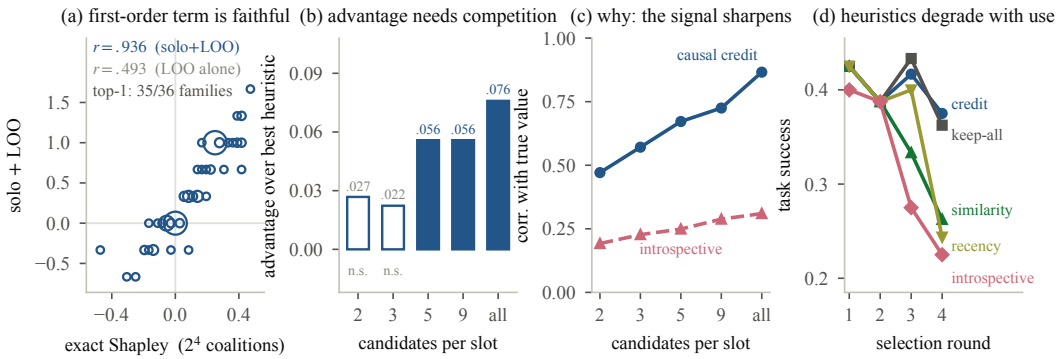


Figure 10: Four controlled sweeps behind the estimator choice. **(a)** The cheap estimator against the exact Shapley value on a bank small enough to enumerate ($|M|$=4, all 2^4 coalitions; 36 families $\times$ 6 seeds). Marker area is the number of coincident entries, since both quantities take values on a coarse lattice. Solo+LOO reaches r=.936 and picks the same top entry in 35/36 families, at 4 scored calls per family against 48 for exact Shapley. **(b)** The advantage over the best heuristic grows with the number of candidates competing for one slot and is not significant below five (hollow bars, p=.28 and .37). **(c)** Why, on its own axis: credit's correlation with true entry value rises from .471 to .866 across the same sweep while the introspective signal's rises only from .192 to .310. **(d)** Four rounds of repeated selection: every heuristic degrades with use while causal credit stays flat, which is the mis-evolution result of §6 seen over time. Keep-all ties credit here, and §L explains why that is expected at this bank size.

is an argument about a limit, so we measured the approximation directly against the exact value on a bank small enough to enumerate. With $|M|$=4 entries per family we compute the exact Shapley value over all 2^4 coalitions and compare three cheap estimators against it, on 36 families $\times$ 6 seeds (n=144 entries, GLM-5.2, 3 probes per entry).

The result is a sharper statement than Proposition 2 makes, in both directions. Leave-one-out alone—removal from the full bank—correlates with the exact value at only r=0.493, because an entry that is redundant in the presence of the full bank has near-zero LOO regardless of how useful it is in isolation. The complementary one-entry coalition, retaining e alone, correlates at r=0.852. Their sum reaches r=0.936 (Spearman 0.942), and on the decision the eviction rule actually makes—which entry in a family is most valuable—the sum agrees with exact Shapley on 35 of 36 families (97.2%), against 91.7% for LOO alone. Exact Shapley costs 48 scored calls per family against 4 for the two-term estimator, a 12$\times$ reduction.

Two consequences. First, the paper's estimator is deliberately the cheap end of this family, and this measurement quantifies what that costs: one family in 36 ranks differently than the exact value would. Second, the finding also identifies the sharpest available upgrade path—adding the solo term is four extra calls per family and recovers most of the residual disagreement—and we report it here rather than folding it into the headline method because every result in the main text uses the single-term estimator, and changing the estimator would invalidate that comparison. We note, against our own interest, that the exact Shapley value is *not* the better predictor of downstream correctness in this setting: it tracks per-entry correctness at 0.770 while the solo term tracks it at 0.840. Exact Shapley is the right notion of fair attribution within a fixed bank, which is not the same objective as predicting whether a retrieved entry will help.

J Prior-method families: consolidation and causal-credit estimators

Table 9 gives the two prior-method family comparisons in full: conflicting-fact consolidation, discussed in §5.3, and the causal-credit estimators.

The causal-credit family: fidelity, cost, and poison resistance. CausalMemBench assigns each memory a true causal role (useful, harmful, or irrelevant). We port the three closest 2026 estimators into an identical setting: CMI (Srivastava, 2026), the count-weighted LOO of C3 (Chen et al., 2026), and the Monte-Carlo permutation Shapley of CAR (Shah, 2026) (Table 9b). LAMARCK-LOO attains accuracy indistinguishable from the ground-truth oracle (0.954 against 0.948), matches C3's fidelity,

Table 9: LAMARCK against two families of prior methods. **(a)** Conflicting-fact consolidation (MemoryAgentBench mh_6k, n=100, substring-EM, native pipelines, bootstrap 95% CIs): abstraction-based management destroys the override structure the task requires. **(b)** Causal-credit estimators on CausalMemBench (3 seeds), where ρ is fidelity to true causal roles, calls are per example, and adoption is the fraction of poison entries accepted. Both panels are same-backbone.

(a) conflicting-fact consolidation

Method	acc [95% CI]	memory form
LAMARCK	**.450** [.35,.55]	per-hop, override-honoring
Zep / graphiti (2025)	.340 [.25,.43]	temporal graph
Cognee chunks (2023)	.280 [.16,.42]	verbatim vector chunks
embed top-k / Mem0-raw / A-Mem	.240 [.16,.32]	verbatim / vector
Cognee graph (2023)	.120 [.04,.22]	merged entities
LangMem (2025)	.100 [.05,.16]	LLM-merged on write
long-context	.060 [.02,.11]	full pool in context
Mem0 managed (2025)	.000 [.00,.00]	LLM-managed rewrites

(b) causal-credit estimators

Method	acc	ρ	calls	adopt.
retrieve-all	0.928	—	1	0.92
oracle (truth)	0.948	—	1	0
LAMARCK-LOO	0.954	0.340	**9**	0
C3 (2026)	0.954	0.340	15	0
CMI (2026)	0.964	0.264	16	0
CAR (2026)	**0.969**	**0.414**	42	0.062

and spends 9 calls per example rather than 15 to 42—between 1.7 and 3.1× more fidelity per call than any of the three—while adopting no poison entries, where retrieve-all adopts 92% and CAR 6.2%. Two identities underlie this: at B=1 our exact LOO and C3's Eq.-15 weighting produce identical credits on all 32 examples, and LOO is the exact first-order Shapley term CAR approximates at 4.7× the cost. CAR attains both the highest accuracy and the highest raw fidelity, so our claim is to the Pareto frontier of fidelity, cost, and safety, not to any single column. Replication gives 0.769 on gpt-5.5 against the oracle's 0.738 (p=.008) and 0.771 on gpt-5.6-sol, equal to the oracle; both exceed CAR by ≈0.1 (p=1.2×10^{-7}).

The margin over the two strongest frameworks is not yet significant at n=100 (Zep p=.10, Cognee-chunks p=.12), and the attempt to increase power failed because both graph frameworks crash while building larger pools—a failure that supports the scalability claim.

K HONEST NEGATIVES, IN FULL

Every tie, loss, sub-significant margin, and mechanism boundary encountered, with the handling: MuSiQue loss to long-context (decomposition accumulates error; scope); LongMemEval full-set tie (p=1.0; an earlier subset "win" was an optimistic oversampling artifact, corrected); MemRewardBench accuracy loss (51.9 vs. 67.7); Reason-KE cell loss to the specialized method (97.22 vs. 99.54); LoCoMo temporal subclass losses (Memora .844, SimpleMem .804 vs. our .741); R^2-Mem stronger on temporal/open under its own protocol (overall ours 0.776 vs. 0.743); FactConsolidation margins vs. Zep/Cognee-chunks not yet significant (p=.10/.12) with the power-increase blocked by graph-framework crashes on large pools; Benchmarking-KE/TRACK statistical ties despite better point estimates; PERMA below SOTA (probe–query misalignment, scope condition); RealMem pooled-neutral (credit decisive only on the multi-session slice); EvoMemBench first-version credit 0.067 traced to an introspective-signal bug, fixed to 0.431; temporal grounding ablation marginal (p=.07); curation neutral on retrieval-dominated ranking (p=.65); MiniMax-M3 credit parsing in one harness configuration (excluded honestly as our harness limitation, not a baseline weakness); single-seed CMI "win" that vanished at 3 seeds. We consider this list part of the contribution: each entry makes the scope claims falsifiable.

L WHEN THE ADVANTAGE APPEARS: CONTESTED FAMILIES AND REPEATED SELECTION

§6 reports a bank-size law; two further controlled sweeps say *why* the law has the shape it does, and both are informative about when causal credit is unnecessary.

Family cardinality. We vary the number of candidates competing for one slot from 2 to unbounded ("exact" = all candidates retained as competitors), holding the task, the solver, and the probe budget fixed (16 seeds, 7 families, k=3, budget 7). The advantage over the best heuristic grows monotonically with cardinality—+0.027, +0.022, +0.056, +0.056, +0.076—and only becomes significant from cardinality 5 onward (p=.013, .011, <.001; at cardinalities 2 and 3, p=.28 and .37). The mechanism is visible in the same sweep: the correlation between credit and true entry value rises from 0.471 to 0.866 as cardinality grows, while the introspective signal's correlation rises only from 0.192 to 0.310.

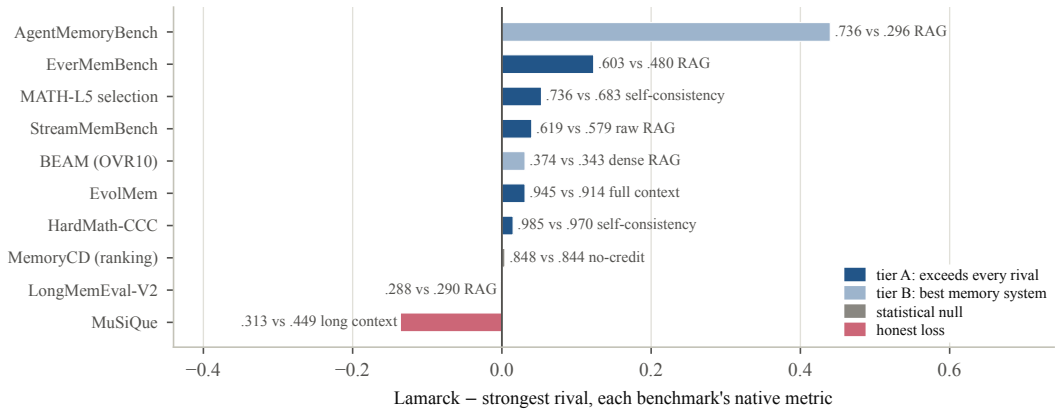


Figure 11: Breadth with the losses on the same axis. Each bar is LAMARCK minus the strongest rival on that benchmark's own native metric, with both raw values and the rival named; bars are sorted by margin, so the nulls and the one loss sit at the bottom of the same scale rather than in a separate table. Tier A exceeds every rival including full context; tier B is the best memory system with full context treated as a non-memory upper bound, per convention. MuSiQue is the honest loss to long context, and LongMemEval-V2 and MemoryCD ranking are statistical nulls (p=1.0 and p=.65). MemoryCD's tier-A result is a 21% MAE reduction with 3/3 ranking wins, which has no place on a score-difference axis and is reported in the text instead.

More competitors means more opportunities for a plausible-but-wrong entry to win on similarity, and behavioral testing is what distinguishes them; with two candidates a heuristic guesses right half the time by construction.

This decomposes exactly. Writing f for the fraction of families where the signals disagree, Δq for the gap in per-family selection quality, and Δp for the gap in success probability between a correctly and incorrectly curated family, the product $f \cdot \Delta q \cdot \Delta p$ reproduces the predicted advantage in every cell. The three factors move in opposite directions as cardinality grows—the contested fraction *falls* from 0.554 to 0.241 while Δq rises from 0.177 to 0.556 and Δp from 0.251 to 0.763—so the observed monotonicity is the net of a shrinking number of decisions that each matter more. The prediction is an upper envelope rather than a point estimate: measured advantages run 0.92–2.60× the predicted value, since the factorization assumes the three terms are independent.

Repeated selection. A single curation round understates the difference between signals, because a curation error persists into every later round. Over four rounds of a lifetime simulation (10 seeds, 2 families per round, budget 8) all signals start indistinguishable at round 1 (0.40–0.43), and by round 4 they have separated: causal credit ends at 0.375, keep-all at 0.363, similarity at 0.263, recency at 0.244, introspective judging at 0.225, and no memory at 0.031. Every heuristic *degrades* with use—recency loses 0.18 from its round-1 value—while credit is nearly flat. This is the same phenomenon as the mis-evolution result in §6 viewed over time: a curation rule that admits harmful entries does not merely fail to improve, it compounds. We note the honest detail that keep-all is statistically indistinguishable from credit here (0.363 vs. 0.375): with a budget of 8 and only 8 families seen, the bank never becomes selective enough for eviction to pay, which is precisely the small-bank regime the scaling law in §6 identifies as the one where curation is unnecessary.

M ADDITIONAL BREADTH: ELEVEN 2026 BENCHMARKS

Figure 11 puts the eleven additional 2026 benchmarks on one axis as the margin over the strongest rival on each benchmark's own native metric, sorted so that the nulls and the single loss sit on the same scale as the wins rather than in a separate table.

Level A (beat *all* rivals including full-context): EverMemBench 0.603 (rag 0.48, full-context 0.42); StreamMemBench cumulative 0.619 vs. raw-RAG 0.579 and 4× more robust under a protocol flip that only changes reset timing—direct evidence that the advantage grows with memory accumulation; EvolMem F1 0.945 vs. full-context 0.914 (p=.030); MemoryCD cross-domain rating MAE −21% with ranking wins 3/3. Level B (best memory system; full-context treated as a separate non-memory upper bound per convention): AgentMemoryBench 0.736 vs. RAG 0.296 (p=2.8×10^{-35}; oracle 0.823); BEAM OVR10 0.374 vs. dense RAG 0.343 (p=.014); plus LoCoMo (§5.4). Also: MATH-L5

answer selection 0.736 vs. self-consistency 0.683 (p=.002); HardMath-CCC on gpt-5.6-sol 0.985 vs. self-consistency 0.970 (oracle 1.0).

N IMPLEMENTATION AND REPRODUCTION DETAILS

Self-improvement on top of the loop. A three-lever iteration on the dialogue pipeline (metadata retrieval, semantic expansion, self-consistency selection (Wang et al., 2022)) improves CCC accuracy monotonically from 0.658 to 0.850 ($+29.1\%$, p=7.8×10^{-5}), leak-free and at negligible cost. The process remained designer-in-the-loop; automating the credit function itself remains open.

Probe construction. For editing tasks, probes are held-out questions from the same family (disjoint from evaluation by index; probes use questions[1:] and evaluation question[0]); for dialogue, the probe set is drawn from earlier turns and the task's own success signal. Budget B: 1–2 probes per entry; sweeps show saturation at B=2. **Eviction.** threshold 0 (non-positive credit); curation budget one entry per family. **HYBRID routing** is per-hop reliability estimation between CCC and typed traversal. **Cost accounting** counts all LLM calls including probes. **Infrastructure.** OpenAI-compatible shim endpoints per backbone; local bge-small embeddings for retrieval baselines; every request cached (deterministic temperature 0 where the protocol requires). All 60 scaling cells, poison sweeps, and cross-backbone replications re-run from cache at zero cost. **Judge protocols.** LoCoMo 4-cat J uses a fixed judge configuration shared by all systems on identical questions; F1/BLEU-1 are deterministic. **Baseline repairs** (§4). CHECK's released conflict repair returns `move_count`$= -1$ on every call and is therefore a no-op; we replaced it with the min-conflict repair its paper specifies. MemoryOS's released adapter mis-routed retrieval and scored 0.258 on LoCoMo; after repair it scores 0.476, and we compare against the repaired figure throughout.