PREDICTION IS NOT PROTECTION: WHAT MODELS FORGET DURING CONTEXT COMPACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-running LLM agents replace old conversations with compact memory. Protecting selected facts can improve overall recall while making information left unprotected less reliable. We study this tension through a paired evaluation that links each fact's pre-compaction forecast, untreated omission, and outcome after key-only protection. With supplied importance labels and controlled fact budgets, gains in overall important-fact recall coexist with losses on the same unprotected facts under self-selected, random, and cue-based policies. At the widest memory-loop budget, external protection retains 76–93% of raw-log recall at 10–29% of its storage, but a simple repetition rule remains competitive; self-ranking is model-dependent. Naturalistic tests qualify these findings. LoCoMo establishes neither general cue superiority nor a best selector. On 12 Multi-Session Chat dialogues at approximately matched measured storage, total point gains occur without detectable unprotected-fact loss. Additional key-free controls do not establish a benefit from stronger wording; self-selection retains a nominal gain over plain compaction. These limited-source results do not establish the absence of collateral cost or a universal selector advantage. Prediction accuracy, overall recall, and reliability on the same unprotected facts therefore require separate evaluation.

1 INTRODUCTION

Long-running agents accumulate constraints, decisions, and unfinished work. Compaction replaces older turns with a summary; omitted details can no longer be recovered from the discarded source. Model-directed compaction is already used in agent scaffolds (Li et al., 2026b; Verma, 2026; Du, 2026), despite reported failures such as erased safety constraints (Chen, 2026) and fabricated work histories (Tamba, 2026). A memory policy should therefore be evaluated not only by what it preserves, but also by what its intervention puts at risk.

Our central question is whether protecting selected facts can improve overall memory while making the *same unprotected facts* less reliable. Aggregate recall alone cannot answer this question: gains on protected facts can outweigh losses elsewhere. Omission prediction is a separate test. A forecast concerns an untreated summary; naming keys changes that summary, and predictive accuracy need not imply decision value (Elmachtoub & Grigas, 2022). Predictions may guide useful protection, but their accuracy alone establishes neither selector superiority nor the absence of collateral loss.

These questions connect complementary evaluation traditions. Self-knowledge studies test forecasts of a model's correctness (Kadavath et al., 2022; Lin et al., 2022; Yona et al., 2024); memory benchmarks test what remains answerable (Maharana et al., 2024; Wu et al., 2025). We connect forecasts to untreated omissions and protection outcomes, scoring identical fact cohorts in each policy contrast and measuring the storage actually written. Controlled logs with unguessable values make losses observable, at the cost of natural semantic importance. Naturalistic dialogue tests the limits of those findings.

Contribution and findings. We contribute a fact-level paired evaluation workflow, supported by a controlled workload generator, reference protection and memory-loop baselines, and analysis code. Its findings distinguish three questions. *First, which facts gain or lose?* Under controlled fact budgets and supplied importance labels, total recall gains coexist with losses on the same unprotected facts, including under random and external selection; the cost is not specific to

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.

self-report. *Second, does predictive information establish a better selector?* Self-ranking is heterogeneous, recorded cues predict many omissions, and hidden states add predictive value in two models, but the fitted protection rule has no established advantage over a simple repetition rule. *Third, how far does the result extend?* LoCoMo establishes neither general cue superiority nor a best selector. Multi-Session Chat (MSC) shows positive total point gains without detectable collateral loss at its sample size; key-free controls qualify the prompting explanation without proving key necessity. Self-selection's MSC gain does not survive three-policy correction. The contribution is thus a way to expose gains and losses that a single score can conceal, together with evidence delimiting its observed effects, not a universal protection rule or a proven mechanism (Section 6).

2 PAIRED EVALUATION OF PREDICTION AND PROTECTION

Materials and scale. Controlled episodes are multi-day user–assistant work logs with $N = 12$ target facts among ordinary chatter. Random ports, paths, and codes prevent prior-knowledge guessing. Repetition and interference assignments are randomised independently of item-list order; day and question order are separately shuffled. Measured positions remain correlated with day and recency: the six cues are predictors, not independent causal factors. Devops is the primary domain; wet-lab logs change keys and vocabulary, not construction. The primary measurement covers six backbone settings, 585 logs, and 7,004 label-bearing fact outcomes. Three open-weight backbones also receive interventions and five-block memory-loop evaluation. LoCoMo uses seven models, 24 windows, and 143 facts from nine sources; MSC supplies a second naturalistic boundary study. Reusing materials for diagnostics does not create independent datasets.

Paired tests and outcomes. Before compaction, the model ranks facts from most to least likely to be forgotten with the intended capacity stated. It then summarises without flags. We check reference-value survival and answer each question from the summary; correctness requires a complete-value match, and refusals count as incorrect. A full-log control distinguishes prior answering failures from compaction loss. Protection names keys, not values, and recompacts the original log (Figure 1). A correct full-log answer followed by an absent summary value is a write-side omission; a present value followed by a wrong answer is a stored-information use failure. The memory loop also checks retrieved context. These are observable stages, not neural mechanisms. Of controlled facts present in summaries, 84–100% are answered correctly, making writing the main loss stage here.

Summary survival measures writing; answer recall measures final utility. Prediction AUROC scores ranking, not retention rate; tables specify its outcome. Stored characters measure actual cost because equal requested limits need not yield equal writing. Recall changes are absolute: $+0.10$ means ten percentage points. For illustration, AUROC 0.70 is an ordering score, not a claim that 70% of facts survive. Recall 0.70 means 70% of the scored questions are answered correctly. Appendix F gives the complete metric guide.

Figure 1 illustrates the distinction: forecasting an omitted meeting time differs from requesting "keep the meeting time" and measuring its subsequent recall. The complete worked example remains in Appendix F.

Baselines, capacity, and uncertainty. The six settings comprise three frontier API models and Qwen3-14B, GLM-4-9B, and Microsoft's 14B Phi-4. Unstable character-limit difficulty motivates requested fact limits for the open-weight group (Table 1). Self-ranking is compared with a training-selected single cue and logistic regression over day, first/last mention position, repetition, interference, and value length. Each episode is held out of fitting and single-cue/sign selection; LoCoMo holds out sources. The repetition rule chooses lower recorded repetition. Controlled repetition/interference are generator annotations, not audited occurrence counts: final-day assignments may have no realised repetition or decoy. These baselines are annotation-assisted (Appendix A.1).

Qualification requires readable full-log questions and both retained and omitted outcomes. GLM-5.1's 0.235 outcome rate misses the 0.25 lower gate; it is retained for transparency, not positive cue evidence. LoCoMo failures are reported below; MSC re-passed the gate for its own backend. Intervals resample controlled episodes or LoCoMo sources; MSC pairs dialogues, and policy contrasts pair the same groups. Memory results receive five-block sensitivity. Reviewer-driven joint-readout, cohort, and multiplicity checks are post hoc (Appendix A).

3 PREDICTING UNTREATED OMISSIONS

Self-ranking and recorded cues. Self-ranking AUROC spans 0.462–0.614; 4 of six episode-clustered intervals include 0.5. Qwen3-14B is above chance, 0.614 [0.567, 0.659], and Phi-4 below, 0.462 [0.436, 0.487]. The latter is a reversed ordering under this interface, not an explanation of its cause. Table 1 reports valid-ranking sample sizes. For fact-limit survival, one cue reaches 0.607–0.641 and six cues 0.650–0.760; Qwen's repetition/six-cue scores are 0.621/0.760 versus self-ranking 0.614. Character-limit rows use a different outcome and show inconsistent cue performance, establishing neither universal external advantage nor a model-size trend. Cross-model and second-domain checks retain some predictive structure but identify no universally necessary cue (Appendices B.3 and B.4). These are diagnostic baselines, not a new algorithm.

Table 1: Controlled prediction AUROC. Rank logs: valid/all; cue fits hold out episodes and use all listed label-bearing facts. Character limits score answers; fact limits score survival. †: outcome rate 0.235 misses the 0.25 positive-cue gate. Common-valid contrasts: Appendix A.2.

model	limit	rank logs	facts	outcome	model [95% CI]	1 cue	6 cues	main cue
Qwen3.7-Max	160 chars	43/45	540	0.61	0.539 [0.488, 0.590]	0.515	0.627	-day
GLM-5.1†	40 chars	45/45	540	0.24	0.509 [0.476, 0.543]	0.422	0.442	-value_len
Doubao-Seed-2	90 chars	43/45	540	0.38	0.499 [0.470, 0.531]	0.395	0.401	-depth_frac
Qwen3-14B	keep 6	60/150	1800	0.49	0.614 [0.567, 0.659]	0.621	0.760	+repeats
GLM-4-9B	keep 6	124/150	1786	0.59	0.502 [0.473, 0.532]	0.607	0.650	-depth_frac
Phi-4	keep 4	136/150	1798	0.52	0.462 [0.436, 0.487]	0.641	0.696	-depth_frac

Interface and saved-state diagnostics. Only 40% of Qwen's original rankings are complete: cached outputs are keep-sized lists, not truncation/parser damage. Tying unlisted items across all 150 episodes gives AUROC 0.578; constrained ordering gives 0.560. Completion selection inflates but does not create the signal. These post-hoc, Qwen-only checks do not validate other interfaces (Appendix A.3). A weak stand-alone probe need not lack incremental value (Hewitt & Liang, 2019; Belinkov, 2022). A post-hoc joint linear-kernel readout with five outer and three inner episode folds improves over same-fold six-cue logistic fits by $+0.070$ [+0.053, +0.087] for Qwen and $+0.029$ [+0.018, +0.040] for Phi-4; GLM's gain is uncertain. Gains also hold against matched ridge. This shows incremental predictive value against the tested baselines, not causal self-knowledge, a better elicitation interface, or protection benefit. In particular, the ability of a trained readout to extract a signal does not establish that the model can report it or that using the signal to select protected keys improves memory. Both original and joint probe protocols are reported together in Appendix B.

4 WHAT CHANGES WHEN FACTS ARE PROTECTED?

4.1 CONTROLLED PROTECTION ON THE SAME IMPORTANT FACTS

The harness randomly marks half the facts important without showing that label to the compactor. Each selector protects three important keys, not values, under the same requested K-fact limit. We compare six-cue external, self-ranking, random, repetition, recency, and hindsight selection based on untreated omissions; all score survival on the same complete important-fact set (Table 2).

Table 2: Important-fact survival after three key-only flags, with the same episodes, important facts, and requested capacity. Repeat cue uses lower annotated repetition; hindsight uses untreated losses. Contrasts pair episodes; *: 95% interval excludes zero. External leads in point estimates but not reliably over repeat cue.

	rule choosing the 3 protected facts						paired difference	
model	external	repeat cue	recency	model	random	hindsight	external − random estimate [95% CI]	model − random estimate [95% CI]
Qwen3-14B	**0.718**	0.712	0.666	0.669	0.666	0.716	$+0.052$ [+0.031, +0.073]*	$+0.003$ [-0.020, +0.026]
GLM-4-9B	**0.778**	0.756	0.764	0.747	0.694	0.749	$+0.083$ [+0.052, +0.114]*	$+0.052$ [+0.019, +0.084]*
Phi-4	**0.643**	0.622	0.598	0.603	0.594	0.616	$+0.049$ [+0.007, +0.091]*	$+0.009$ [-0.030, +0.049]

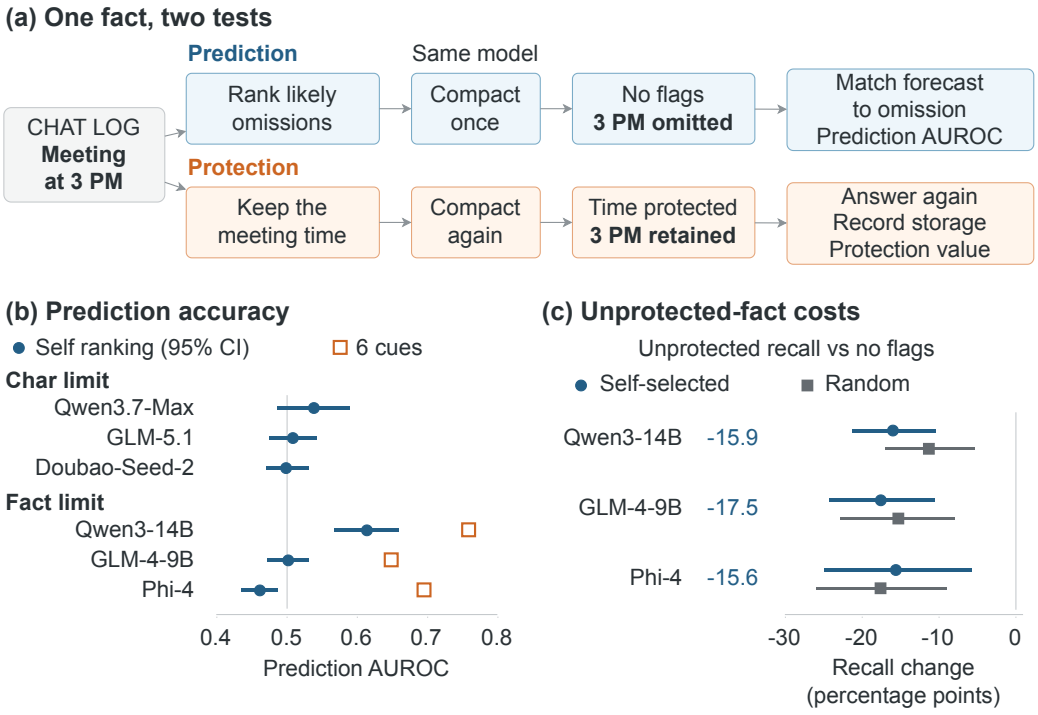


Figure 1: Prediction versus protection. (a) Prediction scores untreated omission; protection names keys, then tests recall and actual storage. (b) Character-limit rows score answers; fact-limit rows score survival. 4 self-ranking intervals include 0.5; Qwen is above, Phi-4 below. Squares: annotation-assisted six-cue estimates. (c) L1/L2 recall changes on identical unprotected facts versus no flags, in percentage points; blue labels denote self-selection. Policy-specific sets differ, so this is not a self-minus-random comparison. Intervals resample episodes; memory has only five blocks.

External-minus-random gains are $+0.052$, $+0.083$, and $+0.049$, with all episode intervals excluding zero. External-minus-repeat gains ($+0.006$, $+0.022$, $+0.021$) are uncertain, making repetition an essential control. Self beats random for GLM-4-9B, $+0.052$ [+0.019, +0.084], including valid rankings only (Appendix C.1); Qwen/Phi-4 contrasts remain uncertain. Hindsight protection has no established advantage over the external rule. External point estimates are at least as high, but paired differences remain uncertain. This neither proves equivalence nor falsifies an oracle ceiling. Moreover, policies preserve 91–100% of marked facts without reliably obeying the requested fact limit. We therefore measure actual outcomes and writing rather than assume a strictly zero-sum allocation or infer policy quality from untreated-loss prediction alone.

4.2 Memory loop: total benefit and unprotected-fact cost

Sixteen sessions share a project-tagged store. Character-bigram BM25 (Robertson & Zaragoza, 2009) fills a common 1,200-character read budget; answers use retrieved memory only. Five independent blocks give 80 sessions and 960 questions per setting. Write requests $K = 3, 6, 9$ are L0/L1/L2. We measure stored characters and exclude cross-session value collisions from scoring, though all questions compete during writing.

Summary bullets or sentences form retrievable entries, whereas raw-log entries are user–assistant turns. BM25 uses $k_1 = 1.5$ and $b = 0.75$, with the question's project anchor identifying the relevant session. The read-character budget is fixed across write policies. These implementation choices make the stored-content, retrieval, and answer checks comparable within this loop (Appendix C.2).

For important, value-unique facts I, policy-a flags F_a, unprotected set $U_a = I \setminus F_a$, recall R_a, and no-flags policy 0, define

$$\Delta_a^{\text{total}} = R_a(I) - R_0(I), \qquad \Delta_a^{\text{unprotected}} = R_a(U_a) - R_0(U_a).$$

The first effect measures total benefit; the second measures the change for unprotected facts, which is a cost when negative. Crucially, U_a is used in *both* arms of the second contrast. Otherwise, a change in which easy or difficult facts are scored could look like a policy effect. Different selectors may leave different facts unprotected, so their respective unprotected effects are not a direct selector comparison. Appendix C.3 gives the decomposition and common-cohort comparisons.

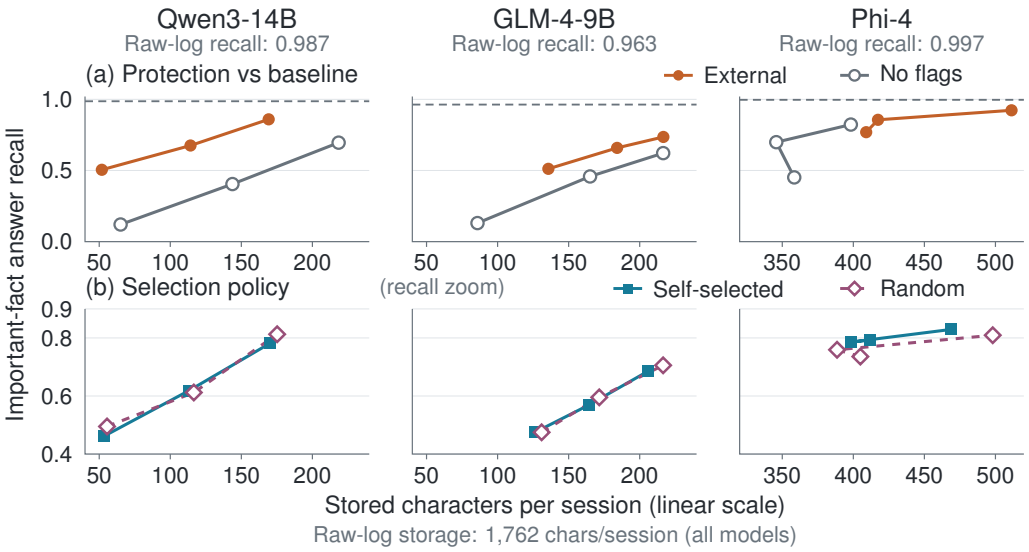


Figure 2: Measured storage and recall: (a) external versus no flags; (b) self versus random. Linear storage windows each span 200 characters; Phi-4's window is shifted. Row (b) zooms recall to 0.4–0.9. Dashed horizontal lines in (a) mark raw-log recall; raw-log recall and storage are also listed numerically. Points are unjittered; lines follow $K = 3, 6, 9$, even when storage decreases. External uses six recorded cues. Overlap does not establish equivalence; paired uncertainty and L0 saturation qualify comparisons.

Table 3: L1/L2 matched-cohort recall changes with 95% episode-clustered intervals. Both arms score identical important, value-unique facts. Each unflagged row uses that policy's unflagged set also for no flags; differing sets preclude cross-selector inference.

comparison	facts scored	Qwen3-14B	GLM-4-9B	Phi-4
external − none	all important	+0.217 [+0.171, +0.264]	+0.157 [+0.100, +0.211]	+0.129 [+0.070, +0.190]
self − none	all important	+0.151 [+0.113, +0.189]	+0.087 [+0.031, +0.143]	+0.047 [-0.013, +0.108]
random − none	all important	+0.162 [+0.121, +0.201]	+0.110 [+0.060, +0.162]	+0.023 [-0.036, +0.085]
self − none	same unflagged	−0.159 [-0.213, -0.104]	−0.175 [-0.242, -0.105]	−0.156 [-0.248, -0.057]
random − none	same unflagged	−0.113 [-0.169, -0.053]	−0.152 [-0.228, -0.080]	−0.175 [-0.259, -0.090]
external − none	same unflagged	−0.231 [-0.291, -0.173]	−0.141 [-0.222, -0.064]	−0.103 [-0.169, -0.038]

Benefit and cost coexist. External adds +0.129–+0.217 total recall at pooled L1/L2. Self/random total point estimates also improve, though not all intervals exclude zero. Self, random, external, and repetition policies all lose recall on their respective same unprotected facts (Table 3). This is neither self-report-specific harm nor negative total utility. Redundant flags are one candidate explanation: self flags overlap untreated retained facts more than external, but not reliably more than random. No self-specific redundancy mechanism is established. Losses with near-matched or larger mean write size exclude a uniform reduction-in-writing account, not all length/format interactions (Appendix C).

Storage trade-off and inferential limits. At the widest limit, external retains 76–93% of raw-log recall at 10–29% of storage (Figure 2). Raw logs answer 0.963–0.997 of important questions at 1762 characters/session and remain preferable when feasible. At matched measured storage, no-flags model compaction exceeds value-bearing-turn retention by +0.159–+0.281 recall and recent turns by +0.172–+0.343 (Table 17). This supports constrained compaction, not a uniformly equal-byte

flag effect. L0 requests three facts and flags three, so selector comparisons use L1/L2. Self's pooled unprotected loss survives each block deletion, but five blocks allow a minimum two-sided sign-flip $p = 0.0625$; episode intervals are not five independent confirmations. Lexical retrieval, short tagged stores, and read-budget pressure limit generality (Appendix C.4).

The extractive and recent-turn baselines receive the measured write budget of unprotected model compaction at each level. Extraction detects turns containing digits, including decoys and numeric filler; it neither matches generator-specific fact-introduction phrases nor receives the correct value list. These are measured-cost baseline comparisons. Flagging policies instead report the characters actually written, so their differences from no flags need not be equal-byte effects.

Retrieval checks distinguish missing values from misbinding. With four episode summaries in one context and no per-entry tags, $P(\text{recovered} \mid \text{kept})$ falls from 0.965 and 0.843 to 0.872 and 0.784 for two backbones; Phi-4 remains at 0.925. Among in-summary failures, 24–42% answer with another project's value for the same key, exposing a binding failure rather than simple omission. Recovery when the value is absent stays at 0.005–0.017. The tagged loop alleviates this ambiguity. A fourfold tighter read budget, 300 instead of 1,200 characters, was tested on two backbones: no cell changes by more than 0.04, and the direction is upward. The third model lacks this replication. These controls characterise a lexical, tagged store, not arbitrary large-scale retrieval (Appendix C.4).

5 Transfer to naturalistic dialogue

5.1 LoCoMo: prediction, protection, and grading

LoCoMo's LLM-generated, human-edited conversations have semantic structure absent from random-value logs (Maharana et al., 2024). All seven models receive the same 24 windows and 143 facts from nine sources, with source holdout/resampling and varying valid-ranking counts. This is an external boundary study, not another controlled replication.

Table 4: LoCoMo: 143 facts, 24 windows, nine sources. Valid n/w/s counts self-report facts/windows/sources; cue fits use the full cohort with source holdout. Answer|survived conditions on exact survival. Common-valid and source-deletion checks: Appendix D.

model	valid n/w/s	survival	model ranking [95% CI]	one cue	six cues	full-log	Answer|survived
Qwen3-14B	143/24/9	0.245	0.484 [0.386, 0.577]	0.485	0.539	0.839	0.257
GLM-4-9B	75/13/8	0.392	0.490 [0.351, 0.653]	0.648	0.655	0.839	0.446
Phi-4	143/24/9	0.231	0.467 [0.330, 0.547]	0.536	0.369	0.867	0.606
DeepSeek-V4-Pro	135/23/9	0.336	0.618 [0.478, 0.716]	0.599	0.614	0.860	0.479
Qwen3.7-Max	143/24/9	0.231	0.592 [0.465, 0.710]	0.514	0.555	0.874	0.333
GLM-5.1	143/24/9	0.224	0.547 [0.442, 0.666]	0.379	0.363	0.839	0.344
Doubao-Seed-2	143/24/9	0.308	0.536 [0.428, 0.639]	0.387	0.484	0.860	0.636

Self-ranking is 0.467–0.490 for 9–14B models and 0.536–0.618 for four frontier models; all seven source-clustered intervals include 0.5. In three frontier settings, self exceeds both cue baselines in point estimates. All unprotected answer rates (0.063–0.203) miss the 0.25 lower gate. Nine sources and failed gates make this exploratory, establishing neither general cue superiority nor a model-size law.

Three models receive interventions at two flag budgets. External, self, and random increase all-fact recall over no flags by $+0.119$–$+0.287$, but requested capacity does not bind writing. External-minus-random/self/mention contrasts are uncertain. DeepSeek self/random write 0.96–0.98 times no-flag characters with $+0.133$–$+0.287$ recall: a near-equal-write prompting benefit, not a best-selector result. Same-unprotected-fact self effects range from -0.105 to 0.000; some nominal intervals are negative, but multiplicity and source-sign-flip checks give mixed support. Hindsight does not reliably beat random, and different point-estimate winners do not establish a population ranking reversal (Appendix D.2).

A blinded audit grades 420 archived answers (60/model), weighted to 1001 responses. Strict-grader agreement is 0.969 [0.952, 0.987]; all 15 disagreements are false negatives. Human grading raises accuracy without restoring general cue advantage: frontier six-cue-minus-self AUROC is -0.165

[-0.259, -0.029], averaging within-model AUROCs rather than score scales. This validates answer grading, not full-sample summary-survival labels (Appendix D.3).

5.2 Multi-Session Chat at matched measured storage

LoCoMo leaves two things unresolved: requested capacity did not bind actual writing, and its unprotected-fact effects were mixed. We therefore ran four policies over Multi-Session Chat (MSC), crowdworker five-session dialogues (Xu et al., 2022), with storage controlled by measurement rather than by request. 12 development dialogues each carry six human-checked facts. Every policy writes one free-prose memory record per dialogue aimed at 200 reference tokens of a pinned reference tokenizer, accepting 190–210. A draft outside the band receives up to two rewrites whose feedback contains only measured length, never answer outcomes; the reader receives the first in-band draft or otherwise the one closest to target. The writer sees the transcript and protected key names, without values supplied in the protection instruction; the reader sees only the record. This yields 288 logical fact conditions from 264 unique answers, because identical protected-key sets within a dialogue share one realisation. The single frontier backend first re-passed the full-log gate on the same 72 questions (per-source recall 1.00, or 0.986 if the undecided answer is counted wrong); the earlier backend's qualification was not carried over.

Storage matching and total gain. 42 of 48 records land in the accepted band. For the paired comparison, the relevant quantity is the spread between the four policies of one dialogue: its median is 10 reference tokens. It meets the tolerance implied by the band for 11 of 12 dialogues, reaching 25 on the remaining one. We retain that dialogue rather than dropping it: "matched storage" therefore means within tolerance in 11 of 12 cases, not uniformly equal cost.

Table 5 reports total and unprotected-fact effects. All three protection policies raise total recall over no protection. Self-selection has the largest point gain, $+0.231$, with 10 dialogues improving and 2 worsening. Its exact two-sided signed-rank $p = 0.020$ becomes 0.059 after accounting for the three policies: this is a directional result, not a corrected significant one. Setting every undecided label to wrong gives $+0.194$ with $p = 0.065$.

Table 5: MSC human-scored recall: 12 dialogues, 288 logical conditions. Total effects compare all six facts; unprotected effects compare the same policy-specific four facts against no protection. w/l/t: dialogue wins/losses/ties. p: exact two-sided paired signed-rank, uncorrected across three policies.

policy	recall, all six	recall, protected two	Δtotal	w/l/t	p	Δunprotected (p)
none	0.507	–	–	–	–	–
random	0.597	0.833	$+0.094$	7/3/2	0.371	$+0.000$ (1.00)
self	0.735	0.913	$+0.231$	10/2/0	0.020	$+0.097$ (0.30)
text cue	0.586	0.750	$+0.083$	8/3/1	0.344	$+0.042$ (0.72)

No detectable unprotected cost. Each policy leaves four facts unprotected; compared against no protection on those same four facts, no point effect is negative: random $+0.000$, self $+0.097$, and text cue $+0.042$. The collateral loss observed under the controlled budget is not detected here. With 12 paired dialogues and four facts per comparison, a loss of one fact in at least 6 dialogues with no offsetting dialogue would have been detected at the 0.05 level. Smaller losses remain possible, so this is non-detection rather than evidence of no cost. As a selection diagnostic, the model's two picks were facts that untreated compaction actually lost 0.583 of the time. This is a hit rate against observed omissions, not a ranking evaluation.

Key-free controls separate wording from selection. Two additional arms test *plain* compaction without the key-list mechanism and *emphatic* compaction, which adds a completeness request to the original empty-key *none* prompt. Over the same 12 dialogues, emphatic-minus-none is $+0.053$ ($p = 0.59$); self-minus-plain is $+0.164$ (10/12 wins, $p = 0.036$). Random-minus-none is $+0.094$ ($p = 0.371$), and self-minus-random $+0.136$ ($p = 0.078$). Thus stronger wording has no established benefit, while selected keys retain a positive nominal contrast against plain compaction. These comparisons are exploratory, not proof that keys are necessary or self-selection is superior after

correction. The new arms were separately generated and single-rated; measured storage is imperfectly matched (Appendix H).

Annotation reliability and remaining ambiguity. For the original four arms, a second reviewer independently relabelled 96 stratified decidable answers: correctness agreement is 95.8%, $\kappa = 0.89$ [0.77, 0.98]. Retention agrees on all 47 mutually decidable answers, but 84 retention and 8 correctness judgments remain undecided among 264 unique answers because memory fails to bind facts to sessions. This limits retention evidence; full reviewer coverage and resolutions remain in Appendix G.

Both self-selection and the text cue see the questions when choosing what to protect, which is informative but not deployment-realistic. The 3 answers that violated the answer schema were retained and flagged, not repaired or resampled. Provider-side reasoning accompanied 96 of 348 responses; it never entered stored records or reached readers. These arms consequently match storage cost, not provider compute. Appendix G collects the full execution and annotation account.

LoCoMo supports prompting benefit but not general cue advantage or consistent fixed-cost unprotected losses; MSC finds gain without detectable cost. Only controlled logs demonstrably show collateral loss. Other dialogue candidates failed qualification or task-fit screening and add no confirmatory evidence (Appendix E).

6 Discussion and limitations

Comparing a new memory policy. Compare no protection, random protection, and a candidate on identical histories with known answers and importance labels. Report survival, overall important-fact recall, same-item unprotected effects, and actual storage. Selector superiority requires comparable resources, not merely identical capacity requests. Include a key-free instruction control when attributing gains to named facts. This evaluates a policy; it does not choose the best policy without testing.

Diagnosing a weak result. The full-log, summary-content, retrieved-context, and answer checks locate the observed failure stage before changing the model or retriever. Weak self-ranking should be separated from lack of predictive information: the joint readout adds value in two models, but has not been tested as a protection policy. The workload generator, key-only protection prompts, memory-loop reference implementation, and analysis code support these comparisons (Appendix A.1). The numerical supplement reproduces saved results; testing a new model requires a generation backend and fresh validity checks.

Limits of reuse. The controlled design makes omissions observable at the price of realism. Importance is supplied by the harness, not inferred from users, and random values and independently assigned manipulations do not reproduce natural semantic importance. The memory loop is a tagged summary store with lexical retrieval, not a named production framework. The repetition heuristic uses supplied annotations, not a verified literal mention counter. Phi-4 exceeds requested fact caps; five blocks limit memory-system inference. LoCoMo has few sources, low unprotected answer rates, non-binding capacity, and only one frontier model with both measurement and intervention. Its unlicensed mirror is not redistributed. The human audit cannot replace unavailable complete-summary labels.

MSC (Section 5.2) binds storage by measurement, but uses crowdworker dialogue rather than production logs, 12 development sources, one backend, and policies that see the questions before choosing. Labels are double-rated on a stratified sample with $\kappa = 0.89$ for answer correctness, yet 84 retention judgments stay undecided after two independent passes. Its central negative claim is a non-detection at $n = 12$: under the stated one-fact-loss, no-offset scenario, at least 6 affected dialogues are needed for detection, so a smaller collateral cost is not excluded. One dialogue's between-policy storage spread exceeds the band's tolerance, and hidden provider-side reasoning means the arms match stored tokens rather than compute. The added key-free controls also leave cross-run and residual storage differences; their nominal tests do not establish equivalence or identify a unique mechanism.

Questions left open. Ranking-interface sensitivity beyond Qwen requires new elicitation experiments. The incremental readout is a post-hoc grouped-CV diagnostic, not the original LOEO protocol, causal introspection, or a tested memory controller. Whether the collateral cost of protection depends on budget tightness, as the controlled and MSC settings jointly suggest, is a hypothesis this paper cannot resolve. It requires the same policies at several budgets in one naturalistic setting and cross-model replication; changing settings also changes materials and protocols, not just storage pressure. More independent naturalistic sources and actual equal-cost interventions are needed for stronger external claims. Multiplicity sensitivities do not provide global confirmatory error control. These limitations remain even when numerical reproduction succeeds.

7 Related work

Self-knowledge and calibration. Models can predict some of their own correctness (Kadavath et al., 2022) and verbalise uncertainty (Lin et al., 2022), yet may communicate intrinsic uncertainty poorly (Yona et al., 2024). Other work tests model-specific metacognition (Wang, 2026; Moran & Whiting, 2026), behavior-predictive explanations (Mayne et al., 2026), hidden-state forecasts (Ahdritz et al., 2024; Ashok & May, 2025), and predictions of the model's own actions (Laine et al., 2024; Naphade et al., 2026). Those studies forecast outputs. Ours is a judgment-of-learning design (Nelson & Dunlosky, 1991; Dunlosky & Metcalfe, 2009): it forecasts per-fact loss caused by the model's next summary, then intervenes on that event.

Compaction and agent memory. Long-horizon agents use memory streams and recursive summaries (Park et al., 2023; Zhang et al., 2022); unlike one-shot prompt compression (Jiang et al., 2023; 2024), self-compaction determines future context. Systems use model-triggered compaction (Li et al., 2026b; Verma, 2026), editing or preservation commitments (Li et al., 2026c; Trukhina & Vashkelis, 2026), parallel compaction (Cim et al., 2026), trained policies and tools (Liu et al., 2026; Li et al., 2026a; Zhang et al., 2025; Yu et al., 2026), or paging (Packer et al., 2023). Surveys and rate–distortion analyses map this space (Du, 2026; Colaco & Lahjouji, 2026); benchmarks test recall (Maharana et al., 2024; Wu et al., 2025), while failures include erased constraints, fabricated work, and invalid summaries (Chen, 2026; Tamba, 2026; Chen et al., 2026). QA-based summary evaluation also tests factual consistency (Wang et al., 2020). We connect pre-event forecasts to omission and separate intervention outcomes on matched fact cohorts.

Selective retention below the text channel. KV eviction and smoothing manage bounded memory (Zhang et al., 2023; Xiao et al., 2024; Li et al., 2024; Vemula & Gajula, 2026), but evict tokens; here the model drops facts. Both expose positional effects (Liu et al., 2024).

8 Conclusion

Overall memory gains can conceal losses on the same unprotected facts. Our controlled experiments demonstrate this combination under supplied importance labels and fact budgets, alongside substantial storage savings. The finding does not establish a universal trade-off: LoCoMo leaves selector advantages unresolved, and MSC shows positive total point gains without detectable collateral loss at its sample size. Key-free controls do not prove key necessity; budget dependence and the loss mechanism remain open, while session ambiguity limits retention judgments. The practical lesson is to test prediction, total benefit, and same-fact collateral effects separately rather than infer reliable protection from one prediction or recall score.

AI Use Statement

Generative AI tools assisted language polishing, figure presentation, and editorial feedback on the interpretation of reported results. The authors take full responsibility for the final text and scientific content.

REFERENCES

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*, 2024.

Dhananjay Ashok and Jonathan May. Language models can predict their own behavior. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Shiyang Chen. Governance decay: How context compaction silently erases safety constraints in long-horizon LLM agents. *arXiv preprint arXiv:2606.22528*, 2026.

Zhuofu Chen, Rui Pan, Yinwei Dai, and Ravi Netravali. Slipstream: Trajectory-grounded compaction validation for long-horizon agents. *arXiv preprint arXiv:2605.08580*, 2026.

Musa Cim, Burak Topcu, Chita Das, and Mahmut Taylan Kandemir. Parallel context compaction for long-horizon LLM agent serving. *arXiv preprint arXiv:2605.23296*, 2026.

Ashwin Gerard Colaco and Nada Lahjouji. What to keep, what to forget: A rate–distortion view of memory compaction in LLMs and agents. *arXiv preprint arXiv:2607.08032*, 2026.

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

John Dunlosky and Janet Metcalfe. *Metacognition*. SAGE Publications, 2009.

Adam N. Elmachtoub and Paul Grigas. Smart "predict, then optimize". *Management Science*, 68(1): 9–26, 2022. doi: 10.1287/mnsc.2020.3922.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13358–13376, 2023. doi: 10.18653/v1/2023.emnlp-main.825.

Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LongLLMLingua: Accelerating and enhancing LLMs in long context scenarios via prompt compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1658–1677, 2024. doi: 10.18653/v1/2024.acl-long.91.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and AI: The situational awareness dataset (SAD) for LLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ruoran Li, Xinghua Zhang, Haiyang Yu, Shitong Duan, Xiang Li, Wenxin Xiang, Chonghua Liao, Xudong Guo, Yongbin Li, and Jinli Suo. Self-Memory policy optimization for long-horizon agents. *arXiv preprint arXiv:2603.00680*, 2026a.

Tianjian Li, Jingyu Zhang, William Jurayj, Xi Wang, Chuanyang Jin, Mehrdad Farajtabar, Eric Nalisnick, and Daniel Khashabi. Self-Compacting language model agents. *arXiv preprint arXiv:2606.23525*, 2026b.

Xiaochuan Li, Ryan Ming, Meng Chu, Shuai Shao, Rong Jin, and Chenyan Xiong. ACM: Agentic context management for long horizon tasks. *arXiv preprint arXiv:2607.23809*, 2026c.

Yuhong Li, Yingbing Huang, Bowen Yang, Bharat Venkitesh, Acyr Locatelli, Hanchen Ye, Tianle Cai, Patrick Lewis, and Deming Chen. SnapKV: LLM knows what you are looking for before generation. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22947–22970, 2024. doi: 10.52202/079017-0722.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024. doi: 10.1162/tacl_a_00638.

Ziyan Liu, Zhezheng Hao, Yeqiu Chen, Hong Wang, Jingren Hou, Ruiyi Ding, Yongkang Yang, Wence Ji, Wei Xia, and Feng Liu. Meta-Cognitive memory policy optimization for long-horizon LLM agents. *arXiv preprint arXiv:2605.30159*, 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. doi: 10.18653/v1/2024.acl-long.747.

Harry Mayne, Justin Singh Kang, Dewi Gould, Kannan Ramchandran, Adam Mahdi, and Noah Y. Siegel. A positive case for faithfulness: LLM self-explanations help predict model behavior. *arXiv preprint arXiv:2602.02639*, 2026.

M. Moran and Mark Whiting. LLMs show no signs of individuated metacognition. *arXiv preprint arXiv:2605.24299*, 2026.

Atharv Naphade, Samarth Bhargav, Sean Lim, and Mcnair Shah. Me, myself, and π: Evaluating and explaining LLM introspection. *arXiv preprint arXiv:2603.20276*, 2026.

Thomas O. Nelson and John Dunlosky. When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-jol effect". *Psychological Science*, 2(4): 267–270, 1991.

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.

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.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2009. doi: 10.1561/1500000019.

Hiroki Tamba. Compaction as epistemic failure: How agentic LLM tools fabricate confirmed results from killed processes. *arXiv preprint arXiv:2607.13071*, 2026.

Natalia Trukhina and Vadim Vashkelis. Compress the context, keep the commitments: A formal framework for verifiable LLM context compression. *arXiv preprint arXiv:2605.17304*, 2026.

Maruthi Vemula and Neeraj Praneeth Gajula. Eviction as estimation: A fixed-lag smoothing view of test-time memory, and when measuring beats accumulating. *arXiv preprint arXiv:2607.24667*, 2026.

Nikhil Verma. Active context compression: Autonomous memory management in LLM agents. *arXiv preprint arXiv:2601.07190*, 2026.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, 2020. doi: 10.18653/v1/2020.acl-main.450.

Jason Z. Wang. MIRROR: A hierarchical benchmark for metacognitive calibration in large language models. *arXiv preprint arXiv:2604.19809*, 2026.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Jing Xu, Arthur Szlam, and Jason Weston. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 5180–5197, 2022.

Gal Yona, Roee Aharoni, and Mor Geva. Can large language models faithfully express their intrinsic uncertainty in words? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7752–7764, 2024. doi: 10.18653/v1/2024.emnlp-main.443.

Yi Yu, Liuyi Yao, Yuexiang Xie, Qingquan Tan, Jiaqi Feng, Yaliang Li, and Libing Wu. Agentic memory: Learning unified long-term and short-term memory management for large language model agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 21457–21483, 2026. doi: 10.18653/v1/2026.acl-long.981.

Yusen Zhang, Ansong Ni, Ziming Mao, Chen Henry Wu, Chenguang Zhu, Budhaditya Deb, Ahmed Awadallah, Dragomir Radev, and Rui Zhang. SummN: A multi-stage summarization framework for long input dialogues and documents. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1592–1604, 2022. doi: 10.18653/v1/2022.acl-long.112.

Yuxiang Zhang, Jiangming Shu, Ye Ma, Xueyuan Lin, Shangxi Wu, and Jitao Sang. Memory as action: Autonomous context curation for long-horizon agentic tasks. *arXiv preprint arXiv:2510.12635*, 2025.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

A Protocol and statistical checks

This appendix specifies the common evaluation and its statistical units. Appendix B groups prediction diagnostics; Appendix C groups protection and memory controls; Appendix D contains the complete LoCoMo boundary study. Qualification attempts outside those two materials are recorded separately in Appendix E.

A.1 Controlled logs, labels, and prompts

The generator draws values, checks within-episode collisions, and places facts into a multi-day log. Independent item subsets are assigned repetition and interference conditions. Day assignment and question presentation receive separate permutations. Actual first- and last-mention positions are measured from the rendered log; they are not independent of day or each other. The devops and wet-lab variants share this protocol, not vocabulary. Project anchors distinguish sessions in the memory loop. The random important-fact labels are supplied to selectors, not inferred from user preferences.

Recorded cues versus realised mentions. The controlled `repeats` and `interference` fields record assigned conditions. A confirmation or decoy is emitted only on a later day; a final-day assignment can therefore remain unrealised. Reconstructing the three main open-weight inputs reproduces all saved feature vectors and identifies 106 of 600 nominally repeated facts per model with no added confirmation; these are the same 150 underlying logs, not independent replications. The single-field selector consequently tests the recorded repetition cue, not an audited literal counter. Runner/table labels such as *mention* and *mention-count* refer to that policy. We retain its observed outcomes without changing the feature values or asserting that a literal counter would give the same results. LoCoMo instead derives repetition from reference-value occurrences in the dialogue. The numerical audit explains this distinction; no corrected-counter policy was rerun.

The two outcomes are `in_summary`, a bounded value match in the summary, and `success`, correctness of the answer from that summary or retrieved memory. Full-log controls test readability before compaction. An explicit refusal is not counted correct merely because it echoes a target value. The naturalistic answer audit is described in Appendix D.3; its labels do not replace summary-survival labels.

Exact prompt templates are included with the numerical supplement's reference harnesses. In `leg_a.py`, `RANK_PROMPT_SELECT` and its character-limit counterpart request a complete most-to-least-forgotten order and state the intended limit; `SUMMARY_PROMPT_SELECT` requests at most K concrete facts, while `QA_PROMPT` asks for an answer using the summary only, or an explicit unknown. The controlled templates are in Chinese; these descriptions are summaries, not verbatim English prompts. `intervene.py` adds key-only protection instructions. `real_arm.py` and `real_interv.py` provide the LoCoMo adaptations. These entrypoints document the recorded experiments; model generation is disabled in the anonymous numerical replay.

Reusing the evaluation components. `materials.py` constructs controlled logs and fact questions; `leg_a.py` defines pre-compaction prediction, full-log and summary-answer controls. `intervene.py` supplies the key-only protection comparisons, and `memory_arm.py` adds storage, retrieval, and answering. `metrics.py` and `reviewer_audit.py` provide grouped comparisons and same-item accounting. These components support three checks: compare forecasts against untreated outcomes, compare policies on fixed facts and measured costs, and diagnose where a failed answer lost access to its evidence. Adapting them to a new domain requires fact references, importance labels, grading validation, and an appropriate model backend. The distributed numerical-only client intentionally blocks fresh generation; it is not an immediately deployable memory service.

Programmatic truncation is excluded from the primary measurement: it makes omission largely a position label rather than a model's selection decision. Verbal probabilities also collapsed to few distinct values, motivating the ranking interface. Character limits led to topic-only or near-complete outputs in some open-weight settings; requested fact limits supplied a usable outcome range but were not always obeyed. A requested fact limit is therefore not treated as an exact byte constraint.

A.2 Qualification, effective samples, and exploratory inference

The base-rate and readability checks were fixed before the corresponding model runs. Subsequent reviewer-driven analyses were designed after feedback and initial inspection: they are post hoc, not preregistered confirmations. The main self-ranking intervals include 4 containing 0.5, 1 above it, and 1 below it. Counts are checked from the endpoints and in Figure 1's text.

Table 6: Design diagnostics for the six main cells.

backbone	base	ICC	design eff.	eff. n	episodes	full-log	strict order	distinct
Qwen3.7-Max	0.61	0.46	6.04	89.46	45	1.00	0.96	9
GLM-5.1	0.24	0.65	8.10	66.70	45	1.00	1.00	10
Doubao-Seed-2	0.38	0.85	10.37	52.08	45	1.00	0.96	12
Qwen3-14B	0.49	0.00	1.00	1800.00	150	1.00	0.40	6
GLM-4-9B	0.59	0.20	3.20	558.95	150	0.98	0.83	2
Phi-4	0.52	0.23	3.56	505.07	150	1.00	0.91	4

In Table 6, effective n is item count divided by an outcome-based design effect. It is not an AUROC power calculation. The distinct-probability and strict-order columns describe different interface failures; neither converts these settings into independent null replications. Table 7 separates full construction size from valid ranking size. Cue fits train with the original held-out-episode protocol; common-valid contrasts score the same items as self-report. The displayed 90% interval inclusion in $[0.45, 0.55]$ uses an illustrative margin chosen post hoc. It is not preregistered equivalence evidence or a claim of absent metacognition. We do not infer AUROC power from an outcome-only effective sample size: score ties, class balance, and score–outcome dependence also matter.

Table 7: Valid ranking counts and exploratory uncertainty checks. The last two columns are cue-minus-self AUROC on common valid items, with 95% episode intervals.

model	valid/all logs	rank facts	self 90% CI	inside band	1 cue − self	6 cues − self
Qwen3.7-Max	43/45	516	[0.496, 0.583]	no	−0.029 [-0.104, +0.060]	+0.082 [+0.020, +0.149]
GLM-5.1	45/45	540	[0.482, 0.536]	yes	−0.087 [-0.139, -0.024]	−0.067 [-0.119, -0.002]
Doubao-Seed-2	43/45	516	[0.475, 0.525]	yes	−0.102 [-0.150, -0.037]	−0.097 [-0.143, -0.028]
Qwen3-14B	60/150	720	[0.578, 0.651]	no	+0.016 [-0.050, +0.080]	+0.135 [+0.076, +0.192]
GLM-4-9B	124/150	1475	[0.478, 0.526]	yes	+0.089 [+0.038, +0.142]	+0.135 [+0.091, +0.180]
Phi-4	136/150	1630	[0.441, 0.483]	no	+0.179 [+0.133, +0.227]	+0.231 [+0.190, +0.274]

Ratio contrasts in the post-hoc audit use 4,000 paired cluster resamples. The exploratory two-sided tail value is twice the smaller bootstrap mass on either side of zero, with a finite-sample correction; it is not an exact randomization p value. Holm sensitivities are reported within named families: six chance tests; 12 common-valid controlled cue contrasts; three full and three valid-only self-minus-random contrasts; 24 per-budget contrasts per memory outcome; 12 pooled memory contrasts per outcome; 18 selector contrasts per cohort view; 12 overlap contrasts; 14 LoCoMo cue contrasts; 18 LoCoMo same-item policy contrasts (six additional valid-only self contrasts); 30 transfer contrasts; 36 matched-protocol leave-one-cue-out transfer contrasts; and six incremental probe contrasts. These post-hoc families do not supply global confirmatory error control. Other intervals remain nominal where stated; a star alone is not multiplicity-adjusted evidence.

For prediction contrasts, bootstrap uncertainty is conditional on fitted out-of-fold scores; it does not rerun learner selection within every bootstrap draw. Source-deletion refits and block-deletion checks are sensitivity diagnostics, not cures for few independent groups. In particular, five-block sign flips assume independent symmetric block contributions and are not randomized policy-assignment tests.

A.3 Ranking-interface robustness

On the same 150 Qwen3-14B episodes and frozen 1,800 labels, cache reconstruction matches all original completion flags. The 90 incomplete outputs contain no truncation, illegal ID, refusal, or parser loss: 81 are valid lists of exactly `keep`=6 IDs and nine contain 11, so this is total-order task misunderstanding. Retaining the stated partial order and tying every unlisted item gives 0.578 [0.548, 0.606] AUROC over all episodes. A post hoc constrained permutation gives 0.560 [0.530, 0.588]; on the original 60 complete episodes it gives 0.598 versus 0.614 (mean Kendall τ =0.707). A direct top-3 prompt is 150/150 compliant and improves retrospective loss-capture over random by 0.076 [0.055, 0.097]. Thus completion selection inflates but does not create Qwen's weak signal. These Qwen-only post hoc interfaces do not replace the common protocol or establish policy benefit. Deployment effects are evaluated separately; interface compliance does not establish policy value.

B Prediction diagnostics

These analyses answer variants of the pre-compaction prediction question. They do not add independent datasets or establish the value of a deployed protection policy.

B.1 Stand-alone hidden-state probe

Kernel (dual) ridge on mean-pooled hidden states over each fact's value token span, layer and λ selected inside each training fold, position residualised out by an algebraic identity on the Gram matrix, float64 normalised Grams with a Cholesky solve and an LU fallback, training labels centred

per fold (equivalent to an unpenalised intercept). Reported against a within-episode label-permutation null, whose mean is 0.518 rather than 0.500 for this estimator. Three synthetic self-tests — true signal, pure noise, pure position label — fix the pipeline's sensitivity, its leak-freedom, and the effectiveness of the residualisation.

Table 8: Hidden-state probe on the compaction forward pass, under both labels. "resid." projects token position out of the representation. The probe beats the model's own report on all three backbones for end-to-end recovery and on two of three for survival; its comparison to the nested one-feature baseline is mixed, and it never beats the six-feature fit as a stand-alone predictor. This does not test incremental information; the joint diagnostic is in Appendix B.2. * marks a paired self-report difference whose episode-clustered CI excludes zero; per-item probe scores were not archived for a corrected paired one-feature CI.

backbone	label	layer	probe	resid.	self-report	1-feat.	surface	− self	− 1 feat.
Qwen3-14B	recovered	30	0.701	0.689	0.618	0.625	0.764	$+0.089^*$	$+0.063$
GLM-4-9B	recovered	20	0.663	0.597	0.511	0.596	0.637	$+0.081^*$	$+0.002$
Phi-4	recovered	40	0.684	0.615	0.463	0.641	0.697	$+0.155^*$	-0.025
Qwen3-14B	kept	30	0.699	0.687	0.614	0.621	0.760	$+0.089^*$	$+0.066$
GLM-4-9B	kept	40	0.624	0.514	0.504	0.608	0.650	$+0.011$	-0.094
Phi-4	kept	40	0.685	0.616	0.463	0.641	0.697	$+0.156^*$	-0.025

B.2 Incremental readout: states plus recorded cues

The original stand-alone probe comparison does not test whether hidden states add information to the six features. Our new diagnostic uses the saved value-span states and summary-survival labels, without another model forward pass. Five outer folds hold out whole episodes; three inner folds of training episodes select hyperparameters. This is a post-hoc grouped-CV analysis, not a replacement for the original LOEO probe table.

The joint linear kernel is $K_F + \eta K_H$, where F contains the six measured cues. Training-fold means and standard deviations normalize F; training-fold centering and mean diagonal normalize K_H. We consider five of the nine archived layer positions, indices $0, 2, 4, 6, 8$, $\eta \in \{0, 0.1, 1\}$, and ridge penalties $\{10^{-2}, 10^{-1}, 1, 10, 100\}$. The $\eta = 0$ candidate is the matched six-feature ridge baseline. Inner-fold AUROC selects parameters; outer labels never enter normalization, selection, or fitting. We also fit the six-feature logistic model on the same outer folds.

Table 9: Post-hoc incremental prediction of summary survival, from 1,800 saved states per model. Differences use common outer-test items and 95% episode intervals, conditional on the fitted OOF scores. p_H adjusts the six joint-minus-baseline contrasts.

model	logistic	ridge	joint	joint − logistic [CI]	p_H	joint − ridge [CI]
Qwen3-14B	0.760	0.760	0.830	$+0.070$ [+0.053, +0.087]	0.003	$+0.070$ [+0.054, +0.087]
GLM-4-9B	0.650	0.650	0.653	$+0.003$ [-0.007, +0.013]	1.000	$+0.003$ [-0.008, +0.013]
Phi-4	0.693	0.692	0.723	$+0.029$ [+0.018, +0.040]	0.003	$+0.030$ [+0.019, +0.042]

The Qwen and Phi-4 increments survive this multiplicity sensitivity against both baselines; GLM's does not. Therefore a general claim of no incremental hidden information is not supported for the tested baselines. This is an incremental predictive-performance test, not a conditional-independence test against every possible function of the six cues. These readouts do not establish that the model itself can report the information, that it is specific to self-knowledge rather than another representation of the text, or that the joint signal improves a protection policy. New interfaces or interventions would require separate model experiments. Per-item OOF scores and selected parameters are saved; the numerical supplement replays their statistics but does not redistribute the tensors.

B.3 CROSS-MODEL TRANSFER AND CUE ABLATIONS

A six-cue logistic fit is trained on one model's outcomes and evaluated on another's. All cells here use summary-survival labels. Table 10 resamples target episodes, conditional on the source-trained fit; it does not include source-training uncertainty. The common-valid column compares transferred and self-report predictions on identical destination items. The underlying controlled texts can be shared across models, so this is cross-model prediction, not independent-domain replication. The descriptive transfer matrix and its own-LOEO reference remain in the numerical supplement; the reference fit is not an upper bound.

Table 10: Off-diagonal transfer uncertainty, using summary-survival labels throughout. p_H adjusts 30 common-valid differences.

source	target	full-target AUROC [CI]	valid n	transfer − self [CI]	p_H
Qwen3-14B	GLM-4-9B	0.627 [0.599, 0.654]	1475	+0.116 [+0.078, +0.156]	0.015
Qwen3-14B	Phi-4	0.675 [0.647, 0.703]	1630	+0.210 [+0.170, +0.251]	0.015
Qwen3-14B	Qwen3.7-Max	0.635 [0.590, 0.684]	516	+0.100 [+0.044, +0.161]	0.015
Qwen3-14B	GLM-5.1	0.502 [0.472, 0.537]	540	+0.015 [-0.017, +0.048]	1.000
Qwen3-14B	Doubao-2	0.542 [0.508, 0.587]	516	+0.045 [+0.005, +0.098]	0.273
GLM-4-9B	Qwen3-14B	0.732 [0.709, 0.755]	720	+0.086 [+0.024, +0.145]	0.068
GLM-4-9B	Phi-4	0.701 [0.674, 0.729]	1630	+0.235 [+0.194, +0.278]	0.015
GLM-4-9B	Qwen3.7-Max	0.643 [0.591, 0.697]	516	+0.105 [+0.046, +0.168]	0.015
GLM-4-9B	GLM-5.1	0.507 [0.469, 0.550]	540	+0.020 [-0.020, +0.063]	1.000
GLM-4-9B	Doubao-2	0.554 [0.516, 0.600]	516	+0.051 [+0.007, +0.112]	0.245
Phi-4	Qwen3-14B	0.740 [0.718, 0.763]	720	+0.101 [+0.040, +0.161]	0.020
Phi-4	GLM-4-9B	0.657 [0.628, 0.686]	1475	+0.141 [+0.101, +0.183]	0.015
Phi-4	Qwen3.7-Max	0.636 [0.586, 0.692]	516	+0.100 [+0.040, +0.163]	0.020
Phi-4	GLM-5.1	0.496 [0.461, 0.535]	540	+0.009 [-0.026, +0.047]	1.000
Phi-4	Doubao-2	0.548 [0.515, 0.591]	516	+0.046 [+0.003, +0.105]	0.275
Qwen3.7-Max	Qwen3-14B	0.744 [0.721, 0.768]	720	+0.104 [+0.039, +0.164]	0.054
Qwen3.7-Max	GLM-4-9B	0.650 [0.620, 0.679]	1475	+0.138 [+0.094, +0.184]	0.015
Qwen3.7-Max	Phi-4	0.684 [0.656, 0.712]	1630	+0.219 [+0.177, +0.263]	0.015
Qwen3.7-Max	GLM-5.1	0.530 [0.496, 0.571]	540	+0.043 [+0.006, +0.084]	0.275
Qwen3.7-Max	Doubao-2	0.548 [0.515, 0.591]	516	+0.051 [+0.009, +0.107]	0.224
GLM-5.1	Qwen3-14B	0.585 [0.558, 0.611]	720	−0.017 [-0.073, +0.039]	1.000
GLM-5.1	GLM-4-9B	0.541 [0.516, 0.568]	1475	+0.043 [+0.006, +0.079]	0.273
GLM-5.1	Phi-4	0.509 [0.484, 0.535]	1630	+0.046 [+0.010, +0.081]	0.224
GLM-5.1	Qwen3.7-Max	0.578 [0.542, 0.617]	516	+0.045 [-0.024, +0.112]	1.000
GLM-5.1	Doubao-2	0.523 [0.491, 0.559]	516	+0.027 [-0.014, +0.075]	1.000
Doubao-2	Qwen3-14B	0.677 [0.651, 0.701]	720	+0.040 [-0.020, +0.102]	1.000
Doubao-2	GLM-4-9B	0.643 [0.611, 0.677]	1475	+0.128 [+0.083, +0.174]	0.015
Doubao-2	Phi-4	0.659 [0.632, 0.688]	1630	+0.194 [+0.153, +0.237]	0.015
Doubao-2	Qwen3.7-Max	0.613 [0.559, 0.668]	516	+0.076 [+0.010, +0.150]	0.275
Doubao-2	GLM-5.1	0.524 [0.484, 0.572]	540	+0.037 [-0.008, +0.088]	0.892

Does one shared cue explain transfer? We additionally refit after removing each of the six cues, restricted to the three fact-limit, summary-survival settings. Models fit only source-model labels with the original regularization and no target tuning. Each ablated-versus-full contrast uses the same target facts and resamples target episodes, conditional on the source fit. Table 11 reports all six directions and six omissions, with 36-test Holm sensitivity; full intervals are in the numerical supplement. No removed cue reliably reduces AUROC in every direction. This does not establish an absence of shared structure: correlated cues may substitute for each other, and the models were evaluated on shared controlled materials, not independent new domains.

Table 11: Leave-one-cue-out transfer: AUROC change from the full six-cue transferred fit. A negative change means removal hurts. * marks exploratory $p_H < 0.05$ across 36 tests. "First" and "last" refer to mention position; "length" is value length.

source	target	full	no day	no first	no last	no repeats	no interference	no length
Qwen3-14B	GLM-4-9B	0.627	+0.007	−0.001*	+0.001	+0.006	−0.001	−0.002
Qwen3-14B	Phi-4	0.675	+0.004	−0.001	−0.000	+0.008	−0.001	−0.021*
GLM-4-9B	Qwen3-14B	0.732	+0.001*	+0.006	−0.014*	−0.001	−0.002	−0.006
GLM-4-9B	Phi-4	0.701	−0.000	−0.006	−0.005	−0.001	+0.000	−0.007*
Phi-4	Qwen3-14B	0.740	−0.002*	+0.006*	−0.007	−0.005	−0.001	−0.003
Phi-4	GLM-4-9B	0.657	+0.000	−0.004	−0.005	+0.001	−0.001	−0.002

B.4 SECOND-DOMAIN ROBUSTNESS

The same model, two lexically disjoint generators. "cross" is fitted on the other domain and evaluated here, so it checks sensitivity to vocabulary within the controlled generator family, not transfer to independent naturalistic dialogue.

Table 12: Cross-material transfer. Point estimates under two generated lexical domains. Neither a near-chance point estimate nor a transfer point estimate establishes equivalence or generalisation to naturalistic dialogue.

backbone	materials	n	base	own LOEO	cross	self-report
Qwen3-14B	wet-lab	1200	0.50	0.843	0.695	0.556
Qwen3-14B	devops	1800	0.49	0.760	0.665	0.614
Phi-4	wet-lab	1200	0.53	0.689	0.698	0.458
Phi-4	devops	1798	0.52	0.696	0.702	0.462
GLM-4-9B	wet-lab	1192	0.52	0.693	0.701	0.508
GLM-4-9B	devops	1786	0.59	0.650	0.660	0.502

B.5 FIXED-BUDGET RANKING QUALITY

We spend a fixed budget on the three items assigned the lowest predicted survival. All methods are evaluated on the same strict-ranking-valid episodes. The nested one-feature and six-feature probabilities are genuinely out of fold: by episode for controlled work logs and by source conversation for LoCoMo. A cutoff tie is assigned its expectation under uniform random tie-breaking rather than broken by item order. Loss-capture@3 excludes an episode only when it has no lost item; precision@3 retains every episode. LoCoMo intervals resample source conversations, while work-log intervals resample episodes.

Table 13: Fixed-budget decision quality. "eps." gives valid-ranking/loss-positive episodes. Each metric cell is point estimate / paired difference from random [95% CI]; * marks an interval excluding zero. Random loss capture varies on LoCoMo because window sizes vary.

backbone	material	eps.	selector	loss-capture@3 / Δ random	precision@3 / Δ random
Qwen3-14B	work logs	60/60	self-report	0.330 / +0.080 [+0.044, +0.114]*	0.661 / +0.158 [+0.086, +0.228]*
			nested 1-feature	0.328 / +0.078 [+0.063, +0.091]*	0.658 / +0.156 [+0.126, +0.183]*
			six-feature	0.338 / +0.088 [+0.054, +0.122]*	0.678 / +0.175 [+0.108, +0.242]*
			random	0.250 / +0.000 [+0.000, +0.000]	0.503 / +0.000 [+0.000, +0.000]
Qwen3-14B	LoCoMo	24/24	self-report	0.526 / −0.026 [-0.069, +0.015]	0.736 / −0.023 [-0.073, +0.026]
			nested 1-feature	0.532 / −0.020 [-0.070, +0.040]	0.750 / −0.009 [-0.061, +0.054]
			six-feature	0.574 / +0.022 [-0.020, +0.069]	0.792 / +0.032 [-0.010, +0.087]
			random	0.552 / +0.000 [+0.000, +0.000]	0.759 / +0.000 [+0.000, +0.000]
GLM-4-9B	work logs	124/107	self-report	0.251 / +0.001 [-0.030, +0.032]	0.425 / +0.019 [-0.017, +0.056]
			nested 1-feature	0.388 / +0.138 [+0.094, +0.182]*	0.578 / +0.173 [+0.127, +0.220]*
			six-feature	0.386 / +0.136 [+0.092, +0.181]*	0.575 / +0.170 [+0.124, +0.217]*
			random	0.250 / +0.000 [+0.000, +0.000]	0.405 / +0.000 [+0.000, +0.000]
GLM-4-9B	LoCoMo	13/13	self-report	0.599 / +0.011 [-0.097, +0.143]	0.615 / +0.015 [-0.089, +0.154]
			nested 1-feature	0.751 / +0.163 [+0.093, +0.232]*	0.756 / +0.156 [+0.076, +0.226]*
			six-feature	0.738 / +0.151 [+0.081, +0.230]*	0.744 / +0.143 [+0.065, +0.237]*
			random	0.588 / +0.000 [+0.000, +0.000]	0.600 / +0.000 [+0.000, +0.000]
Phi-4	work logs	136/121	self-report	0.225 / −0.025 [-0.054, +0.006]	0.444 / −0.040 [-0.082, -0.000]
			nested 1-feature	0.367 / +0.117 [+0.077, +0.157]*	0.650 / +0.165 [+0.124, +0.205]*
			six-feature	0.393 / +0.143 [+0.107, +0.181]*	0.706 / +0.222 [+0.186, +0.256]*
			random	0.250 / +0.000 [+0.000, +0.000]	0.484 / +0.000 [+0.000, +0.000]
Phi-4	LoCoMo	24/24	self-report	0.544 / −0.008 [-0.054, +0.034]	0.764 / −0.027 [-0.091, +0.024]
			nested 1-feature	0.589 / +0.037 [-0.001, +0.068]	0.847 / +0.056 [+0.005, +0.104]*
			six-feature	0.546 / −0.006 [-0.038, +0.028]	0.778 / −0.013 [-0.051, +0.022]
			random	0.552 / +0.000 [+0.000, +0.000]	0.791 / +0.000 [+0.000, +0.000]

The work-log result separates the model families rather than manufacturing a uniform claim. Qwen3-14B's valid self-rankings, nested single feature, and six-feature fit all capture about 0.33 of losses versus 0.25 at random. GLM-4-9B self-report is random while the external methods capture 0.39; Phi-4 self-report is no better than random while the external methods reach 0.37–0.39. LoCoMo

is lower powered: the two external selectors separate on GLM-4-9B, Phi-4 has a precision-only one-feature effect, and the remaining intervals cover zero. Among the added frontier LoCoMo cells, DeepSeek-V4-Pro is the exception: its self-report improves loss-capture@3 over random by $+0.123$ [0.031, 0.201] and precision@3 by $+0.103$ [0.001, 0.188], both source-clustered. Yet when its ranking is actually spent at two flags, self-report recall is 0.294, exactly tied with random and below surface at 0.350. Thus even this decision-quality signal does not establish downstream policy benefit.

C Protection and memory controls

C.1 Controlled selectors and valid rankings

The full-sample controlled comparison evaluates all important facts, irrespective of whether the original self-ranking was complete. Incomplete rankings use seeded random fallback for the self-selected policy. The valid-only analysis removes whole episodes with incomplete rankings and is a sensitivity check, not a replacement estimand: completion can select an easier subset. Both analyses retain GLM's positive self-minus-random result.

Table 14: Controlled self-minus-random protection on important-fact survival. Full and valid-ranking-only cohorts are distinct. p_H is the within-family Holm-adjusted bootstrap-tail sensitivity value; nominal 95% episode intervals are shown.

model	full difference [CI]	p_H	valid logs	valid-only difference [CI]	p_H
Qwen3-14B	$+0.003$ [-0.020, +0.026]	1.000	60	$+0.011$ [-0.028, +0.050]	1.000
GLM-4-9B	$+0.052$ [+0.019, +0.084]	0.007	124	$+0.047$ [+0.012, +0.083]	0.033
Phi-4	$+0.009$ [-0.030, +0.049]	1.000	136	$+0.005$ [-0.037, +0.048]	1.000

Table 15: Original external protection contrasts on the same important-fact cohort.

model	external − random [CI]	external − mention [CI]
Qwen3-14B	$+0.052$ [+0.031, +0.073]*	$+0.006$ [-0.009, +0.021]
GLM-4-9B	$+0.083$ [+0.052, +0.114]*	$+0.022$ [-0.009, +0.057]
Phi-4	$+0.049$ [+0.007, +0.091]*	$+0.021$ [-0.021, +0.066]

The hindsight selector uses the untreated run's observed omissions. Its uncertain paired comparison to the external rule does not establish a policy ceiling or disprove one. Descriptive gain ratios are retained with the numerical records, not interpreted as evidence of superiority.

C.2 Memory implementation and measured costs

Each store contains 16 tagged sessions. Summary bullets or sentences form retrievable entries, whereas raw-log entries are user–assistant turns. A character-bigram BM25 retriever uses $k_1 = 1.5$ and $b = 0.75$, with the question's project anchor identifying the relevant session. A fixed read-character budget applies to every write policy. Value collisions within a block are excluded from the important-fact scoring set, but the excluded questions are not removed from the writing process.

Extractive and recent-turn baselines receive the measured write budget of unprotected model compaction at each level. The extractive baseline detects turns containing digits, including decoys and numeric filler; it does not match generator-specific fact-introduction phrases or receive the correct value list. This is a measured-cost baseline comparison. Flagging policies instead report the characters they actually wrote, so their differences from no flags need not be equal-byte effects.

C.3 Same-item effects, redundancy, and grouping

Let I be the fixed important, value-unique facts and $U_a = I \setminus F_a$ the facts not flagged by policy a. The same-item unflagged effect compares that same U_a under policy a and no flags, whereas an

Table 16: End-to-end important-fact answer recall at the middle requested memory limit; actual stored characters per session are in parentheses. Extractive baselines receive the measured storage of unprotected model compaction; protection policies report their actual write sizes.

write policy	Qwen3-14B	GLM-4-9B	Phi-4
write-all, truncate overflow	0.000 (45)	0.027 (81)	0.237 (349)
keep the most recent turns	0.140 (132)	0.174 (146)	0.227 (327)
keep value-bearing turns	0.164 (132)	0.184 (145)	0.291 (330)
model's own compaction	0.405 (144)	0.458 (165)	0.699 (346)
+ random protection	0.612 (117)	0.595 (172)	0.759 (389)
+ **model-selected** protection	0.619 (113)	0.569 (164)	0.786 (399)
+ recency protection	0.632 (114)	0.625 (175)	0.753 (397)
+ mention-count protection	0.666 (114)	0.649 (178)	0.833 (423)
+ **external** protection	**0.676 (114)**	**0.659 (184)**	**0.856 (417)**
keep the raw log, retrieve	0.987 (1762)	0.963 (1762)	0.997 (1762)

Table 17: Compaction versus extractive baselines on the same important, value-unique facts. Differences pool the three write levels, with episode-clustered 95% intervals; each baseline is matched to unprotected model compaction's measured storage at that level.

comparison	Qwen3-14B	GLM-4-9B	Phi-4
model compaction − extractive	$+0.159$ [+0.120, +0.197]	$+0.174$ [+0.123, +0.224]	$+0.281$ [+0.191, +0.373]
model compaction − recency	$+0.172$ [+0.131, +0.213]	$+0.182$ [+0.133, +0.232]	$+0.343$ [+0.256, +0.426]

own-unflagged contrast uses all of I in the comparator:

$$\Delta_a^U = \bar{y}_{U_a}^a - \bar{y}_{U_a}^0, \qquad \Delta_a^{\text{own}} = \Delta_a^U + \underbrace{(\bar{y}_{U_a}^0 - \bar{y}_I^0)}_{\text{composition}}.$$

The whole-cohort difference is the size-weighted mean of the flagged and unflagged same-item differences. The code checks both identities and exact episode/item, importance, uniqueness, feature, and block alignment. Nothing is selected by its observed success. For direct selector comparisons we use either all of I or the intersection $I \setminus (F_a \cup F_b)$, not the two policies' different unflagged sets.

Self flags have greater overlap with the untreated kept set than external flags, but none of the six self-minus-random overlap intervals excludes zero. Thus the proposed explanation that self protection is uniquely redundant is not established. Requiring an already-kept fact does not mechanically consume an additional slot; actual gains and losses must be measured. The stored transitions show that redistribution costs occur under every tested flagging rule. This is not a causal localization of an instruction-following failure.

Mean write-size comparisons do not by themselves identify a length mechanism. GLM's L1 self arm is nearly byte matched and Phi-4's is larger, yet both incur same-item unflagged loss. We therefore reject only a uniform explanation in terms of smaller average output, not every possible interaction between length, output format, and selection. Scores from the anchored source run are reused across memory limits; they are not newly elicited at each budget. L0 is the tightest nominal limit, while L2 is the widest.

Earlier versions compared a treated policy's unflagged set with a different no-flag or selector set. Such own-cohort differences contain a composition term and are not same-item treatment effects. We report the final fixed-cohort analyses here; all original tables remain in the numerical supplement, with their status explained in `NUMERICAL_AUDIT.md`. The old valid-only memory comparison also used different cohorts and is not evidence that all parsing confounds have been excluded.

C.4 Retrieval and read-budget robustness

With four episode summaries in one context and no per-entry tags, $P(\text{recovered} \mid \text{kept})$ falls from 0.965 and 0.843 to 0.872 and 0.784 for two backbones; Phi-4 remains at 0.925. Among in-summary failures, 24–42% answer with another project's value for the same key. This exposes a binding failure, not simply a failure to retain the value. Recovery when the value was absent stays at 0.005–0.017.

Table 18: Per-budget matched changes versus no flags. "Composition" is the additive term above. Positive total effects can coexist with negative same-item unflagged effects.

model	budget	policy	write ratio	all I change [CI]	same U_a change [CI]	composition
Qwen3-14B	L1	self	0.788	$+0.214$ [+0.159, +0.272]	-0.200 [-0.273, -0.123]	+0.083
Qwen3-14B	L1	external	0.795	$+0.271$ [+0.207, +0.332]	-0.320 [-0.403, -0.240]	+0.255
Qwen3-14B	L1	random	0.811	$+0.207$ [+0.154, +0.264]	-0.199 [-0.275, -0.120]	+0.026
Qwen3-14B	L1	mention	0.793	$+0.261$ [+0.202, +0.320]	-0.255 [-0.331, -0.176]	+0.179
Qwen3-14B	L2	self	0.780	$+0.087$ [+0.043, +0.130]	-0.119 [-0.199, -0.043]	+0.023
Qwen3-14B	L2	external	0.775	$+0.164$ [+0.111, +0.218]	-0.143 [-0.219, -0.067]	+0.161
Qwen3-14B	L2	random	0.802	$+0.117$ [+0.070, +0.164]	-0.026 [-0.104, +0.053]	-0.040
Qwen3-14B	L2	mention	0.775	$+0.181$ [+0.134, +0.227]	-0.067 [-0.148, +0.013]	+0.123
GLM-4-9B	L1	self	0.994	$+0.110$ [+0.037, +0.187]	-0.248 [-0.344, -0.151]	+0.051
GLM-4-9B	L1	external	1.115	$+0.201$ [+0.127, +0.270]	-0.148 [-0.255, -0.038]	+0.112
GLM-4-9B	L1	random	1.039	$+0.137$ [+0.069, +0.203]	-0.185 [-0.285, -0.088]	+0.012
GLM-4-9B	L1	mention	1.078	$+0.191$ [+0.109, +0.269]	-0.107 [-0.213, +0.000]	+0.052
GLM-4-9B	L2	self	0.951	$+0.064$ [+0.000, +0.127]	-0.102 [-0.182, -0.013]	-0.030
GLM-4-9B	L2	external	1.001	$+0.114$ [+0.049, +0.184]	-0.134 [-0.221, -0.050]	+0.116
GLM-4-9B	L2	random	1.000	$+0.084$ [+0.017, +0.150]	-0.119 [-0.209, -0.025]	+0.014
GLM-4-9B	L2	mention	1.007	$+0.114$ [+0.051, +0.176]	-0.094 [-0.182, +0.000]	+0.042
Phi-4	L1	self	1.153	$+0.087$ [+0.007, +0.168]	-0.156 [-0.280, -0.038]	-0.010
Phi-4	L1	external	1.207	$+0.157$ [+0.073, +0.239]	-0.152 [-0.254, -0.052]	+0.156
Phi-4	L1	random	1.124	$+0.060$ [-0.019, +0.140]	-0.159 [-0.269, -0.047]	-0.010
Phi-4	L1	mention	1.222	$+0.134$ [+0.047, +0.226]	-0.107 [-0.229, +0.019]	+0.080
Phi-4	L2	self	1.178	$+0.007$ [-0.061, +0.076]	-0.156 [-0.274, -0.035]	-0.038
Phi-4	L2	external	1.284	$+0.100$ [+0.043, +0.160]	-0.055 [-0.121, +0.007]	+0.088
Phi-4	L2	random	1.251	-0.013 [-0.082, +0.057]	-0.192 [-0.297, -0.084]	-0.002
Phi-4	L2	mention	1.257	$+0.054$ [-0.003, +0.114]	-0.094 [-0.178, -0.013]	+0.030

Table 19: Fraction of each flag set already kept without flags. Overlap uses all flagged facts; self-minus-random contrasts use paired episodes.

model	budget	self	random	external	mention	self − random [CI]	p_H
Qwen3-14B	L1	0.463	0.512	0.267	0.342	-0.050 [-0.112, +0.017]	0.708
Qwen3-14B	L2	0.733	0.783	0.604	0.629	-0.050 [-0.112, +0.008]	0.708
GLM-4-9B	L1	0.512	0.525	0.408	0.500	-0.013 [-0.071, +0.046]	1.000
GLM-4-9B	L2	0.721	0.713	0.579	0.675	$+0.008$ [-0.046, +0.062]	1.000
Phi-4	L1	0.758	0.754	0.613	0.683	$+0.004$ [-0.046, +0.050]	1.000
Phi-4	L2	0.875	0.846	0.742	0.812	$+0.029$ [-0.012, +0.071]	0.832

The tagged memory loop alleviates this retrieval ambiguity. A fourfold tighter read budget, 300 rather than 1,200 characters, was tested on two of the three backbones; no cell changes by more than 0.04, and the direction is upward. The third model has no such replication. These are controls within a lexical, tagged store, not a test of arbitrary large-scale retrievers or an implementation of a named agent-memory system.

Table 20: Selector comparisons on common cohorts. The intersection contains only facts flagged by neither policy; its size is not generally the size of either own-unflagged set.

model	budget	contrast	all I change [CI]	p_H	intersection n	intersection change [CI]	p_H
Qwen3-14B	L1	external − self	+0.057 [+0.000, +0.113]	0.564	94	−0.053 [-0.139, +0.035]	1.000
Qwen3-14B	L1	external − random	+0.064 [+0.003, +0.125]	0.564	73	−0.055 [-0.141, +0.028]	1.000
Qwen3-14B	L1	external − mention	+0.010 [-0.034, +0.052]	1.000	113	−0.035 [-0.109, +0.038]	1.000
Qwen3-14B	L2	external − self	+0.077 [+0.034, +0.118]	0.017	94	+0.000 [-0.090, +0.088]	1.000
Qwen3-14B	L2	external − random	+0.047 [+0.000, +0.095]	0.564	73	−0.055 [-0.164, +0.051]	1.000
Qwen3-14B	L2	external − mention	−0.017 [-0.062, +0.027]	1.000	113	−0.027 [-0.114, +0.056]	1.000
GLM-4-9B	L1	external − self	+0.090 [+0.019, +0.160]	0.203	75	+0.173 [+0.041, +0.306]	0.178
GLM-4-9B	L1	external − random	+0.064 [+0.000, +0.129]	0.564	72	+0.125 [-0.014, +0.254]	1.000
GLM-4-9B	L1	external − mention	+0.010 [-0.064, +0.088]	1.000	90	+0.122 [-0.011, +0.258]	1.000
GLM-4-9B	L2	external − self	+0.050 [-0.010, +0.109]	0.780	75	+0.080 [-0.037, +0.195]	1.000
GLM-4-9B	L2	external − random	+0.030 [-0.027, +0.087]	1.000	72	+0.042 [-0.111, +0.186]	1.000
GLM-4-9B	L2	external − mention	+0.000 [-0.057, +0.056]	1.000	90	+0.022 [-0.089, +0.136]	1.000
Phi-4	L1	external − self	+0.070 [+0.010, +0.129]	0.364	59	−0.051 [-0.186, +0.097]	1.000
Phi-4	L1	external − random	+0.097 [+0.037, +0.159]	0.037	71	+0.042 [-0.092, +0.176]	1.000
Phi-4	L1	external − mention	+0.023 [-0.034, +0.079]	1.000	99	+0.040 [-0.076, +0.149]	1.000
Phi-4	L2	external − self	+0.094 [+0.040, +0.147]	0.017	59	+0.220 [+0.103, +0.350]	0.009
Phi-4	L2	external − random	+0.114 [+0.060, +0.166]	0.009	71	+0.141 [+0.014, +0.275]	0.640
Phi-4	L2	external − mention	+0.047 [+0.000, +0.092]	0.564	99	+0.051 [-0.038, +0.137]	1.000

Table 21: Pooled L1/L2 same-item unflagged effects: episode intervals and shared-store-block sensitivity. The five-block sign-flip check assumes symmetric independent block contrasts; it is not based on randomized policy assignment, and its minimum two-sided p is 0.0625.

model	policy	episode difference [CI]	leave-one-block range	sign-flip p
Qwen3-14B	self	−0.159 [-0.213, -0.104]	[-0.177, -0.145]	0.062
Qwen3-14B	external	−0.231 [-0.291, -0.173]	[-0.239, -0.222]	0.062
Qwen3-14B	random	−0.113 [-0.169, -0.053]	[-0.140, -0.085]	0.125
Qwen3-14B	mention	−0.161 [-0.222, -0.101]	[-0.177, -0.147]	0.062
GLM-4-9B	self	−0.175 [-0.242, -0.105]	[-0.184, -0.163]	0.062
GLM-4-9B	external	−0.141 [-0.222, -0.064]	[-0.197, -0.120]	0.125
GLM-4-9B	random	−0.152 [-0.228, -0.080]	[-0.168, -0.124]	0.062
GLM-4-9B	mention	−0.101 [-0.177, -0.025]	[-0.134, -0.071]	0.125
Phi-4	self	−0.156 [-0.248, -0.057]	[-0.173, -0.112]	0.062
Phi-4	external	−0.103 [-0.169, -0.038]	[-0.122, -0.072]	0.062
Phi-4	random	−0.175 [-0.259, -0.090]	[-0.201, -0.156]	0.062
Phi-4	mention	−0.101 [-0.185, -0.019]	[-0.125, -0.054]	0.125

D LoCoMo: external boundary and human validation

D.1 Source grouping and common-valid predictions

The 24 windows are nested in nine source conversations, not 24 independent sources. GLM-4-9B has valid rankings in 13 windows from eight sources. The main table states the actual valid facts/windows/sources for each model. Fits hold out source conversations; Table 22 then scores self-report and both cue predictors on the same valid facts. Source deletion removes a whole conversation and refits the cue models. Deletion ranges are not confidence intervals and cannot create external validity from nine sources.

Table 22: Common-valid LoCoMo contrasts and leave-one-source-out point ranges. Counts are valid facts/windows/sources. All intervals resample sources; deletion ranges are not CIs.

model	valid n/w/s	1 cue − self [CI]	6 cues − self [CI]	deletion range, 6 cues − self
Qwen3-14B	143/24/9	+0.000 [-0.142, +0.166]	+0.055 [-0.052, +0.224]	[+0.016, +0.105]
GLM-4-9B	75/13/8	+0.169 [-0.039, +0.387]	+0.136 [-0.084, +0.352]	[+0.062, +0.236]
Phi-4	143/24/9	+0.068 [-0.057, +0.249]	−0.098 [-0.176, +0.045]	[-0.227, -0.020]
DeepSeek-V4-Pro	135/23/9	−0.006 [-0.130, +0.153]	+0.005 [-0.137, +0.174]	[-0.070, +0.035]
Qwen3.7-Max	143/24/9	−0.078 [-0.271, +0.053]	−0.036 [-0.227, +0.127]	[-0.098, +0.044]
GLM-5.1	143/24/9	−0.169 [-0.292, -0.036]	−0.185 [-0.310, -0.048]	[-0.217, -0.096]
Doubao-Seed-2	143/24/9	−0.149 [-0.268, -0.030]	−0.052 [-0.158, +0.065]	[-0.171, -0.005]

D.2 PROTECTION, ACTUAL COSTS, AND UNPROTECTED FACTS

The intervention retains the same 24 windows and 143 target facts for every arm. Qwen3-14B, GLM-4-9B, and DeepSeek-V4-Pro each have seven arms and two flag budgets. Every target fact is queried by the dataset, so all-fact recall is the primary policy outcome; the two-flag level is primary and three flags supply budget sensitivity. Only DeepSeek completes both measurement and intervention among the frontier models. All comparisons are exploratory because the unprotected measurement gate failed and the requested fact limit does not bind actual writing.

Table 23: LoCoMo end-to-end recall under no flags (N), random flags (R), self-report flags (S), and six-feature external flags (E). Differences use paired source-conversation bootstrap intervals; * excludes zero.

backbone	flags	N	R	S	E	E−N [95% CI]	S−N [95% CI]	E−R [95% CI]
Qwen3-14B	2	0.063	0.266	0.245	0.252	$+0.189$ [+0.123, +0.267]*	$+0.182$ [+0.074, +0.302]*	-0.014 [-0.099, +0.082]
Qwen3-14B	3	0.063	0.336	0.315	0.280	$+0.217$ [+0.117, +0.329]*	$+0.252$ [+0.136, +0.391]*	-0.056 [-0.174, +0.064]
GLM-4-9B	2	0.175	0.343	0.315	0.294	$+0.119$ [+0.015, +0.224]*	$+0.140$ [+0.038, +0.266]*	-0.049 [-0.140, +0.035]
GLM-4-9B	3	0.175	0.322	0.364	0.336	$+0.161$ [+0.066, +0.284]*	$+0.189$ [+0.088, +0.331]*	$+0.014$ [-0.070, +0.085]
DeepSeek-V4-Pro	2	0.161	0.294	0.294	0.350	$+0.189$ [+0.134, +0.250]*	$+0.133$ [+0.082, +0.198]*	$+0.056$ [+0.000, +0.124]
DeepSeek-V4-Pro	3	0.161	0.448	0.434	0.434	$+0.273$ [+0.189, +0.364]*	$+0.273$ [+0.161, +0.391]*	-0.014 [-0.106, +0.053]

External, self, and random protection improve all-fact recall over no flags in all six model/budget cells. External-minus-random, self, or mention intervals touch or cross zero in all 18 selector contrasts. These are not generally equal-byte effects: the displayed flagging arms write 0.47–1.34 times the no-flag length. DeepSeek's self/random arms are near matched, at 0.96–0.98, while its external arm writes more. The result therefore supports a preservation-prompt benefit, not a reliably best selector at a fixed cost.

The same-item unprotected analysis joins exact episode/item IDs, verifies source IDs and features, and evaluates both the treated arm and no flags on the treated arm's same unflagged set. All 18 model/budget/policy cells are shown below. Whole-cohort effects are unchanged by this correction because that cohort was already identical across arms. For self-selected protection, Qwen at three flags has $p_H = 0.018$ and DeepSeek at two flags $p_H = 0.104$; both source-sign-flip sensitivities are 0.0625. The sign-flip check assumes independent symmetric source contributions, not randomized assignment. Neither it nor the bootstrap guarantees nominal coverage with so few sources. Random and external protection can also incur losses on unprotected facts.

Table 24: Same-item unprotected recall changes versus no flags on LoCoMo. Each pair uses identical n facts from nine sources. p_H adjusts all 18 tests; p_{SF} is the source-sign-flip sensitivity. Historical different-cohort values remain in the numerical audit, not in this table.

model	flags	policy	n	same-item difference [95% CI]	p_H	p_{SF}
Qwen3-14B	2	self	95	$+0.000$ [-0.067, +0.071]	1.000	1.000
Qwen3-14B	2	random	95	$+0.000$ [-0.066, +0.075]	1.000	1.000
Qwen3-14B	2	external	95	$+0.000$ [-0.043, +0.039]	1.000	1.000
Qwen3-14B	3	self	71	-0.099 [-0.157, -0.043]	0.018	0.062
Qwen3-14B	3	random	71	-0.042 [-0.098, +0.025]	1.000	0.500
Qwen3-14B	3	external	71	-0.099 [-0.160, -0.034]	0.018	0.062
GLM-4-9B	2	self	95	$+0.000$ [-0.114, +0.152]	1.000	1.000
GLM-4-9B	2	random	95	$+0.011$ [-0.091, +0.126]	1.000	1.000
GLM-4-9B	2	external	95	-0.063 [-0.207, +0.055]	1.000	0.500
GLM-4-9B	3	self	71	-0.056 [-0.143, +0.034]	1.000	0.406
GLM-4-9B	3	random	71	-0.085 [-0.171, +0.000]	0.728	0.156
GLM-4-9B	3	external	71	-0.141 [-0.258, -0.038]	0.180	0.078
DeepSeek-V4-Pro	2	self	95	-0.105 [-0.173, -0.037]	0.104	0.062
DeepSeek-V4-Pro	2	random	95	-0.063 [-0.129, +0.000]	0.916	0.188
DeepSeek-V4-Pro	2	external	95	-0.011 [-0.073, +0.068]	1.000	1.000
DeepSeek-V4-Pro	3	self	71	-0.042 [-0.128, +0.094]	1.000	0.688
DeepSeek-V4-Pro	3	random	71	-0.028 [-0.109, +0.074]	1.000	0.766
DeepSeek-V4-Pro	3	external	71	-0.070 [-0.170, +0.056]	1.000	0.406

Hindsight-minus-random intervals cover zero in all six all-fact cells. Hindsight-minus-no-flags has positive point estimates of $+0.084$–$+0.301$, with five intervals excluding zero and one touching it. Retrospective loss-capture and intervention point-estimate winners also differ across the three models, but their uncertainty does not establish a population ordering reversal. The descriptive ranking table

Table 25: Corrected same-item LoCoMo self-minus-no-flags sensitivity after excluding incomplete rankings. Counts are unflagged facts/windows/sources; p_H adjusts six tests. Subset selection is not a replacement for the full-sample estimand.

model	flags	valid n/w/s	same-item difference [95% CI]	p_H	p_{SF}
Qwen3-14B	2	95/24/9	$+0.000$ [-0.067, +0.070]	1.000	1.000
Qwen3-14B	3	71/24/9	-0.099 [-0.155, -0.039]	0.012	0.062
GLM-4-9B	2	49/13/8	-0.041 [-0.083, +0.000]	0.832	0.500
GLM-4-9B	3	36/13/8	-0.056 [-0.133, +0.059]	1.000	0.625
DeepSeek-V4-Pro	2	89/23/9	-0.101 [-0.176, -0.029]	0.060	0.078
DeepSeek-V4-Pro	3	66/23/9	-0.030 [-0.128, +0.109]	1.000	0.875

and earlier mixed-cohort sensitivity outputs are retained in the numerical supplement, not presented as additional confirmations.

D.3 Human validation of answer grading

A mechanical audit of the three original LoCoMo answer cells finds that 78–90% of automatic zeroes are explicit refusals. Candidate substring false negatives and false positives are each 0–1.4%; strict and relaxed deterministic graders move self-report AUROC by at most 0.054.

We additionally conducted a blinded full-question audit of 420 archived answers, 60 from each of the seven LoCoMo model cells. The rating sheet showed only the complete question, reference answer, and model response. It hid model identity, source conversation, automatic labels, summary presence, and prediction scores. Every automatic success was included; automatic failures were sampled within each model after stratification by whether the reference answer appeared in the summary. Tables 26 and 27 use inverse sampling-probability weights to represent the full 1001-response population. Their intervals resample the nine source conversations as clusters, jointly across model cells. Multi-model AUROCs are equal-weight averages of within-model AUROCs, rather than pooling incomparable score scales.

Table 26: Full-question human audit of the LoCoMo answer grader. "sample/pop." gives audited and represented responses. Rates and agreement are inverse-probability weighted; FN is the strict-grader false-negative rate conditional on a human-correct answer.

group	sample/pop.	automatic	human	agreement [95% CI]	FN
Overall	420/1001	0.128	0.159	0.969 [0.952, 0.987]	0.195
Open-weight	180/429	0.126	0.157	0.969 [0.941, 1.000]	0.199
Frontier	240/572	0.129	0.160	0.969 [0.946, 0.991]	0.191
Qwen3-14B	60/143	0.063	0.077	0.985 [0.961, 1.000]	0.188
GLM-4-9B	60/143	0.175	0.226	0.949 [0.890, 1.000]	0.226
Phi-4	60/143	0.140	0.168	0.972 [0.911, 1.000]	0.169
DeepSeek-V4-Pro	60/143	0.161	0.190	0.971 [0.912, 1.000]	0.153
Qwen3.7-Max	60/143	0.077	0.105	0.972 [0.911, 1.000]	0.270
GLM-5.1	60/143	0.077	0.084	0.993 [0.976, 1.000]	0.083
Doubao-Seed-2	60/143	0.203	0.261	0.942 [0.850, 1.000]	0.222

The grader agrees with human judgment on 96.9% of the represented population. All 15 disagreements in the sample are false negatives, and none is a false positive. Estimated answer accuracy increases from 0.128 to 0.159. The 180 entries shared with the earlier shortened-question audit reproduce both correctness and semantic-paraphrase labels exactly (180/180); because the displayed questions differed, we treat this as label reproducibility rather than a clean inter-rater coefficient. Seven decisions contained ambiguity notes. A conservative sensitivity that sets all seven to incorrect does not change the qualitative model-cell conclusions.

The audit does not convert the LoCoMo result into a universal external-prediction gain. Only GLM-4-9B has positive cue-minus-self intervals for both external predictors. In the open-weight macro average both cue differences include zero; in the frontier macro average the one-cue difference includes zero and the six-cue difference is negative. Human judgment therefore confirms that exact matching is conservative while preserving—and, for the six-cue frontier comparison, strengthening—the external-validity boundary. The audit checks archived final-answer correctness only: it cannot

replace the full-population `in_summary` labels because complete summaries were not stored per item.

Table 27: Prediction AUROC after replacing the automatic answer label with audited human correctness. "rank/audit" reports responses with valid self-ranks. Group rows macro-average within-model AUROCs. Differences are cue minus self-report; intervals resample source conversations. These are secondary end-to-end labels, not replacements for the full-sample summary-survival analysis.

(a) AUROC under the audited answer label.

group	rank/audit	self-report [95% CI]	one cue [95% CI]	six cues [95% CI]
Overall	414/420	0.591 [0.555, 0.636]	0.587 [0.492, 0.649]	0.527 [0.493, 0.588]
Open-weight	174/180	0.522 [0.442, 0.622]	0.612 [0.536, 0.684]	0.595 [0.538, 0.678]
Frontier	240/240	0.642 [0.560, 0.699]	0.568 [0.432, 0.655]	0.477 [0.410, 0.572]
Qwen3-14B	60/60	0.594 [0.478, 0.866]	0.603 [0.441, 0.778]	0.605 [0.455, 0.815]
GLM-4-9B	54/60	0.480 [0.296, 0.658]	0.703 [0.617, 0.820]	0.713 [0.621, 0.814]
Phi-4	60/60	0.493 [0.283, 0.670]	0.529 [0.305, 0.658]	0.468 [0.246, 0.636]
DeepSeek-V4-Pro	60/60	0.661 [0.482, 0.802]	0.718 [0.531, 0.856]	0.682 [0.530, 0.823]
Qwen3.7-Max	60/60	0.732 [0.596, 0.858]	0.653 [0.247, 0.800]	0.448 [0.207, 0.751]
GLM-5.1	60/60	0.609 [0.487, 0.763]	0.466 [0.306, 0.697]	0.335 [0.220, 0.543]
Doubao-Seed-2	60/60	0.565 [0.377, 0.688]	0.436 [0.293, 0.588]	0.440 [0.283, 0.643]

(b) Direct AUROC differences against self-report.

group	one cue − self [95% CI]	six cues − self [95% CI]
Overall	-0.004 [-0.115, $+0.072$]	-0.063 [-0.125, $+0.007$]
Open-weight	$+0.090$ [-0.070, $+0.222$]	$+0.073$ [-0.059, $+0.204$]
Frontier	-0.074 [-0.218, $+0.046$]	-0.165 [-0.259, -0.029]
Qwen3-14B	$+0.009$ [-0.300, $+0.190$]	$+0.010$ [-0.227, $+0.178$]
GLM-4-9B	$+0.224$ [$+0.019$, $+0.463$]	$+0.233$ [$+0.034$, $+0.467$]
Phi-4	$+0.036$ [-0.140, $+0.186$]	-0.026 [-0.263, $+0.264$]
DeepSeek-V4-Pro	$+0.057$ [-0.154, $+0.298$]	$+0.021$ [-0.152, $+0.276$]
Qwen3.7-Max	-0.080 [-0.522, $+0.100$]	-0.284 [-0.576, $+0.092$]
GLM-5.1	-0.143 [-0.376, $+0.143$]	-0.274 [-0.478, -0.011]
Doubao-Seed-2	-0.128 [-0.333, $+0.105$]	-0.124 [-0.339, $+0.184$]

E ADDITIONAL DATASET QUALIFICATION

These resources were considered for external validation but do not provide additional successful replications. We retain the exclusions and failed pilots to distinguish dataset search from evidence supporting the paper's conclusions.

Table 28: Additional resources and why they are not counted as external confirmation.

resource	evaluation stage	outcome and evidence status
LongMemEval-S	structural screening	One question per distinct history does not supply the within-history fact ranking used here. No compaction experiment is claimed.
PersonaMem-v1	full-log/summary pilot	Readability or an identifiable compression-loss gap failed; the full arm was stopped.
RUMBA	full-input QA qualification	Semantic, output-format, and context-documentation gates were not all met. No forgetting forecast or protection effect was tested.
Multi-Session Chat	full-log QA qualification, then protection	Gates met on this setting's own backend, and the four-policy intervention ran at matched measured storage. Counted as external evidence and reported in Section 5.2.

PersonaMem and structural screening. PersonaMem-v1 offered 37 independent contexts and 388 questions under our construction. A two-context, 24-question pilot gave Qwen3-14B full-log/summary accuracy of 0.458/0.167 and DeepSeek-V4-Pro 0.625/0.625. The latter has no paired compression-loss gap, so a full intervention would not identify the intended effect. LongMemEval-S has one question per history (Wu et al., 2025); other screened candidates lacked the evidence or fact-level questions required by this protocol. We did not redefine the task after observing results.

RUMBA qualification, not another memory experiment. We used the English synthetic-dialogue release at revision `a20471ab`, with the full pin and MIT license in the supplement (`https://github.com/ai-forever/RUMBA`). Following earlier exploratory local-model diagnostics, a later pilot used 14 data-reviewed questions, one per distinct source not used in those runs. It tested full-input QA only. The gateway-served Qwen3.7-Max alias was not independently mapped to a public model release.

One human reviewer judged 12 answers correct, 1 wrong, and 1 unresolved. A second reviewer inspected only the unresolved answer and also left it unresolved: an added causal statement was not established by the dialogue, and the rubric did not settle its effect on overall correctness. This is not a fully double-rated sample. Keeping all questions yields semantic bounds of 85.7–92.9%, not a confidence interval. Only 13/14 responses were valid JSON; the malformed response was semantically correct, and its format status was not repaired.

The original gate required every approved question to be answered correctly without malformed or truncated output. It was not passed. Provider-confirmed context documentation was missing, and the attached receipt did not match the configured window. A reported long-input acceptance test lacked raw responses and cannot exclude silent truncation. The fixed labels, unresolved case, and numerical closure are retained without excluding difficult questions or relaxing the gate. This pilot supplies no confirmatory evidence for compaction, self-predicted forgetting, or protection.

F METRIC GUIDE AND WORKED EXAMPLE

As an illustration, suppose a chat says that a meeting starts at 3 PM. Its pre-compaction ranking is evaluated against whether 3 PM appears in the untreated summary. A protection policy may instead say "keep the meeting time," naming the key but not supplying 3 PM. We summarise the original chat again and ask the same question. The first test measures whether the model anticipated an omission; the second measures whether the protection policy changed the outcome for that fact (Figure 1).

Locate the observed failure stage. In the controlled logs, an incorrect full-log answer indicates a readability or answering problem that cannot be attributed to compaction alone. If the full-log answer is correct but the reference value is absent from the summary, we record a write-side omission. If the value is present but the answer is wrong, the failure occurs when using the stored information. The memory loop additionally records whether that value reached the retrieved context, separating retrieval misses from answer failures. These checks distinguish observable stages, not internal neural mechanisms. On controlled materials, 84–100% of facts present in the summary are then answered correctly, making writing the main loss stage under this protocol.

Metrics and their roles. Summary survival isolates what was written; answer recall measures what the resulting memory can answer. Prediction AUROC measures how well the ranking separates positive from negative outcomes; each table states whether that outcome is summary survival or correct answering. For illustration, AUROC 0.70 is an ordering score, not a claim that 70% of facts survive. Recall 0.70 means 70% of the scored questions are answered correctly. Stored characters measure actual cost, since equal requested limits do not guarantee equal writing.

Table 29: How to read the main quantities. A change of $+0.10$ in recall means ten percentage points, not a ten-percent relative improvement.

quantity	question answered	interpretation
summary survival	Was the reference value written into the summary?	Fraction of facts present in the summary; higher means fewer write-side omissions.
answer recall	Can the model answer using only stored memory?	Fraction of questions answered correctly; higher means better final recall.
prediction AUROC	Does the forecast rank retained or answerable facts above lost ones?	0.5 is random ordering and 1.0 is perfect ordering; this is not a retention rate.
stored characters	How much memory did the policy write?	Actual characters per session; smaller is cheaper at the same recall.

G MULTI-SESSION CHAT: FULL EXECUTION AND ANNOTATION DETAILS

LoCoMo leaves two things unresolved: its requested capacity did not bind actual writing, and its unprotected-fact effects were mixed. We therefore ran the same four policies over Multi-Session Chat, crowdworker five-session dialogues (Xu et al., 2022), with storage controlled by measurement rather than by request. 12 development dialogues each carry six human-checked facts. Every policy writes one free-prose memory record per dialogue aimed at 200 reference tokens of a pinned reference tokenizer, accepting 190–210; a draft outside the band receives up to two rewrites whose feedback contains only measured length, never answer outcomes, and the draft handed to the reader is the first in band or otherwise the one closest to the target. The writer sees the transcript and the protected key names, never their values; the reader sees only the record. This yields 288 logical fact conditions from 264 unique answers, because identical protected-key sets inside a dialogue share one realisation. The backend is a single frontier model that first re-passed the full-log gate on the same 72 questions (per-source recall 1.00, 0.986 if the one undecided answer is counted wrong); the earlier backend's qualification was not carried over.

Storage is matched by measurement, and mostly within tolerance. 42 of 48 records land in the accepted band. The quantity that licenses the paired comparison is the spread between the four policies of one dialogue: its median is 10 reference tokens, and it stays within the tolerance the band implies for 11 of 12 dialogues, reaching 25 on the remaining one. We report that dialogue rather than dropping it, so "equal cost" here means equal measured storage in 11 of 12 cases, not uniformly.

Total gain appears without a detectable unprotected cost. Table 5 gives both effects. All three protection policies raise recall over no protection, and self-selection is the largest at $+0.231$ with 10 of 12 dialogues improving and 2 worsening. Its exact two-sided signed-rank p is 0.020, which becomes 0.059 once the three policies are counted, so this is a directional result rather than a corrected significant one; setting every undecided label to wrong moves it to $+0.194$ with $p = 0.065$. On the four facts each policy leaves unprotected, compared against no protection on those same four, no arm is negative: $+0.000$, $+0.097$ and $+0.042$. This is the cost that the controlled budget produced, and it does not appear here. With 12 paired dialogues and four facts per comparison, a loss of one fact in at least 6 dialogues with no offsetting dialogue would have been detected at the 0.05 level; smaller losses remain possible, so this is a non-detection and not a demonstration of no cost. As a selection diagnostic, the model's two picks were facts that unprotected compaction actually lost 0.583 of the time, which is a hit rate against observed omissions and not a ranking evaluation.

What this setting cannot settle. The dialogues are development sources. Two reviewers labelled them: one judged all 288 conditions, and a second independently relabelled a stratified sample of 96 decidable answers without seeing the first reviewer's work, agreeing on answer correctness 95.8% of the time ($\kappa = 0.89$, bootstrap [0.77, 0.98]). Fact retention agreed on all 47 answers both reviewers were willing to judge, but that is the smaller half of the story: on the 264 unique answers, 84 retention judgements and 8 correctness judgements remain undecided after both passes, because the stored record does not bind the fact to the session the question names. The second reviewer resolved only 1 retention and 1 correctness case and abstained on the rest, so this is a property of the material rather than of one grader, and it is why the retention column is weaker evidence than the recall column. Both self-selection and the text cue see the questions when choosing what to protect, which is informative but not deployment-realistic. 3 of 288 answers violated the answer schema and were retained and flagged rather than repaired or resampled. Finally, the backend emitted provider-side reasoning alongside 96 of 348 responses; that text never entered the stored record or reached the reader, but it means the arms are matched on storage cost and not on provider compute.

H KEY-FREE INSTRUCTION CONTROLS ON MULTI-SESSION CHAT

Prompt variants and comparisons. The original four policies share the writer-system template and vary the key list: random-minus-none tests naming random facts; self-minus-random tests selection. Added *plain* removes the key-list mechanism and `protected_keys` payload. Added *emphatic* augments the original empty-list *none* scaffold with a within-budget completeness request. Thus emphatic-minus-none targets wording; emphatic-minus-plain changes wording, scaffold, and payload, not just one sentence.

Both prompts already request factual accuracy and speaker/session preservation; the tested incremental emphasis is:

> Within the stated budget, retain the archive's facts as accurately and completely as possible; prefer covering more distinct facts over elaborating any single one.

No arm is required to keep all facts while obeying an incompatible fact-count cap.

Execution and grading. The new arms were generated in `opus5_ladder_v10`, separately from the original `opus5_protection_v7` four-arm run, on the same frozen 12 dialogues and six questions per dialogue, with the reported same backend, reader prompt, and length controller. The additional run records 185 calls, no transport failures, and 144 answers across 24 memory records. One reviewer scored the new answers; 43 retention judgments remain undecided owing to session binding. The four malformed answers were retained for semantic grading, not repaired or resampled. Cross-run contrasts pair source dialogues but do not remove possible run effects.

Recall and prompt contrasts. Reported recall for the new arms is 0.569 (plain) and 0.556 (emphatic). Table 30 gives the six paired dialogue-level contrasts, which are distinct from differences between rounded pooled recall rates in Table 5.

Table 30: MSC instruction ladder, paired over 12 dialogues. Differences are in answer recall; w/l/t counts source wins/losses/ties. The first three rows compare key-free prompts. All p values are nominal, without correction for the six displayed contrasts; the additional self-minus-plain contrast is reported below. Rows involving plain or emphatic and an original policy are cross-run comparisons.

contrast	prompt or key change	Δ recall	w/l/t	p
emphatic − plain	Two key-free prompt variants	-0.014	3/4/5	1.000
none − plain	Empty-list scaffold vs plain compaction	-0.067	1/7/4	0.195
emphatic − none	Added completeness request	$+0.053$	5/4/3	0.590
random − none	Two random keys vs empty list	$+0.094$	7/3/2	0.371
self − random	Selected vs random keys	$+0.136$	5/2/5	0.078
self − none	Selected keys vs empty list	$+0.231$	10/2/0	0.020

Self-minus-plain is $+0.164$ (10/12 wins; nominal $p = 0.036$), a baseline sensitivity check, not a multiplicity-corrected finding. Plain has the highest observed key-free mean, but is not a universally hardest comparator.

Measured storage and interpretation. Mean storage is 205.4 reference tokens for plain and 209.7 for emphatic. Their accepted-band counts are 11/12 and 7/12, respectively; all out-of-band records exceed the upper limit. Table 31 summarizes these archived measurements without filtering the out-of-band records.

Table 31: Measured reference-token storage for the two new key-free arms. The accepted band is 190–210 tokens. All out-of-band records are above it. These are storage observations, not human recall scores.

arm	mean tokens	minimum	maximum	in band
plain	205.4	197	220	11/12
emphatic	209.7	194	228	7/12

Longer emphatic memories without an established gain do not exclude length-mediated mechanisms: storage is imperfectly matched, batches differ, and only 12 sources are tested. Self-minus-plain supports targeted preservation directionally, not key necessity, key-free equivalence, or corrected selector superiority. This total-recall extension narrows the prompting explanation; it neither identifies a unique mechanism nor independently replicates controlled memory-loop collateral loss.