DIAGNOSING FAILURE REGIMES IN LLM AGENT MEMORY WITH EXTRACTION-PASS DISAGREEMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM agents build long-term memory by consolidating interaction history into a structured store, and this consolidation is lossy in two ways. In the first, a stochastic pass drops information that another pass may keep, and computation removes the error. In the second, a capability-limited backbone compresses history into a low-rank sketch, and the bias it introduces is beyond the reach of computation. The two failures look the same in aggregate accuracy, so existing systems apply the same remedy to both. We propose Regime-Routed Construction (RRC), a training-free configuration layer whose routing follows the diagnosed failure regime. RRC repeats consolidation on the same history and reads how much the passes disagree. The disagreement estimator, combined with backbone capability, routes the vote budget, the retrieval depth, and the answer policy. A noisy reading brings more extraction passes merged by majority vote. A quiet reading deepens retrieval and calibrates the answer policy when the backbone sits above its capability threshold, and declines what the store cannot support below it. A Fano-type argument shows that below a capability threshold, error on relational questions is irreducible. Across two benchmarks and backbones from 1.5B to frontier API models, RRC improves on the strongest published LoCoMo result by $+4.7$ F1 and outperforms the RL-trained Memory-R1 at matched backbones without gradient updates. On LongMemEval, the gains are small and insignificant, consistent with a diagnosis of little headroom.

1 INTRODUCTION

LLM agents that serve users over hours and days must remember what happened during interactions. However, context windows are bounded while raw interaction history keeps growing. Practical memory systems consolidate it into structured stores: MemoryBank and Mem0 distill history into summaries, Graphiti maintains a temporal knowledge graph, and MemOS organizes memory as an operating system (Zhong et al., 2024; Xu et al., 2025; Rasmussen et al., 2025; Chhikara et al., 2025; Li et al., 2025). Unfortunately, consolidation is inherently lossy, and question-time answering reads the store instead of the raw history. The agent cannot recall what the store loses.

Despite much effort to mitigate it, the essential challenge is that consolidation fails in two ways that are easy to conflate and respond to opposite remedies. In the first kind of loss, a stochastic extraction pass drops information that is still present in the raw history. Repeating the pass and voting recovers what one pass dropped, as self-consistency voting does for sampled reasoning paths (Wang et al., 2023; Chen et al., 2023). In the second kind of loss, a capability-limited backbone compresses history into a low-rank sketch of its relational structure (Woodruff, 2014) and destroys whatever falls outside the sketch. No retrieval depth and no number of extra passes can restore this loss, because the backbone itself cannot represent what falls outside the sketch. Figure 2 in Appendix C walks one LoCoMo conversation through both regimes: a history window that a 7B backbone consolidates with recoverable pass noise is consolidated by a 1.5B backbone into a loss that no pass can see. Aggregate accuracy applied in previous studies (Zhong et al., 2024; Xu et al., 2025; Rasmussen et al., 2025; Chhikara et al., 2025; Li et al., 2025) cannot distinguish them.

The closest related work comes in two lines, and each covers only part of the problem. Test-time-compute methods treat agreement among repeated samples as a signal (Wang et al., 2023; Chen et al., 2023; Farquhar et al., 2024; Kuhn et al., 2023), but they act on answers rather than on construction, and they average the statistic away instead of keeping it as a measurement. Scaling studies show that memory quality depends heavily on backbone capability (Wu et al., 2026a) and that rate-distortion analysis bounds what context compression can preserve (Arumugam & Van Roy, 2021; Nagle et al.,

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.

2024). However, no prior work joins the two: it neither computes agreement over the construction process itself nor routes memory consolidation by the failure regime that the reading reveals.

Joining the two parts is non-trivial: a failing construction carries no label for why it fails and the store holds only what survived, so the missing content is invisible from inside the system. Moreover, a store of bounded rank preserves bounded information about the history, and Fano's inequality (Cover & Thomas, 2006) turns this bound into a hard floor on error: below a capability threshold, part of the loss is irreducible no matter how retrieval and answering are designed. Nor can pass-level disagreement see this loss: whatever the sketch drops, it drops identically in every pass. Our insight is to use the construction's own stochasticity as the measuring instrument. Disagreement among repeated passes exposes the recoverable part of the loss. A quiet reading excludes it, leaving irreducible loss as the only explanation.

We build the method on these two instruments: the disagreement reading and the capability threshold. Regime-Routed Construction (RRC) is a training-free configuration layer that treats memory construction as diagnose-then-configure. It probes the backbone with repeated extraction passes and reads disagreement among them as the estimator. When the reading is noisy, RRC spends more extraction votes. When the reading is quiet and the backbone sits above its capability threshold, RRC deepens retrieval and calibrates the answer policy. Below the threshold the loss is irreducible rather than undiagnosed, and RRC declines the questions the store cannot support.

We make three contributions.

- We identify two failure regimes of lossy memory consolidation that previous systems conflate, and aggregate accuracy cannot distinguish.
- We propose RRC and prove three properties of it: the estimator tracks the first regime only, a capability threshold exists below which error is irreducible, and a degrading backbone loses relational structure before isolated facts.
- We evaluate on two benchmarks and backbones from 1.5B to 72B. On LoCoMo, RRC reaches 57.8/58.5 overall F1, $+4.7$ over Kumiho (Park, 2026), the strongest reproducible published result, and exceeds the RL-trained Memory-R1 at matched backbones without any gradient updates. On LongMemEval, the gains are small and insignificant, consistent with a diagnosis of little headroom.

2 Related Work

Memory consolidation systems. A long line of studies explores this design space, from MemoryBank (Zhong et al., 2024) and A-MEM (Xu et al., 2025) through Zep/Graphiti (Rasmussen et al., 2025), Mem0 (Chhikara et al., 2025), and MemOS (Li et al., 2025), to GAM (Wu et al., 2026b), the closest neighbor in architecture, and Kumiho (Park, 2026), the strongest published LoCoMo result. However, none of them provides a statistic that separates failure regimes or a routing policy that follows from it, and GAM performs single-pass extraction without a noise estimator, cross-scale study, or decline policy. Synapse (Jiang et al., 2026) spreads activation over an episodic-semantic graph and rejects low-confidence retrievals. Its cross-backbone appendix is the closest precedent for the backbone ladder we study. Finally, Wu et al. (2026a) evaluates ten memory architectures across open backbones and concludes that the quality is heavily backbone-dependent.

Test-time compute and agreement as signal. The inference-time-compute literature mainly acts on answer construction. Self-consistency (Wang et al., 2023; Chen et al., 2023) ensembles sampled reasoning paths and reads agreement as an implicit confidence score, and semantic entropy (Farquhar et al., 2024; Kuhn et al., 2023) reads disagreement among generations to detect hallucination.

Information-theoretic limits and scaling thresholds. Memory consolidation is lossy compression: the backbone's capacity caps how much the store preserves. Rate-distortion analyses of context compression (Nagle et al., 2024) derive impossibility bounds of the same genre as ours. Capacity thresholds for transformer memory appear in Allen-Zhu & Li (2025) and Madden et al. (2025). Inverse scaling (McKenzie et al., 2023; Wei et al., 2023) and the emergent-abilities debate (Schaeffer et al., 2023) show capability moving non-monotonically with scale. Our consolidation-as-sketching model extends this footing to relational memory.

3 Method

3.1 Setup and two regimes of forgetting

We abstract the memory construction stage as an operator. Let $\mathcal{H}$ be the space of interaction histories. The *backbone* is the served LLM that runs extraction and answers questions. A *consolidation operator* $S_C : \mathcal{H} \to \mathcal{M}$ maps a history H to a memory structure $M = S_C(H)$, where C is a

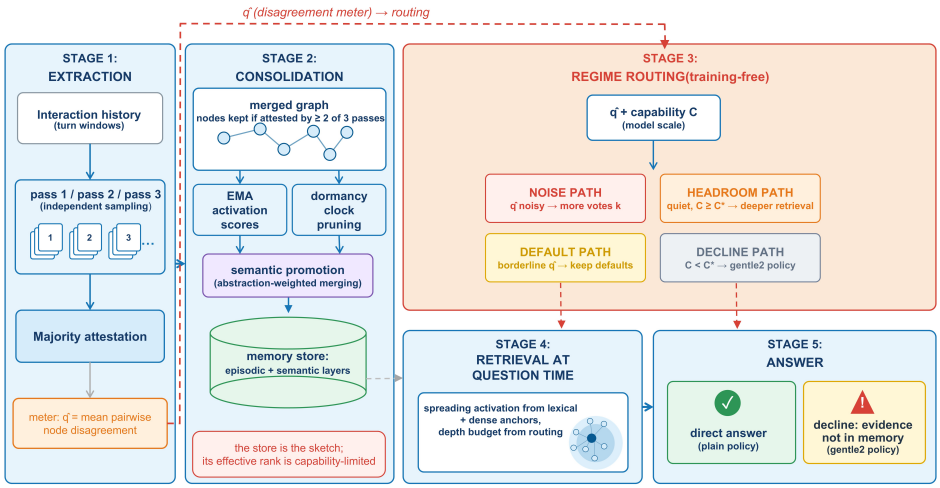


Figure 1: Regime-Routed Construction. Extraction runs k independent passes and merges by majority vote, and recorded disagreements define the estimator $\hat{q}$. The routing layer reads $(\hat{q}, \text{capability})$ and sets three parameters, vote budget, retrieval depth, and the answer policy.

scalar capability index that is monotone in measured backbone quality. We model S_C as a sketch: a representation $I(H)$ of the history's information content, ideally the node-edge graph an unconstrained extractor would write, is passed through a rank-limited projection Π_r of effective rank $r = g(C)$. We instantiate C as log parameter count, with closed models placed above the open ladder (Kaplan et al., 2020; Hoffmann et al., 2022; Wei et al., 2022; McKenzie et al., 2023). The consolidated memory keeps the component of $I(H)$ that lies inside the sketch and loses the rest, $S_C(H) = \Pi_{g(C)}\big(I(H)\big) + \varepsilon$, and the two regimes live in the two terms. The decomposition separates what computation can remove from what it cannot. These asymmetric definitions guide the configuration and routing in our method.

Definition 1 (Regime I: extraction noise). Extraction is a stochastic algorithm, and independent passes $i = 1, \ldots, k$ yield $S_C^i(H) = \Pi_{g(C)}(I(H)) + \varepsilon_i$ with pass-specific noise ε_i. Equivalently, each pass draws an independent perturbation of the same sketching operator. Because the information missed by a single-pass still exists in H, losses in this regime are reducible by computation.

Definition 2 (Regime II: representational loss). The systematic loss is $L_{II} = I(H) \backslash \text{span}\big(\Pi_{g(C)}\big)$. Whatever the sketch's rank budget cannot afford is lost at consolidation time. They are irreducible by any downstream computation as a retrieval-and-answering scheme is a function of M alone.

3.2 Diagnosing the two regimes

The first loss is detected by statistics on repeated runs of the pipeline's own extraction stage: run extraction k times with independent sampling and average a node-level disagreement measure $d(\cdot, \cdot)$ between the extracted graphs (§4.1). The *extraction-noise estimator*, or meter, is the U-statistic $\hat{q}_k = \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} d\big(S_C^i(H), S_C^j(H)\big)$ over pass pairs: all passes share the same systematic component $\Pi_{g(C)}(I(H))$, so the pairwise disagreement reflects the pass-specific noise ε_i of §3.1, and therefore the estimator indicates Regime I directly.

As detecting Regime II requires labels on the outcomes, the proposed method never directly obtains this coordinate, and operationalizes failure by the capability threshold $C^*(d)$ (§3.3) instead. The quiet cut is calibrated once from the meter's own readings, with no labels: the readings of any set of systems under the same extraction configuration separate into a quiet group and a noisy group, with a gap in between. A system reads noisy when its $\hat{q}$ interval lies entirely above the gap, quiet when entirely below, and borderline when the interval straddles it. Any cut inside the gap routes every system identically.

3.3 Regime-Routed Construction

RRC is the configuration layer (Figure 1). Its input is the diagnosis $(\hat{q}, C, d)$ and its output is a configuration of three parameters: the vote budget, the retrieval depth, and the answer policy. The pipeline it configures has five stages. **Extraction** and **consolidation** determine what the store M

contains: they suppress Regime I noise and incur Regime II loss, while **retrieval** and the **answer policy** operate on M and can only manage the second. The full procedure of RRC is introduced in Algorithm 1 (Appendix B).

Extraction. Extraction turns raw dialogue into candidate memory content, and its design goal is to suppress Regime I noise. It opens with a probe of k_0 passes (Algorithm 1), and each history window of W turns is extracted k times in total, pass i producing a graph $S_C^i(H)$. The merged graph keeps only the content that receives a strict majority of the k votes,

$$v \in M \;\Leftrightarrow\; \sum_{i=1}^{k} \mathbf{1}\{v \in S_C^i(H)\} > k/2,$$

so a node or edge survives only if more than half of the passes agree. Each pass records its node-name set N_i, and pass-to-pass disagreement is summarized as

$$\hat{q} \;=\; \frac{2}{k(k-1)} \sum_{1 \le i < j \le k} \left(1 - \frac{|N_i \cap N_j|}{|N_i \cup N_j|}\right),$$

the mean Jaccard distance over pass pairs. The merged graph serves retrieval, and $\hat{q}$ serves routing below. The vote budget could be set by two rates the theory supplies: Proposition 2 makes majority error decay exponentially in k, so extraction accuracy saturates after a few passes, while Proposition 6 gives the estimator confidence width of order $k^{-1/2}$, making the estimator the rate-limiting consumer of votes.

Consolidation. Consolidation decides what the store M keeps as the dialogue grows. Its core quantity is a per-node activation score a_v, an exponential moving average of node v's recent retrieval use,

$$a_v \leftarrow (1-\alpha)\, a_v + \alpha\, \mathbf{1}\{v \text{ retrieved in the current window}\},$$

together with a dormancy clock. A node whose a_v stays below threshold for W_{dorm} consecutive windows is pruned, and the survivors are promoted to a semantic layer where nodes are merged with weights that favor abstraction edges. Consolidation is where Regime II loss is incurred: whatever pruning and promotion remove is gone before retrieval runs, and no later stage can recover it. The loss is administered here, since any capability-limited store must drop the same content and the ranking only decides what the budget keeps.

Routing. Routing turns the diagnosis into a configuration of three parameters, the accumulated vote budget k, the retrieval depth, and the answer policy π,

$$(k,\ \text{retrieval depth},\ \pi) \;=\; R(\hat{q}, C, d).$$

The vote budget k is the total number of independent extraction passes including the probes, the retrieval depth is the top-K cutoff and the spread budget T, and the policy π is one of two answer policies defined below. Routing reads three quantities, the estimator $\hat{q}$, the capability index C, and the reasoning depth d read from the question taxonomy. As RRC reads no outcome and no label, it cannot use the diagnostic plane's outcome axis, and the threshold comparison takes its place. The capability threshold is

$$C^*(d) \;=\; g^{-1}\big((s_d - \log 2)/\kappa\big),$$

from Proposition 3, with the map g of §3.2, the structure dimension s_d of the task axiom (§3.4), and the per-dimension sketch rate κ. Routing runs once per backbone, immediately after the probe, and the selected configuration is held fixed for the entire stream. It moves only the three parameters above and never edits the store M. Table 4 (Appendix B) restates the same rule by diagnostic signature. Then routing selects one of four paths, each stated below with its trigger, signature, and prescription. Only the decline path changes the answer policy.

Noise path. Triggered by a high $\hat{q}$ at or above the threshold. The signature is Regime I, extraction fluctuation, and the prescription is to spend votes. The probe adds extraction passes until the bootstrap interval of $\hat{q}$ tightens, and the tightened budget is the routed k. A wide interval means the noise level is measured imprecisely, and more passes tighten the estimate. And extra passes are safe because majority merge removes only what the passes disagree on.

Headroom path. Triggered by a quiet $\hat{q}$ together with $C \ge C^*(d)$. The signature is Regime II with headroom, and the prescription is to deepen retrieval by raising the top-K cutoff and the spread budget T. The quiet estimator excludes noise and the threshold comparison says the backbone can sketch the depth-d content, so the loss lives at retrieval selection and deeper retrieval is the parameter that reaches it.

Decline path. Triggered by $C < C^*(d)$, regardless of $\hat{q}$. The signature is Regime II below the capability threshold, where the loss is irreducible, and the prescription is the gentle2 policy, which marks exactly this risk by declining whenever the retrieved support is missing. This path indicates a backbone whose sketch cannot cover the depth-d requirement answers from bias.

Default. A borderline $\hat{q}$ at or above the threshold triggers no path. The pipeline runs as written, under the plain policy, with no loss that routing can address.

Retrieval. Retrieval fetches the content a question needs from M: its input is a question x and its output is a ranked subgraph. The two budget parameters, spread depth and inhibition strength, are set by routing stage. First, seeding selects seed nodes in M from lexical anchors and dense nearest neighbors of x. Second, spreading activation runs from the seeds along typed edges for a budget of T steps, updating node scores as $a_u \leftarrow \sum_{v \to u} w_{\tau(v,u)}\, a_v$ with edge-type weights w_τ, and renormalizing competing branches so that one branch cannot starve the rest. Third, ranking orders nodes by accumulated activation and returns the top-K subgraph to the answer stage. The returned subgraph is the candidate evidence, and the part of it that is relevant to the question is its support.

Answer policy. The answer stage converts the retrieved subgraph into the reply the user sees, and its design question is what to do when the support is missing. The *plain* policy answers directly from the retrieved subgraph. When retrieval returns no supporting content, it still answers from the backbone's parametric knowledge, which is a guess the reply does not mark as such. The *gentle2* policy answers only with retrieved support and otherwise states that the evidence is not in memory, which is where honest declining lives.

3.4 Theoretical analysis of regime separation and routing

Section 3.1 defined two losses and §3.3 built a routing rule on them. The theoretical analysis answers four questions. First, can the first loss be removed with more computation? Proposition 2 shows that more computation removes it at an exponential rate. Second, is there a level below which the second loss cannot be removed? Proposition 3 gives the level: below the capability threshold $C^*(d)$, no amount of extra computation removes the loss, and Proposition 4 says what to do there. Third, which content is lost first as capability drops? Proposition 5 shows that higher-order structure is lost first. Fourth, can the estimator tell the two losses apart? Proposition 6 says it tracks the first loss alone, which is why a quiet estimator reads as an exclusion of Regime I. Proposition 1 shows the split is complete: every construction error is one of the two terms. Detailed proofs are in Appendix A.

The ensemble behind the estimator. Take a window w with candidate-node universe U_w. Pass i of extraction produces a vote field $X_i \in \{0,1\}^{U_w}$, with $X_i(v) = 1$ when pass i votes for node v, and consolidation merges fields by majority vote. Let $p_v = \mathbb{P}[X_i(v) = 1]$ be the per-pass vote probability and $q_v = 2p_v(1-p_v)$ the probability that two independent passes disagree on v. This p_v is the node-level counterpart of the fact-retention probability p_f of Lemma 1. The estimator of §3.2 is the mean Jaccard distance between the passes' node-name sets, averaged over windows and pass pairs. It is a U-statistic, and its expectation is the average pairwise disagreement. The split of construction error into fluctuation and bias mirrors the classical bias-variance decomposition (Geman et al., 1992; Domingos, 2000).

Proposition 1 (**one construction error, two terms: fluctuation plus bias**). *In the voting model with independent passes, fix a node v with truth value $t_v \in \{0,1\}$. If the expected vote lands on the correct side of one half, $p_v > \frac{1}{2}$ when $t_v = 1$ and $p_v < \frac{1}{2}$ when $t_v = 0$, majority vote errs with probability at most $\exp(-k(\frac{1}{2} - q_v))$. If it lands on the wrong side, majority vote errs with probability at least $1 - \exp(-k(\frac{1}{2} - q_v))$. Construction error is therefore the sum of a fluctuation term and a bias term: the correct-side case is Regime I and the wrong-side case is Regime II.*

The fluctuation term falls in k at the rate the estimator tracks. The bias term comes from the shared sketch: whatever it drops is dropped in every pass, so no number of votes can remove it.

Remark 1 (**the estimator measures distance, not direction**). The map $p_v \mapsto 2p_v(1-p_v)$ is symmetric about $p_v = \frac{1}{2}$, so the estimator tells how far the expected vote is from the majority threshold, but not on which side. Regime II is the wrong-side case: the expected extractor is itself a biased sketch, and majority vote converges to that bias with certainty (White, 1982). This symmetry is the reason for the limits of Proposition 6: disagreement measures how decisively the backbone extracts, never whether what it extracts is true.

Assumptions. The analysis rests on five assumptions. Passes are independent copies of one sketch (Assumption 1), relaxed by Proposition 2 to a common-correlation model. Capability buys sketch rank monotonically (Assumption 2). Reasoning depth carries an information dimension (Assump-

tion 3). A rank-limited store carries bounded information (Assumption 4). Structure competes for rank in a fixed order (Assumption 5).

Assumption 1 (**passes are independent copies of one sketch**). (Wang et al., 2023; Wood et al., 2023). Across independent passes, the noise terms ε_i in the decomposition are independent and identically distributed with bounded second moment.

Assumption 2 (**capability buys sketch rank monotonically**). (Allen-Zhu & Li, 2025; Madden et al., 2025). The rank budget $g(C)$ is continuous and strictly increasing in C.

Assumption 3 (**depth d has an information size s_d**). (Arumugam & Van Roy, 2021). To answer depth-d questions correctly, the system must tell apart N_d history states, where $\log N_d = s_d$.

Assumption 4 (**a rank-r sketch carries at most κr bits**). (Woodruff, 2014). A rank-r sketch carries at most κr bits about the history: $I(H; M) \le \kappa\, r$ for a universal constant $\kappa > 0$.

Lemma 1 (**facts above one half survive majority vote**). *Let a fact f be retained by a single extraction pass with probability $p_f > \frac{1}{2}$. The probability that a majority of k independent passes omits f is at most* $\exp\big(-k(2p_f-1)^2/2\big)$.

Proposition 2 (**Regime I noise shrinks exponentially in the vote budget**). *Under Assumption 1, a fact retained with per-pass probability $p_f > 1/2$ is preserved by k-pass majority vote with failure probability at most* $\exp(-k(2p_f-1)^2/2)$. *If passes have common pairwise correlation ρ in their retention indicators, the bound holds with k replaced by $k_{\text{eff}} = k/(1+(k-1)\rho)$.*

Corollary 1 (**votes needed: $k \ge 2\log(1/\delta)/(2p_f-1)^2$**). *Inverting Proposition 2, k-pass majority vote fails on a fact with per-pass retention $p_f > 1/2$ with probability at most δ whenever $k \ge 2\log(1/\delta)/(2p_f-1)^2$.*

The budget grows with the square of the margin above one half: a backbone that barely clears the threshold needs quadratically more votes than a decisive one. The probe of §3.3 sets the budget from this margin.

Remark 2 (**the majority threshold cuts both ways**). Proposition 2 requires $p_f > 1/2$. Majority vote keeps above-half facts and drives out below-half facts at the same exponential speed, and no number of passes recovers a fact whose per-pass retention sits below one half. The noise path therefore spends votes only when the estimator reads noisy. A quiet reading excludes Regime I.

That is all of Regime I: the noise is bounded, and more passes remove it. Regime II is different, and the next proposition gives its boundary.

Proposition 3 (**below the critical rank, error cannot reach zero**). *Let $r^*(d) = (s_d - \log 2)/\kappa$. If $r < r^*(d)$, every retrieval-and-answering scheme answers depth-d questions with expected error*

$$P_{\text{err}} \;\ge\; 1 - \frac{\kappa\, r + \log 2}{s_d} \;>\; 0. \tag{1}$$

We call $r^*(d)$ the critical rank for depth d.

The bound is positive exactly when $r < r^*(d)$: below the critical rank, no amount of extra computation drives the error to zero. By Assumption 2, capability C buys rank $g(C)$, so the bound reads as a capability threshold. It holds for every $C < C^*(d) = g^{-1}(r^*(d))$.

This bound meets the two-term split of Proposition 1. If the depth-d content sat on the correct side of one half, k-pass majority vote would keep it, contradicting the bound. So below the critical rank the lost content sits on the wrong side: its per-pass vote probability is below one half. That is Regime II, and it is why more passes do not help.

Remark 3 (**the bound speaks only below the threshold**). Proposition 3 speaks only about the below-threshold side. It does not prove that the recipes help above the threshold. That direction is the design claim of Table 4, tested in §4.

The decline path covers the below-threshold side. What a declining policy gains has a classical name, selective prediction (Chow, 1970; El-Yaniv & Wiener, 2010).

Proposition 4 (**declining trades coverage for risk**). *Let $r < r^*(d)$, and let μ be the fraction of depth-d questions whose required content the store carries. The plain policy answers every question and has expected error at least $1-(\kappa r+\log 2)/s_d$ by Proposition 3. The gentle2 policy answers only questions whose retrieved support contains the required content and declines the rest. Its coverage is μ, and its risk, the error on the questions it answers, is only Regime I noise, which votes and deeper retrieval reduce.*

So below the threshold, RRC only answers questions supported by the store.

The second loss does not fall on all content at once. The next proposition gives the order in which structure is lost as the rank budget shrinks. It rests on one more structural property of the sketch.

Assumption 5 (lower-order structure is kept first). When several kinds of structure compete for the rank budget, the lower-order kind is kept first: if the budget lies between r_1^* and r_2^*, the sketch keeps all of the order-1 structure, and only the budget left over is available to order-2 structure.

Any consolidation ranked by evidence concentration keeps this order: order-1 facts receive more votes and more redundant support than cross-session bridges. The merging of §3.3 has this bias.

Proposition 5 (higher-order structure dies first). *Let $s^{(1)} < s^{(2)} \le s_d$ be the intrinsic dimensions of order-1 structure (isolated facts) and order-2 structure (cross-session bridges and temporal chains), and define $r_o^* = (s^{(o)} - \log 2)/\kappa$. A backbone with rank budget $r \in [r_1^*, r_2^*)$ keeps order-1 structure, up to Regime I noise, while order-2 structure keeps positive error bound of Proposition 3.*

So as capability degrades, the loss concentrates on higher-order structure first. Lower-order structure survives down to much lower capability.

The last proposition is about the estimator itself: what it can see and what it cannot.

Proposition 6 (the estimator tracks noise, not truth). *Under Assumption 1, as $k \to \infty$ the estimator $\hat{q}_k$ converges to $\bar{q}_C = \mathbb{E}[d(S_C^i, S_C^j)]$, a fixed number that depends only on how the backbone votes on each pass, with bootstrap confidence width $O(k^{-1/2})$. The distance d never sees gold labels, so a backbone that extracts wrong nodes with the same per-pass probabilities produces the same reading.*

Corollary 2 (routing mistakes vanish as $O(1/(k\Delta^2))$). *Let Δ be the distance of the backbone's limit reading $\bar{q}_C$ from the nearest routing boundary of Table 4 (the quiet threshold or $C^*(d)$). With k probe passes, the probability of selecting the wrong path is $O(1/(k\Delta^2))$ by Chebyshev's inequality.*

Routing mistakes shrink as the probe gathers votes, and routing once per backbone keeps the configuration fixed.

4 Experiments

4.1 Setup

Benchmarks. LoCoMo (Maharana et al., 2024) spans 10 sessions with 1,986 questions under an official five-category protocol. LongMemEval (Wu et al., 2025) contains 500 questions over long multi-session histories. LoCoMo questions are mostly single-span, while LongMemEval's multi-hop and temporal classes require cross-session evidence.

Backbones and hardware. We evaluate on Qwen2.5-1.5B/3B/7B/14B/72B and DeepSeek-V3, with GPT-4o as a reference anchor. Backbones up to 14B are served with vLLM on a single RTX 4090D, 14B quantized to int4 (AWQ); 72B and DeepSeek-V3 backbones run through their APIs.

Metrics. On LoCoMo we report overall F1 against the official answer keys. LongMemEval is scored with weighted micro-F1 over answer strings, with its official LLM-judge protocol reported alongside (Appendix B.1). Δ is the paired bootstrap gain of the RRC row over the same-backbone baseline (10,000 resamples, 95% interval in brackets), and P the two-sided paired-bootstrap p.

Compared methods. Table 1 compares against published memory systems at matched backbones: the strongest published LoCoMo result (Kumiho, 56.5 F1 with GPT-4o), the RL-trained Memory-R1, and A-Mem, GAM, Synapse, Zep, and MemOS. The baseline pipeline is the same system without RRC: a single extraction pass, no estimator, no routing, default retrieval depth, and the gentle answer policy. RRC differs only in vote, estimation, and routing.

Implementation. Unless routed otherwise, extraction uses windows of $W{=}5$ turns, $k{=}3$ passes, temperature 0.1. Consolidation runs every $N{=}5$ windows with duplicate-merge threshold 0.92. Retrieval runs spreading activation for $T{=}3$ steps and returns the top $K{=}50$ nodes. The answer stage runs the plain policy by default and the gentle2 policy when routed. The full hyperparameter set is in Appendix B. Code and vote logs will be released upon acceptance.

4.2 Main Results

On LoCoMo, RRC improves over the same-backbone baseline at every backbone from 3B upward, by $+1.22$ to $+9.93$ F1 (every $P < 0.0001$ except 72B at $P = 0.065$), and the headline configuration reaches $57.8/58.5$ across two runs. This exceeds Kumiho's 56.5 and stands $+4.7$ above its pinned reproducible value 53.1 (its headline folds a 97.5 adversarial-refusal score into the average). Within each scale group, the RRC row is first or tied-first among published systems: at 3B 43.39 against the

Table 1: Main results: overall F1 on LoCoMo (official answer keys) and weighted micro-F1 on LongMemEval. Recipe rows average the four non-adversarial categories; headline rows use the official five-category average. The API headline row runs top-75 retrieval with the gentle policy. Bold is best, underline second best per group. Row provenance follows Appendix B.1. Δ is the paired bootstrap on per-question differences and can differ slightly from the difference of the rounded column means.

System	Backbone	LoCoMo F1	Δ [95% CI]	LME F1	Δ [95% CI]
Small models (1.5B–3B)					
GAM (Wu et al., 2026b)	Llama-3.2-3B	36.27		–	
Synapse (Jiang et al., 2026)	Qwen-1.5B	35.9		–	
Synapse	Qwen-3B	36.6		–	
A-Mem (Xu et al., 2025)	Qwen-1.5B	20.4		–	
MemoryBank (Zhong et al., 2024)	Qwen-1.5B	10.0		–	
MemoryOS (Li et al., 2025)	Qwen-3B	22.1		–	
A-Mem (Xu et al., 2025)	Qwen-3B	16.2		–	
Baseline pipeline	Qwen2.5-3B	39.57		35.08	
RRC (ours)	Qwen2.5-3B	**43.39**	$+3.79$ $[+2.31, +5.25]$	33.72	-1.18 $[-4.16, +1.71]$
Baseline pipeline	Llama-3.2-3B	37.18		–	
RRC (ours)	Llama-3.2-3B	40.36	$+3.18$ $[+2.21, +4.15]$	–	
Medium models (7B–14B)					
GAM (Wu et al., 2026b)	Qwen2.5-7B	40.00		–	
Memory-R1‡ (Yan et al., 2026)	Qwen2.5-7B	43.14		46.70†	
Memory-R1‡ (Yan et al., 2026)	LLaMA-3.1-8B	<u>45.0</u>		45.20†	
Memory-R1‡	Qwen2.5-14B	44.40		–	
Baseline pipeline	Qwen2.5-7B	37.85		39.61	
RRC (ours)	Qwen2.5-7B	43.22	$+5.29$ $[+3.65, +6.95]$	38.69	-0.98 $[-3.40, +1.47]$
Baseline pipeline	LLaMA-3.1-8B	36.75		–	
RRC (ours)	LLaMA-3.1-8B	**46.68**	$+9.93$ $[+8.18, +11.64]$	–	
Baseline pipeline	Qwen2.5-14B	38.69		–	
RRC (ours)	Qwen2.5-14B	42.83	$+4.14$ $[+2.22, +6.22]$	–	
RRC (ours), grounded policy	Qwen2.5-14B	**45.05**	$+6.36$ $[+5.04, +8.10]$	–	
Large open model (72B)					
Long-context	Llama-2-70B-Chat	17.9		–	
MemTree (Wu et al., 2026a)	Qwen2.5-72B	34.85		–	
Zep (Wu et al., 2026a)	Qwen2.5-72B	37.45		–	
MemoryOS (Wu et al., 2026a)	Qwen2.5-72B	39.63		–	
MemOS (Wu et al., 2026a)	Qwen2.5-72B	42.79		–	
Tree+Tier (Wu et al., 2026a)	Qwen2.5-72B	<u>43.87</u>		–	
Baseline pipeline	Qwen2.5-72B	43.33		–	
RRC (ours)	Qwen2.5-72B	**44.27**	$+1.22$ $[-0.08, +2.54]$	–	
API models					
MemOS (Li et al., 2025)	GPT-4o-mini	40.0		–	
Synapse (Jiang et al., 2026)	GPT-4o-mini	40.5		–	
Synapse	GPT-4o	43.4		–	
GAM (Wu et al., 2026b)	GPT-4o-mini	43.14		–	
Kumiho (Park, 2026)	GPT-4o	<u>56.5</u>		–	
Baseline pipeline	DeepSeek-V3	44.33		<u>48.81</u>	
RRC (ours)	DeepSeek-V3	53.24	$+8.53$ $[+6.98, +10.11]$	**50.13**	$+1.37$ $[-0.63, +3.36]$
RRC (ours), headline	DeepSeek-V3	**57.77 / 58.46**		–	

best published 36.6, at 7B 43.22 against the best published 43.14, at 8B 46.68 against Memory-R1's 45.0, at 72B 44.27 against the best published 43.87, and at 14B 45.05.

At matched backbones RRC edges out the RL-trained Memory-R1 without any gradient updates (43.22 vs 43.14 at 7B, 46.68 vs 45.0 at 8B). The gain is not uniform in scale: largest at the API backbone ($+8.53$), moderate through 3B–14B ($+3.79$ to $+6.36$), smallest at 72B ($+1.22$), and negative at 1.5B (-5.09). The pattern matches where the losses live: RRC gains where extraction noise dominates, has little room where the baseline approaches the plateau (72B), and cannot help where the store itself is the bottleneck (1.5B). This is the capability threshold made visible. The margin against Kumiho comes from the retrieval classes, single-hop 69.3 and multi-hop 42.3 against 46.2 and 35.5, while Kumiho keeps the temporal and open-domain classes (Table 10), and RRC remains ahead overall ($57.8/58.5$ vs 56.5).

On LongMemEval the column of Table 1 is intentionally sparse: published systems report this benchmark only under the official LLM judge, whose scores are not comparable across systems, and we ran it on the Qwen ladder and the API backbone only. RRC improves at the largest scale ($+1.37$ at API) but no transfer reaches significance (Δ from -1.18 to $+1.37$, every $P \geq 0.18$, Table 9 in Appendix C). The estimator explains the null transfers. Above 1.5B it reads quiet at every backbone on this benchmark (0.52–0.69, Table 9), so there is little noise to remove, and the residual losses concentrate on the store's limitation. The per-category decomposition localizes them: the single-session class loses most at 3B (-8.09), in unanswered or confusable-alternative errors, with multi-session and temporal close behind (-3.65 and -3.87).

Table 2: Ablations on LoCoMo, F1 per backbone; – not run. The API column is the majority-vote plain-policy arm; Table 1's headline row differs (top-75 retrieval, gentle policy).

Variation	3B	7B	API
RRC full method	43.39	43.22	53.24
w/o majority vote (union merging)	43.75	43.07	51.91
w/o gentle2 decline clause (verify-then-decline)	30.12	–	30.33

Table 3: Case study against our Kumiho reproduction at the RRC headline configuration (DeepSeek-V3 backbone); the last column cites each system's evidence.

Class	Question	Gold	RRC (ours)	Kumiho (repro.)	Why RRC wins
Multi-hop	Which US cities does John mention visiting?	Seattle, Chicago, New York	Seattle, Chicago, New York City	New York City	All three passages retrieved; reproduction keeps only New York.
Multi-hop	What musical artists/bands has Melanie seen?	Summer Sounds, Matt Patterson	Matt Patterson, "Summer Sounds"	Not mentioned	Both mentions retrieved; the reproduction drops both.
Single-hop	Why did John's teammates sign the basketball?	to show their friendship and appreciation	To show our friendship and appreciation	To celebrate his recovery.	Gold passage retrieved and quoted; the memory lacks the gift.

4.3 Ablation

Table 2 ablates each component RRC adds over the baseline pipeline (all cells LoCoMo). Swapping majority vote for union merging does not help at 3B (43.75 vs 43.39, within $\sigma = 0.45$ run-to-run noise), 7B (43.07, $\Delta = -0.15$), or API (51.91 vs 53.24). Replacing the gentle2 decline clause with verify-then-decline collapses 3B from 43.39 to 30.12 and API from 53.24 to 30.33: quiet failures arrive confident. Both match the two-regime picture: no gain where the estimator is quiet, large losses against confident Regime II errors (not run at 14B or 72B).

4.4 Parameter, Policy, and Case Analysis

We test the three parameters of RRC on a fixed stack (Table 12 in Appendix C).

Vote budget. On LoCoMo 3B, k=$1, 3, 5$ score 40.59, 40.26, and 40.09, below the noise floor ($\sigma = 0.45$, three seeds): the merged store is vote-neutral; votes do not denoise it directly.

Answer policy. The gentle policy scores 40.26 against 43.39 plain on LoCoMo 3B, and 37.22 against 43.22 on 7B. Under the official five-category average: the decline branch carries the headline at API (Table 10), so the headline runs the gentle policy at top-75 retrieval. The grounded policy is the largest upside: on 14B it scores 45.05, best overall there.

Retrieval depth. Widening retrieval to top-75 at 3B lowers LongMemEval F1 from 35.08 to 32.66: a quiet estimator means deeper retrieval only adds distractors, and the parameter pays only where it reports headroom (Table 9).

Case study. Table 3 contrasts RRC with our reproduction of Kumiho. Traces are verbatim from run logs, on questions RRC answers correctly and the reproduction does not.

5 Limitations

The analysis rests on a mechanism model with qualitative claims: a capability threshold exists, structure is lost in a fixed order, and the estimator tracks noise only. LongMemEval is scored with token-F1. Under the benchmark's official LLM-judge protocol the conclusions are unchanged (Appendix B.1).

6 Conclusion

This paper studies why long-term memory systems fail and separates the failures into two regimes: pass-specific extraction noise and representational loss in the store. Regime-Routed Construction (RRC) spends a few extra extraction passes on a measurement, the pass-disagreement estimator $\hat{q}$, and routes one of four configurations from that reading. The analysis proves the split is real: noise decays exponentially in the vote budget, error below a capability threshold cannot reach zero. On LoCoMo, RRC reaches 57.8/58.5 F1, $+4.7$ over the strongest published result. On LongMemEval, the gains are small and insignificant, consistent with a diagnosis of little headroom.

LLM USAGE STATEMENT

Large Language Models (LLMs) were used to polish the writing of the paper. All paragraphs were obtained after multiple rounds of human-LLM interaction, with the authors carefully reviewing, editing, and approving all content. The idea, experimental design, implementation, and analysis were entirely conducted by the authors.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *International Conference on Learning Representations (ICLR)*, 2025.

Dilip Arumugam and Benjamin Van Roy. Deciding what to learn: A rate-distortion approach. In *International Conference on Machine Learning (ICML)*, pp. 373–382, 2021.

Xinyun Chen, Renat Aksitov, Uri Alon, Jie Ren, Kefan Xiao, Pengcheng Yin, Sushant Prakash, Charles Sutton, Xuezhi Wang, and Denny Zhou. Universal self-consistency for large language model generation. *arXiv preprint arXiv:2311.17311*, 2023.

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

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006.

Pedro Domingos. A unified bias-variance decomposition and its applications. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 231–238, 2000.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(8):1605–1641, 2010.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630, 2024.

Stuart Geman, Élie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Hanqi Jiang, Junhao Chen, Yi Pan, Ling Chen, Weihang You, Yifan Zhou, Ruidong Zhang, Yohannes Abate, and Tianming Liu. SYNAPSE: Empowering LLM agents with episodic-semantic memory via spreading activation. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023.

Zhiyu Li et al. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*, 2025.

Liam Madden, Curtis Fox, and Christos Thrampoulidis. Next-token prediction capacity: General upper bounds and a lower bound for transformers. *IEEE Transactions on Information Theory*, 71 (9):7134–7148, 2025. doi: 10.1109/TIT.2025.3584013.

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 (ACL)*, pp. 13851–13870, 2024.

Ian R. McKenzie, Alexander Lyzhov, Michael Pieler, Alicia Parrish, Aaron Mueller, Ameya Prabhu, Euan McLean, Aaron Kirtland, Alexis Ross, Alisa Liu, Andrew Gritsevskiy, Daniel Wurgaft, Derik Kauffman, Gabriel Recchia, Jiacheng Liu, Joe Cavanagh, Max Weiss, Sicong Huang, The Floating Droid, Tom Tseng, Tomasz Korbak, Xudong Shen, Yuhui Zhang, Zhengping Zhou, Najoung Kim, Samuel R. Bowman, and Ethan Perez. Inverse scaling: When bigger isn't better. *Transactions on Machine Learning Research*, 2023.

Alliot Nagle, Adway Girish, Marco Bondaschi, Michael Gastpar, Ashok Vardhan Makkuva, and Hyeji Kim. Fundamental limits of prompt compression: A rate-distortion framework for black-box language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Young Bin Park. Graph-native cognitive memory for AI agents: Formal belief revision semantics for versioned memory architectures. *arXiv preprint arXiv:2603.17244*, 2026.

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

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Robert J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley, 1980.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Jason Wei et al. Inverse scaling can become u-shaped. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2023.

Halbert White. Maximum likelihood estimation of misspecified models. *Econometrica*, 50(1):1–25, 1982.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24 (359), 2023.

David P. Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends in Theoretical Computer Science*, 10(1–2):1–157, 2014.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. LongMemEval: Benchmarking chat assistants on long-term interactive memory. *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.10813.

Y. Wu, T. Lin, Y. Zhou, F. Zhang, Q. Guo, X. Zhou, S. Wang, X. Liu, Y. Ma, and Y. Fang. Memory in the LLM era: Modular architectures and strategies in a unified framework. *arXiv preprint arXiv:2604.01707*, 2026a.

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu, and Hongwei Wang. GAM: Hierarchical graph-based agentic memory for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2026b.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-MEM: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schütze, Volker Tresp, and Yunpu Ma. Memory-R1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 12805–12825, 2026.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 19724–19731, 2024. doi: 10.1609/aaai.v38i17.29946.

A PROOFS

This appendix proves the claims of §3.4 in the order they are used. The first group concerns Regime I. Proposition 1 decomposes construction error into a fluctuation term and a bias term. Lemma 1 and Proposition 2 bound the fluctuation term and show that it decays exponentially in the number of passes. Corollary 1 inverts the bound into a vote count. The second group concerns Regime II. Proposition 3 uses Fano's inequality to show that below a capability threshold the error is bounded away from zero for every decoder. Proposition 4 states what an answer policy can do in this regime. Proposition 5 orders the per-category losses. The third group concerns the estimator. Proposition 6 shows that $\hat{q}_k$ concentrates at rate $O(k^{-1/2})$ and is invariant to systematic error. Corollary 2 shows that routing mistakes are rare. The proofs assume independent extraction passes except where stated. The correlated-pass case is covered by Proposition 2.

Proposition 1 (two-term decomposition). Fix a node v with truth value t_v and write $A_k(v) = \sum_{i=1}^{k} X_i(v)$, which is $\mathrm{Binomial}(k, p_v)$ under independent passes. Majority voting emits the sign of $A_k(v) - k/2$. If $\mathrm{sign}(p_v - \frac{1}{2}) = 2t_v - 1$, an error is the event $\bar{X} = A_k(v)/k \le \frac{1}{2}$, and Hoeffding's inequality (Hoeffding, 1963) at distance $p_v - \frac{1}{2}$ gives

$$\Pr(\text{error}) \;\le\; \exp\big(-2k\,(p_v - \tfrac{1}{2})^2\big) \;=\; \exp\big(-k\,(\tfrac{1}{2} - q_v)\big),$$

using $q_v = 2p_v(1-p_v)$, so that $2(p_v - \frac{1}{2})^2 = \frac{1}{2} - q_v$. If $\mathrm{sign}(p_v - \frac{1}{2}) \neq 2t_v - 1$, a correct output requires $\bar{X} \ge \frac{1}{2}$, and the same tail bound applied to the wrong side gives

$$\Pr(\text{correct}) \;\le\; \exp\big(-2k\,(\tfrac{1}{2} - p_v)^2\big) \;=\; \exp\big(-k\,(\tfrac{1}{2} - q_v)\big),$$

hence $\Pr(\text{error}) \ge 1 - \exp(-k(\frac{1}{2} - q_v))$, converging to systematic error. Summing the per-node bounds over nodes decomposes construction error into the fluctuation term, controlled by q_v and reducible in k, and the bias term, which is fixed by p_v and hence by $\mathbb{E}[X]$ alone. □

Lemma 1 (majority recall bound). Let $X = \sum_{i=1}^{k} Z_i$ with $Z_i \sim \mathrm{Bernoulli}(p_f)$ independent. Majority omission is the event $X \le k/2$, equivalently $\bar{Z} = X/k \le 1/2$. Since $\mathbb{E}[\bar{Z}] = p_f$, Hoeffding's inequality (Hoeffding, 1963) gives

$$\Pr\big(\bar{Z} \le \tfrac{1}{2}\big) \;=\; \Pr\big(\bar{Z} - p_f \le \tfrac{1}{2} - p_f\big) \;\le\; \exp\big(-2k\,(p_f - \tfrac{1}{2})^2\big) \;=\; \exp\big(-k(2p_f - 1)^2/2\big).$$

Proposition 2 (noise is reducible). The independent case is Lemma 1. For correlated passes, let Z_i have mean p_f and pairwise correlation ρ. Then

$$\mathrm{Var}(\bar{Z}) = \frac{p_f(1-p_f)}{k}\big(1 + (k-1)\rho\big).$$

Define $k_{\text{eff}} = k/(1 + (k-1)\rho)$. The standardized gap between p_f and $1/2$ is $(p_f - 1/2)/\sqrt{p_f(1-p_f)/k_{\text{eff}}}$. A variance-matched Gaussian tail gives $\exp\big(-k_{\text{eff}}\,(p_f - 1/2)^2/(2p_f(1-p_f))\big)$. Since $p_f(1-p_f) \le 1/4$, this is at most $\exp(-k_{\text{eff}}(2p_f - 1)^2/2)$, which is the stated bound. It is exact when the standardized sum is Gaussian with equicorrelation ρ, and a variance-matched proxy otherwise. □

Proposition 3 (a threshold exists). Let $\delta : \mathcal{M} \to \mathcal{A}$ be any decoder from memory to answers, where $\mathcal{A}$ is the answer space of depth-d questions. Let $\{H_1, \ldots, H_{N_d}\}$ be the distinguishing family of Assumption 3, with answers pairwise separated under the task metric. Consider the Markov chain $H \to M \to \hat{A}$. Fano's inequality (Cover & Thomas, 2006) gives

$$\Pr(\hat{A} \neq A) \;\ge\; 1 - \frac{I(H;M) + \log 2}{\log N_d}.$$

Assumption 4 bounds $I(H;M) \le \kappa r$ with $r = g(C)$. Assumption 3 sets $\log N_d = s_d$. Substitution gives equation 1. The bound is strictly positive whenever $\kappa r + \log 2 < s_d$, equivalently $r < r^*(d) = (s_d - \log 2)/\kappa$. Since g is continuous and strictly increasing (Assumption 2), the preimage $C^*(d) = g^{-1}(r^*(d))$ exists and the irreducibility regime holds for all $C < C^*(d)$. □

Proposition 4 (declining trades coverage for risk). Partition the depth-d questions into Q_{carry}, whose required content the store carries, and Q_{miss}, whose content it does not, with $|Q_{\text{carry}}|/|Q| = \mu$. On Q_{miss} the hypothesis $r < r^*(d)$ puts us in the regime of Proposition 3: every answering scheme, the plain policy included, errs with probability at least $1 - (\kappa r + \log 2)/s_d$, and this mass is unavoidable. The gentle2 policy declines exactly Q_{miss}, since its decline condition is the absence of retrieved support, so its coverage is μ and its risk, conditioned on answering, is the error on Q_{carry}. On Q_{carry} the required content reaches the answer stage, and the residual error is the pass-specific noise of extraction and the selection noise of retrieval over carried content; the first is Regime I and

Table 4: Configurations by diagnostic signature. All five signatures are occupied by real backbones in §4.2.

Diagnostic signature	Dominant regime	Configuration
noisy $\hat{q}$, failing	Regime I (noise)	spend votes until $\hat{q}$ interval tightens
noisy $\hat{q}$, succeeding	Regime I, already quiet	minimal config, estimator monitoring only
quiet $\hat{q}$, failing	Regime II (bottleneck)	deepen retrieval, plain policy
quiet $\hat{q}$, succeeding	neither active	minimal config (single pass, shallow retrieval, plain)
below $C^*(d)$, any	Regime II below threshold	gentle2 decline

is suppressed by Proposition 2, and the second is addressed by retrieval depth. The plain policy's risk mixes in the Q_{miss} mass at the Fano floor, which is the trade. □

Proposition 5 (what dies first). Decompose the information content as $I(H) = U_1 \oplus U_2 \oplus U_{\text{rem}}$, where U_1 is the span of order-1 facts with $\dim U_1 = s^{(1)}$, U_2 is the span of order-2 relational bridges with $\dim U_2 = s^{(2)} - s^{(1)}$, and U_{rem} is the remainder. A rank-r sketch with $r \geq r_1^*$ carries U_1 in full, so order-1 recall is limited only by Regime I noise, which Proposition 2 suppresses at rate $\exp(-k(2p_f - 1)^2/2)$. For a backbone with $r < r_2^*$, apply Proposition 3 with s_d replaced by $s^{(2)}$. The Fano bound is strictly positive on the order-2 component. Mapping categories to structural orders, single-hop questions read order-1 facts and multi-hop and temporal questions read order-2 bridges. Per-category losses therefore order as $\Delta_{\text{SH}} \approx 0$ ahead of $\Delta_{\text{MH}}, \Delta_{\text{Temp}} < 0$ throughout the intermediate capability range $r \in [r_1^*, r_2^*)$. □

Proposition 6 (the estimator separates the regimes). The statistic $\hat{q}_k$ is an order-two U-statistic with kernel $h(i, j) = d(S^i, S^j)$. Under Assumption 1 and bounded second moments of h, the U-statistic law of large numbers (Hoeffding, 1948) gives $\hat{q}_k \to \bar{q}_C$ almost surely, and the U-statistic central limit theorem (Serfling, 1980) gives a Gaussian limit with variance $O(1/k)$. The bootstrap width of a U-statistic at level k is therefore $O(k^{-1/2})$. For the kernel as implemented, d is computed from the two vote realizations alone and the truth values t_v never enter it, so $\bar{q}_C$ is a functional of the vote law only and is invariant to systematic error: Regime II changes which nodes are extracted, not the value of the estimator, provided the per-pass vote probabilities are matched. In the sketch model of §3.1 with a norm-induced d, the same invariance reads as cancellation of the shared systematic part $\mu_C(H) = \Pi_{g(C)}(I(H))$ in pairwise differences, since $d(S_C^i, S_C^j) = d(\varepsilon_i, \varepsilon_j)$. □

Corollary 1 (votes needed). Set the failure bound of Proposition 2 at δ: $\exp(-k(2p_f - 1)^2/2) \leq \delta$. Taking logs, $k \geq 2\log(1/\delta)/(2p_f - 1)^2$. □

Corollary 2 (routing is stable). Let B be the nearest routing boundary to $\bar{q}_C$, at distance Δ, and assume first that B is an estimator-axis boundary, since the capability coordinate is read exactly. A wrong path requires $\hat{q}_k$ and $\bar{q}_C$ to fall on opposite sides of B, hence $|\hat{q}_k - \bar{q}_C| \geq \Delta$. Proposition 6 gives $\mathbb{E}[(\hat{q}_k - \bar{q}_C)^2] = O(1/k)$ under the bounded-second-moment condition, so Chebyshev's inequality yields $\Pr(\text{wrong path}) \leq O(1/(k\Delta^2))$. □

B Implementation details

Extraction. Windows of W turns are extracted with an instruction-tuned schema covering entities, attributes, events, and temporal links. Passes differ by sampling temperature. Majority voting keeps nodes and edges supported by at least $\lceil (k+1)/2 \rceil$ passes. Vote logs record per-pass node sets. The estimator $\hat{q}$ is the mean, over all windows of a run and all $\binom{k}{2}$ pass pairs within a window, of the Jaccard distance between the two passes' node-name sets: one minus the intersection size over the union size, with an empty pair scored as zero disagreement.

Consolidation. The full hyperparameter set for every reported configuration is generated by Algorithm 1 step 3 and is listed in Table 5. Consolidation runs every N windows. Duplicate merging uses cosine threshold τ_{dup}. Activation follows an EMA with coefficient α. Dormancy prunes nodes whose activation falls below ϵ for W_{dorm} consecutive windows. Abstraction and temporal edge weights are a λ triple $(0.55, 0.30, 0.15)$ for the winner configuration.

Retrieval. Spreading activation propagates over T steps with spreading coefficient S, decay δ, inhibition strength β applied over the top B activated nodes, and steepness γ around threshold θ. Anchors select the top lexical and semantic neighborhoods. Retrieval returns the top-K nodes, with $K = 30, 50, 75$ across configurations.

Serving. Local backbones are served with vLLM. The serve environment and the driver environment are separated. A separate embedding service on the same GPU serves the dense encoders. Context-

Algorithm 1 Regime-Routed Construction

Require: dialogue history H, chunked into turn windows $w \in \mathcal{W}$; backbone with capability index C; threshold surface $C^*(d)$ given by Proposition 3; reasoning-depth coordinate d per question class, read from the taxonomy or measured from evidence annotations (Appendix B.1); probe slice $\mathcal{W}_{\text{probe}} \subseteq \mathcal{W}$

Ensure: per-question answers under the routed configuration

Stage 1 — Probe (measure the estimator).

for $w \in \mathcal{W}_{\text{probe}}$ **do**
  run the default probe of k_0 independent extraction passes on w, yielding vote fields $X_1, \ldots, X_{k_0}$ over its candidate nodes
  merge by majority vote, keeping node or edge v with $\sum_{i=1}^{k_0} \mathbf{1}\{v \in X_i\} > k_0/2$
  append the per-pass node sets to the vote log
end for
compute the estimator $\hat{q}$ as the mean pairwise Jaccard distance over pass pairs and windows, with a bootstrap confidence interval of width w_q ▷ Regime I strength

Stage 2 — Diagnose (select the path).

$\sigma \leftarrow$ path selected by R from $(\hat{q}, C, d)$, *decline* if $C < C^*(d)$, *noise* if the $\hat{q}$ interval lies above the quiet cut, *headroom* if below it and $C \geq C^*(d)$, else *default* ▷ no outcome is read

Stage 3 — Route (prescribe the configuration).

$(\text{vote budget } k,\ \text{retrieval depth } (K, T),\ \text{policy } \pi) \leftarrow R(\sigma)$ ▷ the paths of Table 4, votes under noise, depth under headroom, decline below $C^*(d)$

Stage 4 — Run (configured pipeline).

The base store runs unchanged; routing never edits the memory it produces.

for $w \in \mathcal{W}$ in stream order **do**
  consolidate and index w with the base store under the routed vote budget k
end for
for each question q **do**
  retrieve by spreading activation under the routed retrieval depth (K, T)
  answer q under the routed policy π, declining when the support is missing
end for

length fallbacks follow a 32768, 16384, 8192 ladder with a per-question trim-and-retry guard. The 14B backbone is served at int4 quantization with official AWQ weights on a single 24-GB card.

Table 5: Hyperparameters. One row per Algorithm 1 configuration class. Values are the winner-configuration defaults. Full per-run values are checkpointed with each run.

Stage	Parameters
Extraction	W=5 turns, k=3 passes, temperature 0.1 (votes 0.7), schema v1
Consolidation	N=5, τ_{dup}=0.92, α=0.5, ϵ=0.01, W_{dorm}=10, λ=(0.55, 0.30, 0.15)
Retrieval	S=0.8, T=3, δ=0.5, β=0.15, B=7, γ=5.0, θ=0.5, anchors 15+15, K=50
Answer policy	plain or gentle2, dual-mode map if routed

B.1 Evaluation protocols

LoCoMo answer keys. We report numbers against the official released answer keys without modification, matching the protocol of the published systems we compare against. Every gain we report is a paired contrast between two configurations scored under the identical key set, so key noise acts on both arms equally and the bootstrap difference is robust to it.

Per-conversation analysis. LoCoMo's 1,986 questions are grouped by their source conversation (10 in total), F1 is computed within each conversation, and the per-conversation gain is the RRC-versus-baseline F1 difference inside that conversation. Categories are per-question labels that cross-cut conversations, so the per-conversation and per-category views are orthogonal.

LongMemEval protocols. The benchmark's official metric is LLM-judge accuracy; we report weighted micro-F1 over answer strings, with exact match alongside, and the 20 adversarial questions are excluded from the F1 and reported separately, following the benchmark's convention. Under the official judge prompts with a DeepSeek-V3 judge, the API winner and its baseline are statistically indistinguishable (66.7 vs 67.1 on the 480 non-adversarial questions, 66.0 tied with adversarial in-

cluded, and a Qwen2.5-72B judge ties them at 61.8), and the capability ladder is intact (3B 39.6, 7B 51.3), so the token-F1 conclusions replicate under the official protocol. Judge scores are not comparable across judges, and we do not rank against leaderboard entries scored by other judges.

Row provenance. Published values are overall token F1 as reported by each paper, except that GAM's and Synapse's values are per-category averages excluding the adversarial category. The A-Mem, MemoryBank, and MemoryOS rows at 1.5B and 3B are Synapse's cross-backbone reruns under its unified protocol. Memory-R1's LongMemEval values (†) use its own subset and scoring, and ‡ marks reinforcement-learning training. Kumiho's headline folds a 97.5 adversarial-refusal score into the average (pinned reproducible value 53.1); its per-category row is our end-to-end reproduction with DeepSeek-V3. Zep and MemOS report LongMemEval only under the official judge (71.2 and 77.8) and are not ranked there. The Llama rows, the bge-base upgrade at LLaMA-3.1-8B, and the int4 14B pair are ours (Appendix C.1, §4.4); LongMemEval was not run on the Llama backbones.

Category mapping. LoCoMo's five official categories map to reasoning depth d as follows. Single-hop is order-1. Multi-hop and temporal are order-2. Open-domain and adversarial are reported separately. LongMemEval categories map identically, with knowledge-update treated as order-2 following its session-spanning structure.

Measuring the depth coordinate. The depth coordinate d is a structural property of a question's evidence, not a fitted parameter, and it can be measured without answer labels. Table 6 reports the session span of LoCoMo's gold evidence annotations: the measured ordering reproduces the taxonomy (single-hop $\approx$ open-domain $<$ multi-hop $<$ temporal $<$ adversarial), so d can be assigned from evidence spans on benchmarks that carry annotations but no published taxonomy.

Question class	Mean sessions	Mean turns	Multi-session
single-hop	1.00	1.07	0%
open-domain	1.00	1.03	0%
multi-hop	1.10	1.15	9%
temporal	1.73	1.93	34%
adversarial	2.68	2.81	95%

Table 6: Structural depth measured from evidence annotations on LoCoMo. Session span is the number of distinct dialogue sessions touched by the gold evidence of a question. The ordering reproduces the taxonomy ordering single-hop $\approx$ open-domain $<$ multi-hop $<$ temporal $<$ adversarial without reading any answer label.

B.2 Prompt templates

All stage prompts are reproduced verbatim. Placeholders in typewriter font are filled at run time. The retrieved-context framing replaces a full-history dump and is the method under test, while every instruction sentence is kept unchanged from the official benchmark prompts.

Extraction stage. One system message and one user template are used for every backbone and every pass. Passes differ only in sampling temperature.

```
SYSTEM: You are an expert knowledge engineer building a semantic graph
from conversation history. Your goal is to consolidate episodic details
into structured knowledge nodes.

USER: Input Context: {n} recent conversation turns.

--- CONVERSATION WINDOW ---
{window}
--- END WINDOW ---

Existing anchor concepts already in the graph (may be empty):
{anchors}

Reasoning (Chain of Thought):
1. Analyze: Identify new facts not present in previous context.
2. Classify: Categorize facts into Identity, Preference, Event,
   or Technical.
3. Extract: Form canonical node names (e.g., "likes camping" ->
   "Camping Preference").
```

```
Task 1: Node Extraction. Produce a list of concept nodes. Each node has:
  "name" (canonical concept name, a short noun phrase),
  "type" (one of Identity | Preference | Event | Technical),
  "confidence" (0.0-1.0).

Task 2: Edge Formation. Link new nodes to existing anchors or to each
other. Each edge has "src", "rel" (a short UPPER_CASE relation), "tgt",
"w" (0.0-1.0).

Respond with ONLY a JSON object of the exact form:
{"nodes": [{"name": "...", "type": "...",
            "confidence": 0.9}],
 "edges": [{"src": "...", "rel": "...", "tgt": "...", "w": 0.8}]}
```

Answer stage. All policies share one system message and one base instruction, the official LoCoMo QA instruction used verbatim by the published baselines. Each policy adds the temporal-grounding clause, a decline clause, both, or neither (Table 7), and every policy ends with the line `Question: {question} Short answer:`.

```
SYSTEM: You are a helpful assistant answering questions about a
conversation between {sa} and {sb} based on retrieved snippets of that
conversation. Each snippet is prefixed with the date on which it was said.

BASE: Based on the above context, write an answer in the form of a short
phrase for the following question. Answer with exact words from the
context whenever possible.

TEMPORAL-GROUNDING (grounded, gentle, gentle2, verify): When a question
asks about time, resolve relative time expressions in the context (such
as 'yesterday' or 'last week') against the date shown on that snippet,
and answer with the resulting date or duration.

GENERIC DECLINE (gentle): If the context does not contain the
information needed to answer the question, output 'Not mentioned'.

ANTI-OVER-DECLINE (gentle2): If the context truly does not mention the
relevant information, output 'Not mentioned'. However, if the context
mentions the relevant facts -- even if answering requires simple
counting, arithmetic, dates/durations, or combining two or three
snippets -- do NOT decline: answer the question.

EXPLICIT-MENTION (verify): Is this EXPLICITLY mentioned in the context?
If not, output 'Not mentioned'.
```

Table 7: Policy composition. Each policy is the base instruction plus the listed clauses, in the order shown, followed by the question line.

Policy	Temporal grounding	Decline clause
plain	–	–
grounded	yes	–
gentle	yes	generic
gentle2	yes	anti-over-decline
verify	yes	explicit-mention

B.3 Cost details

Table 8 reports LLM calls and token volumes for representative runs. Extraction-stage calls dominate the budget, and RRC multiplies them by the vote budget and the probe (roughly fourfold on the API LongMemEval run), while the answer stage stays cheap because answers are formed from retrieved subgraphs rather than full histories (9.5k completion tokens on the API LoCoMo run). An early extraction-cache defect once inflated API call counts by roughly $2.7\times$. The corrected counts are reported here.

Table 8: Compute per run. LLM calls and token volumes for representative runs.

Run	Backbone	LLM calls	Prompt tok	Completion tok
LME baseline	7B	58,196	96.8M	29.3M
LME baseline	API	47,397	80.2M	29.8M
LME winner	API	191,863	317.5M	116.0M
LoCoMo winner	API	1,974	6.5M	9.5k

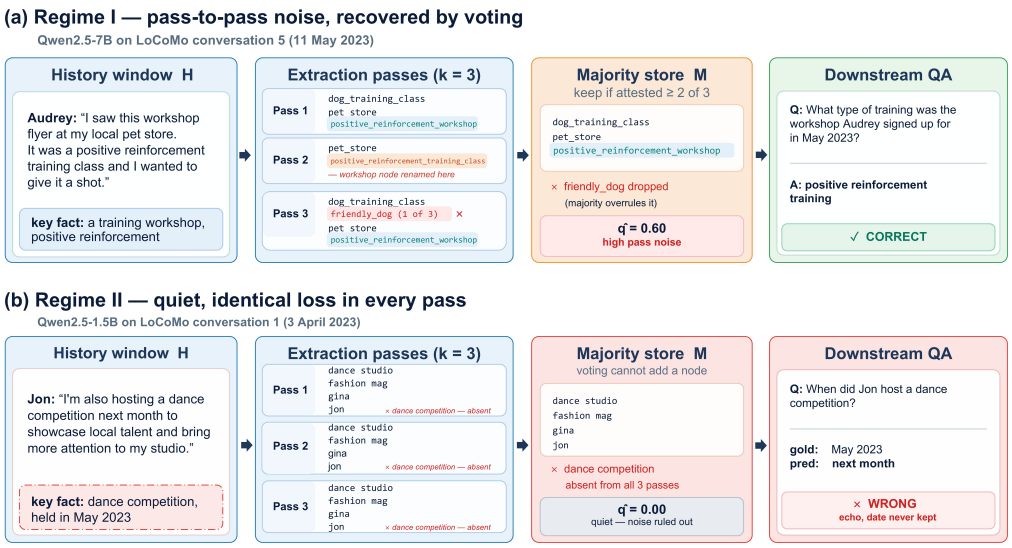


Figure 2: The two regimes of forgetting on real LoCoMo history windows. (a) Regime I at Qwen2.5-7B: the three passes disagree ($\hat{q} = 0.60$), one renaming the workshop node and one inventing a node, and majority vote keeps the workshop node while voting the invention down. (b) Regime II at Qwen2.5-1.5B: the three passes return identical node sets ($\hat{q} = 0.00$) and the dance-competition fact is absent from every pass, so voting has nothing to recover. The same window consolidated by the 7B backbone extracts the missing node in one pass ($\hat{q} = 0.44$).

C Additional Results

This appendix reports the experiments behind the claims of §4.2: the LongMemEval transfer panel, mechanism checks on the quiet cells, per-category scores on LoCoMo, the estimator readings and run variance, a cross-family replication, and the full policy ladder. Each experiment states its purpose first, then its setup, then its result.

Mechanism checks. We run two checks to localize the quiet losses before attributing them to a regime. The first is forensic. For the 7B single-hop loss we inspect the failing questions. Twelve of 150 questions share one pattern. Confusable alternatives recorded in the same dialogue are surfaced by top-50 retrieval against the baseline's top-30, and 14–25 of 50 passages are shared. These errors live at retrieval selection, not consolidation. The second varies retrieval depth. The 3B LongMemEval cell scores 32.66 under top-75 retrieval against the 35.08 baseline. Deep retrieval does not rescue the quiet cells, so the losses are neither pass noise nor a retrieval-depth problem. Figure 3 places the eight backbone-benchmark pairs on the estimator-outcome plane. The quiet-but-failing points sit in the lower-left quadrant, the representational-bottleneck signature of §3.4.

Per-category scores. We report per-category F1 under the official category mapping so that gains and losses can be read category by category (Table 10). The table also carries the below-threshold pair. The configuration that gains $+3.79$ at 3B and $+5.29$ at 7B collapses at Qwen2.5-1.5B, from 37.98 down to 32.81 (-5.09 $[-6.76, -3.46]$, $P(\Delta \leq 0) = 1.0$), with a normal $\hat{q}$ reading, the quiet-but-failing signature of Proposition 6, and is discussed in §4.2. Even in collapse, the RRC row remains above every published system at this backbone except Synapse, and the paired baseline itself tops all of them.

Table 9: LongMemEval transfer panel, baseline F1 repeated from Table 1 where available. Δ, CI, and P are the paired bootstrap on per-question token-F1 differences (10,000 resamples, two-sided P). The 14B row is the mid-ladder replication at int4 quantization, run at top-30 retrieval after the top-50 cell degraded the answer stage under int4 serving, a serving artifact rather than a theory test.

Backbone	Baseline F1	Δ [95% CI]	P	$\hat{q}$	Diagnostic signature
1.5B	30.33	-0.42 $[-2.84, +1.94]$	0.72	0.754	noisy, borderline
3B	35.08	-1.18 $[-4.16, +1.71]$	0.42	0.692	quiet, failing, Regime II
7B	39.61	-0.98 $[-3.40, +1.47]$	0.45	0.518	quiet, failing, Regime II
14B	38.29	-0.28 $[-2.07, +1.53]$	0.77	0.638	quiet, null transfer
API	48.81	$+1.37$ $[-0.63, +3.36]$	0.18	0.689	quiet, succeeding

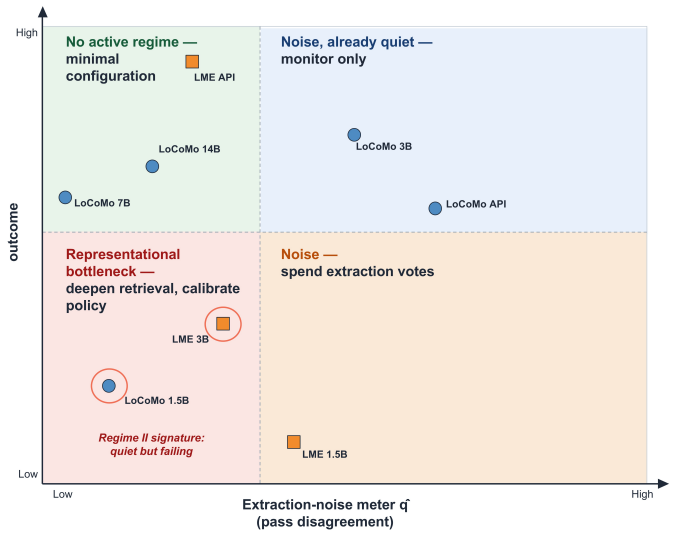


Figure 3: The diagnostic plane. The extraction-noise estimator against outcome for eight backbone-benchmark pairs, circles for LoCoMo and squares for LongMemEval, with the quiet-but-failing points circled. Only the lower-left quadrant diagnoses a representational bottleneck, read by exclusion.

Estimator readings and run variance. We report $\hat{q}$ at every backbone and benchmark to check that the estimator reads distinguishably across regimes (Table 11). The statistic is the mean pairwise Jaccard distance between per-pass node-name sets over $k = 3$ passes. The held-out LongMemEval 7B reading is recovered by deterministic replay of the run's cached vote passes. The Llama reading comes from the cross-family replication of Appendix C.1, and the 14B reading from the int4 mid-ladder replication of §4.2. Run-to-run variance sets the noise floor for interpreting differences between configurations. The 3B headline varies by $\sigma = 0.45$ across three runs, the API headline by $\sigma = 0.78$ across two, and 72B exhibits the largest variance in the panel, flagged in §5.

C.1 Replication across model families

The same configuration runs on two Llama backbones. The purpose is to test whether the capability ladder is a Qwen artifact. On Llama-3.1-8B the baseline 36.75 sits just below the Qwen 3B/7B baselines (39.57/37.85) and the winner 43.60 converges onto the Qwen plateau (43.39/43.22), with every non-adversarial category gaining (multi-hop $+8.18$, single-hop $+7.53$, open-domain $+6.74$, temporal $+3.96$). A bge-base retriever upgrade lifts the winner to 46.68. On Llama-3.2-3B the plain-policy gain is $+0.08$, and the grounded policy (§4.4, the plain policy with the temporal-grounding instruction restored) moves the identical pair to $+3.18$, placing the zero crossing between the 1.5B collapse and the Qwen-3B gain, as on the Qwen ladder.

C.2 Parameter scan

Table 12 scans one parameter at a time on a fixed stack, F1 on the named benchmark. Vote-budget rows share the gentle policy and top-50 retrieval at 3B, and policy rows share one extraction stack. The grounded row is overall F1, its temporal category moving from 24.20 to 38.88.

Table 10: LoCoMo per-category F1 under the official category mapping. †Recipe rows average the four non-adversarial categories; headline rows use the five-category weighted average. The LLaMA-3.1-8B row runs the recipe's default retriever; the bge-base upgrade reported in Table 1 scores 46.68. Adversarial gold answers decline the question: the RRC recipe rows run the plain answer policy, which answers instead, so their adversarial scores sit near zero, while the headline rows run the gentle policy and score high. The recipe's plain policy trades temporal F1 for overall F1 at 7B–72B; the grounded policy recovers temporal at 14B (24.20 → 38.88). Row provenance follows Appendix B.1.

System	Backbone	Multi-Hop	Temporal	Open-Domain	Single-Hop	Adversarial	Avg†
GAM (Wu et al., 2026b)	GPT-4o-mini	35.88	51.96	28.12	56.58	–	43.14
Synapse (Jiang et al., 2026)	GPT-4o-mini	35.7	50.1	25.9	48.9	96.6	40.5
Kumiho (Park, 2026), reproduced	GPT-4o	35.47	53.27	29.03	46.23	97.53	56.53
A-Mem (Xu et al., 2025)	Qwen-1.5B	18.2	24.3	16.5	23.6	46.0	20.4
MemoryBank (Zhong et al., 2024)	Qwen-1.5B	11.1	4.5	8.1	13.4	36.8	10.0
MemoryOS (Li et al., 2025)	Qwen-3B	21.4	26.2	10.2	23.3	–	22.1
A-Mem (Xu et al., 2025)	Qwen-3B	12.6	27.6	7.1	17.2	27.9	16.2
Baseline pipeline	Qwen2.5-1.5B	26.31	34.99	11.09	46.11	16.82	37.98
RRC (ours)	Qwen2.5-1.5B	25.55	15.80	16.78	43.56	0.0	32.81
Baseline pipeline	Qwen2.5-3B	27.37	34.88	6.10	49.26	41.03	39.57
RRC (ours)	Qwen2.5-3B	34.32	34.44	12.40	53.38	1.12	43.39
Baseline pipeline	Qwen2.5-7B	21.47	34.49	7.43	48.11	87.89	37.85
RRC (ours)	Qwen2.5-7B	35.32	25.33	15.44	55.86	0.90	43.22
Baseline pipeline	Qwen2.5-14B	24.81	29.65	6.72	50.62	86.10	38.69
RRC (ours)	Qwen2.5-14B	34.13	24.20	12.55	56.52	0.90	42.83
RRC (ours), grounded policy	Qwen2.5-14B	32.90	**38.88**	10.84	55.38	0.67	**45.05**
Baseline pipeline	Llama-3.1-8B	28.89	14.59	6.12	51.55	73.32	36.75
RRC (ours)	Llama-3.1-8B	37.07	18.55	12.87	59.09	1.57	43.60
Baseline pipeline	Qwen2.5-72B	29.15	43.70	14.15	51.27	86.77	43.33
RRC (ours)	Qwen2.5-72B	34.35	32.82	16.35	55.15	2.24	44.27
Baseline pipeline	DeepSeek-V3	26.96	33.89	15.80	57.39	90.58	44.33
RRC (ours)	DeepSeek-V3	39.45	38.73	27.39	66.36	1.12	53.24
RRC (ours), headline	DeepSeek-V3	41.81	30.19	15.98	68.99	75.56	57.77
RRC (ours), headline, run 2	DeepSeek-V3	42.71	31.46	12.02	69.56	76.91	58.46

Table 11: Extraction-noise estimator $\hat{q}$. Mean pairwise Jaccard distance between per-pass node-name sets over $k = 3$ passes, with vote-log window counts. LongMemEval 7B is the held-out test, and Llama-3.1-8B is the cross-family replication of Appendix C.1, recovered by deterministic replay of the run's cached vote passes. 14B is the mid-ladder replication of §4.2, served at int4 quantization.

	LoCoMo		LongMemEval	
Backbone	$\hat{q}$	windows	$\hat{q}$	windows
1.5B	0.669	2,358	0.754	49,181
3B	0.773	3,504	0.692	61,596
7B	0.538	1,179	0.518	49,549
API	0.817	3,537	0.689	49,549
Llama-3.1-8B	0.787	1,179	–	–
14B	0.673	1,179	–	–

C.3 Policy ladder on LongMemEval

We test how each answer policy trades weighted F1 against adversarial F1 on LongMemEval. Table 13 reports the full policy ladder. The gentle2 policy converts adversarial F1 from 0 to 95–100 on 3B at roughly half of weighted F1.

Table 12: Parameter analysis, one factor at a time on a fixed stack.

Backbone	Benchmark	parameter	Setting	F1
3B	LoCoMo	vote budget	k=1	40.59
			k=3	40.26
			k=5	40.09
3B	LoCoMo	answer policy	plain	43.39
			gentle	40.26
7B	LoCoMo	answer policy	plain	43.22
			gentle	37.22
14B	LoCoMo	answer policy	grounded	45.05
3B	LongMemEval	retrieval depth	default	35.08
			top-75	32.66

Table 13: The policy ladder on LongMemEval. Weighted micro-F1 and adversarial F1. Table 9 reports the headline transfer on the per-question micro-F1 protocol.

Backbone	policy	F1w	Adv. F1
1.5B	plain (winner)	30.00	0
1.5B	gentle2	15.97	85
1.5B	top-75 gentle2	12.31	85
3B	plain (winner)	33.72	0
3B	gentle2	21.04	95
3B	top-75 gentle2	19.26	100
7B	plain (winner)	38.69	5
7B	gentle2	29.70	90
API	plain	50.18	0
API	gentle2	49.51	75