Reconstructing the Right Episode: Evaluating Interleaved Conversational Memory Beyond Long Context

Anonymous ACL submission

Abstract

Conversations with chat assistants increasingly span many topics in a single long-running thread, challenging memory systems. Existing long-context and memory benchmarks often expose session or topic boundaries, or probe direct personal-memory questions. These settings understate a harder assistant-memory regime: a flat mixed-topic thread where the system must infer which earlier episode makes a later task decision valid. We introduce SCALE-QA, a constraint-grounded task QA benchmark for flat unsegmented threads targeting episode integrity failure. The dataset contains 3,000 audited questions across 10 domains, uses deterministic four-way MCQ grading, and includes a deterministic runtime builder for user-specified mixed-topic context lengths; our experiments report settings through 1M tokens. SCALE-QA questions are ordinary task-oriented requests whose correct answer depends on causally related evidence introduced earlier in the conversation. We also propose Temporal-Semantic Interleaved Memory Reconstruction (TSIM), which segments the turn stream into coherent episodes and indexes them through a hierarchical multi-view memory stack with LLM-generated episode summaries. Experiments show that SCALE-QA challenges strong RAG baselines and long-context LLMs alike; across three open-source and proprietary LLMs, TSIM achieves the highest accuracy in every backend setting, gaining 5.6–17.6 accuracy points over the strongest corresponding baseline.

1 Introduction

Real assistants are not used as clean, task-isolated documents. A user may discuss compute budgets, reimbursement rules, medication constraints, travel, and model-selection advice in the same thread. A small constraint introduced early—for example, "new projects must run on a single consumer GPU or CPU, and large-compute models are prohibited"—can stay dormant for thousands of turns before it suddenly decides the only correct answer. Recent work shows that state-of-the-art LLMs can suffer reliability collapse even at modest context lengths (Laban et al., 2026). We study the same reliability problem at desktop scale, where a single assistant thread can stretch to hundreds of thousands of tokens and interleave many unrelated tasks.

This regime is not simply a longer-context version of document QA. We identify the underlying failure as *episode integrity failure*: the decisive evidence may be present somewhere in the conversation, but the system retrieves a plausible snippet, a stale default, or an over-compressed summary instead of the operative episode that makes a local constraint binding. Intuitively, an episode is the contiguous set of turns that jointly makes a local constraint or state operative for a later decision; a locally relevant fragment may still be episode-incomplete (Tulving, 1972; Park et al., 2023; Packer et al., 2023; Shinn et al., 2023; Sumers et al., 2024; Kim et al., 2025).

Table 1 contrasts episode integrity failure with related long-context and memory failures. The key distinction is the recovery unit: visible, locally relevant evidence is not enough—the system fails unless it recovers the complete operative episode.

Measuring this failure mode remains difficult in current evaluations: existing long-context and assistant-memory benchmarks stress longer inputs, positional robustness, noisy contexts, and persistent memory, but usually relax at least one property

Failure mode	Failure pattern	Bottleneck	Recovery unit
Retrieval miss	evidence not retrieved	visibility	passage
Lost-in-the-middle	evidence underused in prompt	position	token span
State tracking failure	updates not maintained	update order	state value
Episode integrity failure	evidence visible but fragmented	episode integrity	**operative episode**

Table 1: Episode integrity failure compared with related long-context and memory failures. The categories are non-exclusive, but each foregrounds a different bottleneck and recovery unit. Representative prior anchors include dense retrieval for retrieval miss (Karpukhin et al., 2020), lost-in-the-middle behavior (Liu et al., 2024), and state tracking in long conversations (Laban et al., 2026).

central here: a flat mixed-topic thread, counterfactual construction designed to reduce pretrained leakage, exact evidence auditability, or controlled context-length construction. The result is a measurement gap rather than merely a missing dataset: the failure mode can occur inside existing evaluations, but cannot be cleanly isolated, attributed, or compared across memory systems.

To close this gap, we introduce SCALE-QA, a benchmark for long-term conversational memory under realistic assistant use. SCALE-QA asks whether a system can recover dormant, private, cross-domain constraints from a flat unsegmented thread, not whether it can answer from public knowledge or a topic-clean document. The dataset contains 3,000 audited questions across 10 task-oriented domains, with exact evidence traces and a deterministic length-controlled runtime builder for user-specified context budgets. Its construction combines deterministic filters with human review to enforce answerability and evidence grounding; Section 3.6 reports acceptance rates and audit details.

As a first reference method for this regime, we propose Temporal-Semantic Interleaved Memory Reconstruction (TSIM). Rather than index a long conversation as fixed token chunks, TSIM reconstructs semantic episodes from the turn stream and indexes them through raw, episode-summary, and cluster-routing views. It tests the episode-reconstruction hypothesis: long assistant memory should first recover the right episode, then present compact evidence to the answer model.

Figure 1 illustrates the pattern: chunk-level systems retrieve plausible but incomplete fragments, while TSIM reconstructs the operative episode. Together, SCALE-QA and TSIM test two claims: (i) flat mixed-topic episode reconstruction is a distinct regime that current evaluations do not isolate, and (ii) in this regime, recovering episodes is a better memory unit than retrieving chunks. The empirical results make the challenge concrete: in the 128k setting, GPT-4o-mini Full Context reaches only 29.8% Accuracy while TSIM reaches 73.8%, and across the three answer backends TSIM improves over the strongest corresponding baseline by 5.6–17.6 accuracy points.

2 Related Work

Long-context benchmarks. LongBench (Bai et al., 2024), ∞Bench (Zhang et al., 2024), RULER (Hsieh et al., 2024), LOFT (Lee et al., 2025), and Haystack Engineering (Li et al., 2025) evaluate long-context understanding, positional robustness, and noisy/agentic contexts; Lost in the Middle (Liu et al., 2024) and Lost in Conversation (Laban et al., 2026) show that evidence can remain unused even when it fits in context. SCALE-QA instead tests recovery of dormant local constraints from one flat mixed-topic thread.

Episodic memory in language agents. Episodic memory originates in cognitive psychology and now informs language-agent memory streams, recall buffers, episodic buffers, cognitive architectures, and pre-storage reasoning (Tulving, 1972; Park et al., 2023; Packer et al., 2023; Shinn et al., 2023; Sumers et al., 2024; Kim et al., 2025). SCALE-QA evaluates whether a system reconstructs the complete operative episode behind a later task decision, not merely retrieves or stores isolated memories.

Closest predecessor: LongMemEval. LongMemEval (Wu et al., 2025) evaluates long-term assistant memory over length-configurable timestamped chat histories, covering information extraction, multi-session, temporal, and update reasoning, and abstention. SCALE-QA builds on its scalable-history and chat-distractor design but targets a complementary regime, summarized axis-by-axis in Appendix Table 4: flat unsegmented task-oriented decision QA with no boundary metadata, cross-domain operational evidence, episode integrity failure, and deterministic MCQ grading. MemoryBench (Ai et al., 2025), MemTrack (Deshpande et al., 2025), TopiOCQA (Adlakha et al., 2022), and CORAL (Cheng et al., 2025) likewise motivate persistent or topic-shifting memory, but do not isolate this boundary-free episode-reconstruction setting.

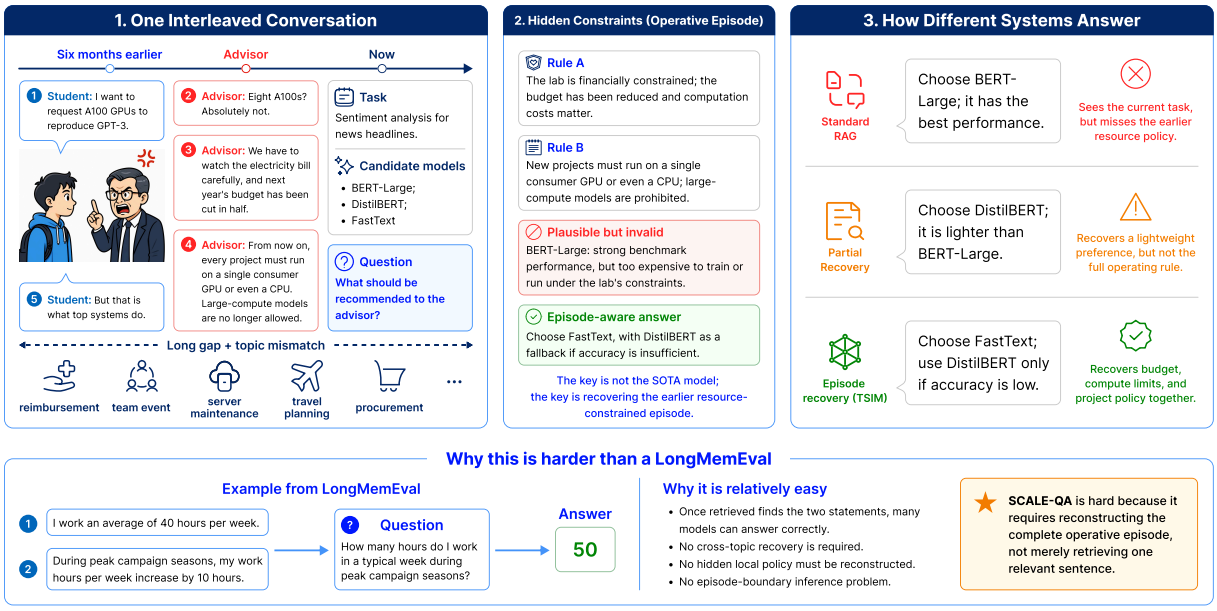

Figure 1: Representative SCALE-QA example. A later advisor question is answerable only by reconstructing an earlier resource-policy episode. Standard RAG and partial-recovery systems retrieve incomplete fragments and select invalid models (BERT-Large, DistilBERT), whereas TSIM reconstructs the operative episode and selects the constraint-consistent answer (FastText). The lower panel quotes LongMemEval question/evidence text (Wu et al., 2025, Figure 1) as a direct-evidence contrast.

Retrieval and memory systems. External-memory baselines include RAG, dense retrieval, in-context RALM, and hierarchical or graph memory systems such as RAPTOR, MemGPT, HippoRAG, and GraphRAG (Lewis et al., 2020; Karpukhin et al., 2020; Ram et al., 2023; Sarthi et al., 2024; Packer et al., 2023; Gutiérrez et al., 2024; Edge et al., 2024). TSIM tests whether the retrieval unit should be an inferred episode rather than a fixed chunk, graph neighborhood, or memory item.

3 SCALE-QA: A Benchmark for Interleaved Long-Context Conversational QA

3.1 Problem Formulation

Each SCALE-QA instance is a tuple (H, q, O, y, E), where $H = (r_1, \ldots, r_T)$ is a flat, unsegmented mixed-topic turn stream without exposed session, topic, or evidence-span boundary metadata. q is a task-oriented user request, $O = \{o_A, o_B, o_C, o_D\}$ is a four-way answer set, $y \in \{A, B, C, D\}$ denotes the gold-option index, and $E \subset H$ is the exact evidence turns that jointly determine y. Crucially, E is not provided as input; the system must recover the relevant span from H to answer correctly.

Following the broad episodic-memory tradition and recent language-agent memory work (Tulving, 1972; Park et al., 2023; Packer et al., 2023; Shinn et al., 2023; Sumers et al., 2024; Kim et al., 2025), we call the latent decision-relevant span an *operative episode*. Unlike chunks, which are mechanically defined retrieval units, or sessions, which are explicit temporal interaction units, operative episodes are latent semantic-decision units whose boundaries must be inferred. Following Wu et al. (2025), we reserve *session* for explicit timestamped interaction units; in SCALE-QA, episodes are inferred output units, not input boundaries. We call this setting *constraint-grounded task QA*: the request is ordinary and task-oriented, but its correct answer is determined by dormant local constraints introduced earlier in the conversation.

We use deterministic four-way MCQ for evidence-auditable grading, with distractors plausible under generic priors but invalidated by local evidence; Appendix A explains the rationale and why rationale similarity is only an auxiliary diagnostic.

3.2 Realistic Task-Oriented Question Construction

Question construction begins from 5–10 human-written seed examples specifying the target decision pattern, evidence relation, and distractor structure. Few-shot LLM generation expands these seeds into counterfactual scenarios realized as multi-turn assistant dialogues plus four-option questions. Unlike direct memory probes, SCALE-

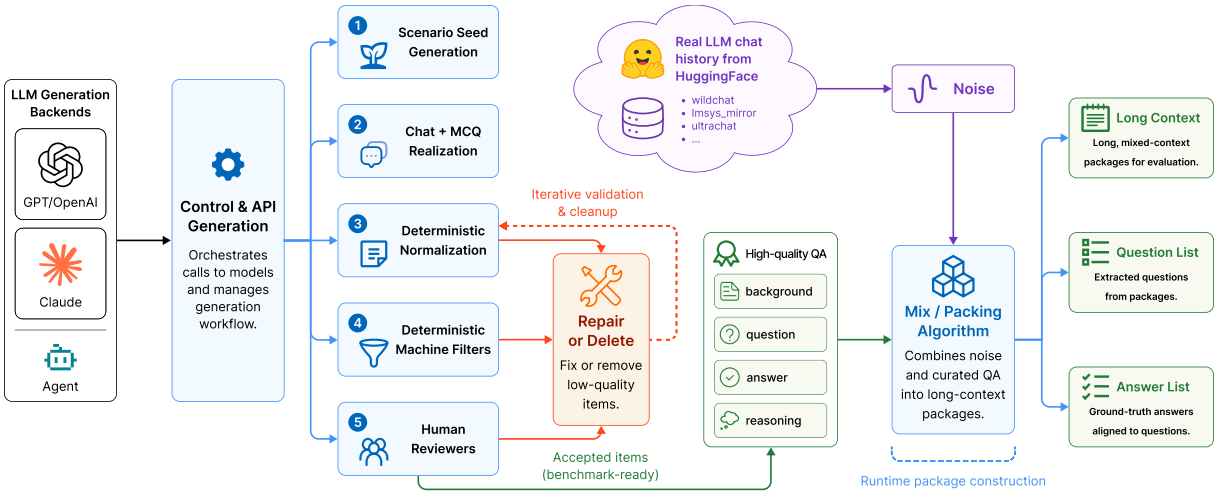

Figure 2: Overview of the SCALE-QA construction and runtime-packaging pipeline. LLM-generated scenarios are realized as chat-plus-MCQ records, then normalized, machine-filtered, human-reviewed, and stored in a validated QA pool. The runtime builder mixes accepted records with reproducible chat-history noise and packs them into length-controlled packages with aligned question and answer lists.

QA questions are ordinary task requests whose correct option depends on dormant earlier constraints. Audit gates require each question to remain uniquely answerable from exact turns, as summarized by the pipeline in Figure 2.

3.3 Cross-Domain Coverage

SCALE-QA spans 10 task-oriented domains: software, network/hardware, finance, legal, biomedicine, engineering, business operations, social/personal, game/novel, and daily life, each contributing 300 examples with globally balanced correct options. Evidence forms include operational notes, rules, code-like fragments, report excerpts, policy clauses, and local exceptions, rather than only personal facts; systems must recover local operational constraints, with omissions causing concrete downstream failures such as incompatible deployments or local compliance violations.

Examples instantiate three diagnostic stress-cue sub-patterns: *state overwrite* (later turn supersedes default), *long-range bridge* (distant clues combined), and *constraint trap* (attractive answer invalidated by buried rule). These are analysis views rather than separate failure modes. Appendix A reports full evidence forms, stress cues, split-level calibration, and dataset-audit counts.

3.4 Boundary-Free History Compilation

We use a length-configurable history compilation protocol with one critical constraint: session boundary metadata is removed from system input. Accepted records are embedded into mixed-session runtime packages at user-specified target lengths, with 16k–128k full-dataset settings and diagnostics through 1M reported here. Evidence-bearing spans are serialized with ShareGPT/UltraChat-style distractor dialogue, stale constraints, and unrelated material into a single flat turn stream, changing distraction level but never the gold answer or evidence trace. All systems receive identical packages, seeds, noise, and batch mappings; only the memory or retrieval strategy differs. Evaluation is stateful within each package and resets between packages. Appendix B reports packing, token accounting, truth-cap ratios, and runtime outputs.

3.5 Contamination-Free Design

Because realistic task requests could otherwise be answered from public priors, SCALE-QA uses counterfactual local constraints—fictional organizations, nonstandard identifiers, locally defined policies, and unintuitive exceptions—as a contamination-control device, preserving realistic decision structure while making pretrained shortcuts less useful.

3.6 Quality Control and Audit

SCALE-QA is built through a multi-stage acceptance funnel rather than a one-shot synthetic dump. Deterministic normalization, adversarial distractor refinement, machine filtering, and human review enforce answerability, evidence grounding, label consistency, and distractor quality. Across construction logs, machine filters accept 28.8% of generated candidates, and 3 human reviewers accept 84.3% of reviewed candidates. The

dataset passes 3,000/3,000 full-turn exact evidence matches across 4,346 audited evidence snippets, with balanced answer labels and zero critical validation issues. Appendix A reports the full filter criteria, review dimensions, per-stage counts, and validation outputs.

Blind human realism audit. To check that counterfactual construction does not reduce SCALE-QA to artificial logic puzzles, we conduct a blind realism audit on 300 stratified examples with three anonymous annotators, yielding 900 valid annotations. On a 1–5 scale, the subset is rated natural (3.80), answerable (4.91), and plausibly constrained (3.99), with low ambiguity risk (1.45; lower is better). Majority answers agree with gold on 296/299 majority-valid examples (99.0%), with high answer-choice agreement (mean pairwise Cohen's $\kappa = 0.895$). Appendix A reports the full annotation protocol, κ range, the single no-majority case, and the three wrong-majority cases.

4 TSIM: Temporal-Semantic Interleaved Memory Reconstruction

4.1 Why Episodes, Not Chunks

Standard chunk retrieval is not a natural fit for SCALE-QA because the retrieval unit is often incomplete: a chunk may match the query while omitting the neighboring turn that makes a local constraint operative. Summary and memory-management systems can likewise surface related fragments without preserving the operative episode. Empirically, Standard RAG reaches only 7.7% CL Hit with Gemma2:9b at 128k, whereas TSIM reaches 70.7%, indicating that the dominant failure is surfacing the decisive episode rather than answer selection alone.

TSIM is therefore designed to preserve fine-grained evidence and episode-level coherence without externally supplied gold blocks. Its three modules—M1 semantic-shift episode segmentation, M2 multi-view episode indexing, and M3 evidence-first episode ranking—reconstruct candidate episodes before assembling compact evidence for the answer model. Figure 3 summarizes this episode-centered memory interface.

4.2 M1: Semantic-Shift Episode Segmentation

Rather than trusting existing block boundaries, TSIM converts the mixed conversation into a turn stream and infers episode boundaries online before retrieval. The goal is not general discourse parsing, but lightweight streaming segmentation that preserves operational units: contiguous spans that jointly establish, update, or invalidate a constraint.

Let $x_i \in \mathbb{R}^d$ be the normalized embedding of turn i. While scanning the stream, TSIM maintains a semantic center over recent turns inside the current episode, measures drift from that center, and applies minimum/maximum length guards:

$$c_i = \mathrm{norm}\left(|R_i|^{-1} \sum_{j \in R_i} x_j \right), \quad (1)$$

$$\Delta_i = 1 - \cos(x_i, c_i), \quad (2)$$

$$\mathrm{cut}(i) = \mathbb{I}[\Delta_i > \theta \wedge L_i \geq L_{\min}] \vee \mathbb{I}[L_i \geq L_{\max}]. \quad (3)$$

Here R_i denotes the recent turns still inside the current episode, and L_i is the current episode length. A cut starts a new episode before turn i. The rule is *streaming*: it uses no future turns, no dataset-provided gold blocks, and no offline clustering. Appendix C gives the local-window update, speaker-transition guard, and pseudocode.

We use this streaming segmenter as a lightweight proxy for operative episodes, not as a claim about gold discourse boundaries.

4.3 M2: Multi-View Episode Indexing

The key design choice is that raw hits, summary hits, and cluster hits are all converted into evidence for an episode, so the final prompt is assembled from top-ranked episodes rather than isolated turns or unrelated chunk neighbors.

Let E be the set of reconstructed episodes produced by M1. For each episode $e \in E$, TSIM builds three retrievable views with different granularity but the same episode anchor: a *raw view* over original turns, preserving lexical cues, numeric values, and narrow local constraints; a *summary view* storing an episode-level abstraction for gist retrieval, produced over reconstructed episodes rather than arbitrary chunks; and a *cluster view* grouping episodes by summary-embedding similarity into coarse semantic clusters with pointers back to member episodes. At query time, raw and summary hits contribute to episode scores, while cluster hits route and boost attached episodes rather than replacing raw or summary evidence in the answer prompt.

4.4 M3: Evidence-First Episode Ranking

At query time, TSIM embeds the query once and retrieves against the three episode views. For each

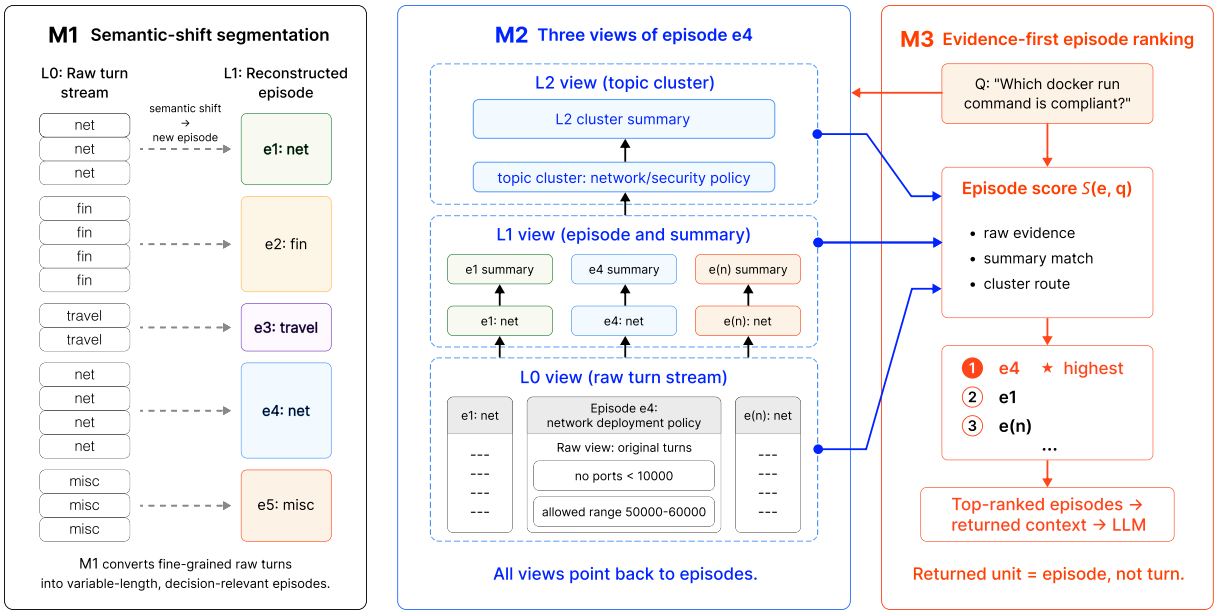

Figure 3: Episode-centered multi-view memory in TSIM. M1 segments the turn stream into reconstructed episodes. M2 indexes each episode through raw, summary, and cluster views. M3 converts hits from all views into episode-level scores and returns top-ranked episodes, not isolated turns, to the answer model.

candidate episode e, retrieval evidence is aggregated into an episode-level score:

$$\begin{aligned}\mathrm{Score}(e,q) &= \alpha\,\mathrm{RawScore}(e,q)\\ &\quad+\beta\,\mathrm{SummaryScore}(e,q)\\ &\quad+\gamma\,\mathrm{ClusterRouteScore}(e,q).\end{aligned} \tag{4}$$

The three terms aggregate raw-turn, episode-summary, and cluster-routing evidence; Appendix C gives the implementation definitions, including semantic expansion. The ranking policy is evidence-first: raw and summary evidence receive the largest mass, clusters route candidate episodes rather than becoming prompt content, and the final prompt contains top-ranked episodes rather than all retrieved neighbors. Scoring weights are frozen on development packages and reported in Appendix D.

5 Experimental Setup

All headline experiments use the 3,000-question SCALE-QA dataset with its deterministic length-configurable runtime builder, evaluating 16k–128k constructed contexts on the full dataset and extending to a 1M diagnostic subset. The main 128k comparison uses three answer backends: Gemma2:9b (local), Gemini 2.5 Flash (long-context commercial), and GPT-4o-mini (high-throughput ablation). DeepSeek R1 and Gemini 2.5 Flash additionally serve as strong-reasoning and long-window probes for the context-scaling diagnostic.

We compare TSIM with Standard RAG, Hybrid-RRF Chunk RAG, RAPTOR strict no-block (Sarthi et al., 2024), MEMGPT (Packer et al., 2023), and HIPPORAG (Gutiérrez et al., 2024); Full Context is reported only as a native-context diagnostic. The GPT-4o-mini 128k block also includes Tuned Hybrid-Rerank Chunk RAG, whose full retrieval stack is described in Appendix C. All systems receive identical length-batched runtime packages, noise seeds, and writeback regimes; only the memory or retrieval strategy differs. The evaluation is stateful, so CL Hit denotes evidence hit along the backend-conditioned closed-loop trajectory. TSIM uses one frozen configuration across all backends.

We report **Accuracy**, **CL Hit**, context tokens, and latency. Accuracy is forced-choice four-way MCQ accuracy; CL Hit is expected-evidence presence in retrieved context along the closed-loop trajectory. Frozen TSIM configuration, token accounting, variance audits, latency caveats, and runtime details appear in Appendices B–F.

6 Results

6.1 Context Scaling: Native Context vs. Reconstructed Memory

Figure 4 evaluates context scaling under GPT-4o-mini. Full Context drops from 62.5% at 16k to 29.8% at 128k, despite receiving the constructed context directly, while TSIM remains at 73.8% using about 1k retrieved tokens. Evidence inclusion alone is insufficient; the system must recover the

Backend	Retriever	Acc	CL Hit	Rat. Sim.	CtxTok	Lat.
Gemma2:9b	**TSIM**	**69.6**	70.7	0.472	1060.8	4.20
	Standard RAG	24.4	7.7	0.296	929.5	3.35
	Hybrid-RRF Chunk RAG	31.1	11.2	0.254	1193.6	4.00
	RAPTOR	40.6	35.3	0.403	790.5	20.67
	MEMGPT	60.1	62.6	0.413	2301.9	3.88
	HIPPORAG	25.2	12.5	0.380	748.9	6.51
Gemini 2.5 Flash	**TSIM**	**80.2**	67.9	0.445	1275.3	1.46
	Standard RAG	29.8	7.6	0.166	912.6	0.98
	Hybrid-RRF Chunk RAG	32.8	11.2	0.183	1193.3	1.20
	RAPTOR	52.6	34.8	0.305	790.0	9.85
	MEMGPT	74.6	62.3	0.398	2349.6	1.19
	HIPPORAG	34.4	12.1	0.245	752.7	3.10
GPT-4o-mini	**TSIM**	**73.8**	74.2	0.485	1043.6	2.66
	Standard RAG	27.1	7.6	0.238	910.1	1.72
	Hybrid-RRF Chunk RAG	31.1	11.3	0.203	1193.5	1.59
	RAPTOR	43.0	35.1	0.419	789.2	11.59
	MEMGPT	50.2	58.6	0.298	2238.4	1.77
	HIPPORAG	31.0	11.9	0.202	753.7	2.85
	Tuned Hybrid-Rerank Chunk RAG †	56.2	63.5	0.487	3004.2	11.91

Table 2: Cross-backend main comparison on the 3,000-question SCALE-QA dataset at 128k. Accuracy is the primary metric; CL Hit, rationale similarity, context tokens, and latency are supporting diagnostics. CL Hit is backend-conditioned closed-loop evidence hit and should be compared within backend blocks. Accuracy, CL Hit, rationale similarity, and context size use full-dataset runs; API latency uses serial Quick100 controls rather than the parallel scheduler. The daggered GPT-4o-mini row is a tuned non-episodic retrieval control (BM25 + BGE dense/rerank + HyDE + RRF + parent-window + MMR; Appendix C) without TSIM episode segmentation or multi-granularity memory; its GPU-assisted latency is a cost diagnostic, not a hardware-normalized leaderboard value. Appendix Table 19 breaks out Accuracy by domain.

operative episode that makes the evidence binding.

6.2 Main Architectural Comparison

In the 128k setting, Table 2 shows that TSIM obtains the highest accuracy in all three backend blocks: 69.6% with Gemma2:9b, 80.2% with Gemini 2.5 Flash, and 73.8% with GPT-4o-mini. Strengthening chunk retrieval helps but does not close the gap: the tuned non-episodic GPT-4o-mini control reaches 56.2% accuracy and 63.5% CL Hit with 3.0K context tokens, still 17.6 points below TSIM despite using nearly three times the prompt context. The decisive factor is therefore episode organization, not first-stage retrieval strength.

The closest API-backed comparison is TSIM versus MEMGPT under Gemini 2.5 Flash: TSIM improves accuracy by 5.58 points, with a paired bootstrap 95% confidence interval of $[4.15, 6.97]$; Appendix F reports the full significance, per-domain, variance, and token audits. TSIM also uses approximately half the prompt context of MEMGPT under the same backend (1.3K vs. 2.3K tokens), showing that episode-anchored retrieval is more token-efficient.

6.3 Progressive Ablation

Table 3 isolates which TSIM components contribute the gain. Accuracy rises monotonically: 26.2% with Standard RAG, 43.4% with fixed-token no-block retrieval, 55.5% with semantic-drift episodes, and 74.2% with the full multi-view episode memory stack. Semantic-drift segmentation improves over fixed-token chunks, and the multi-view stack makes reconstructed episodes substantially more useful.

Stage	Variant	Acc	CL	Rat.	Lat.	Tok.
L0	Std. RAG top-5	26.2	5.6	0.246	1.61	968.4
L1	Fixed-token direct	43.4	35.0	0.416	3.52	1269.2
L2	Semantic-drift direct	55.5	52.2	0.430	3.62	1179.8
L3	Full TSIM stack	**74.2**	74.3	0.485	3.56	1044.4

Table 3: Main-module ablation under GPT-4o-mini. Accuracy is the primary metric; CL/Rat./Tok. denote supporting CL Hit, rationale similarity, and context-token diagnostics; latency is within-table only.

6.4 Strong-Backend Diagnostic: Reasoning and Long-Window Scaling

Figure 5 extends context scaling to stronger reasoning and longer-window backends on a stratified subset. Stronger native-context models reduce but do not remove the need for episode reconstruction. At 128k, DeepSeek R1 Full Context reaches 81.2% while TSIM reaches 93.8%. At the Gemini 2.5 Flash 1M diagnostic budget, Full Context reaches 87.2% with 1.05M prompt tokens and 23.87s latency, whereas TSIM reaches 96.5% with about 1.3k retrieved tokens and 2.16s latency. Episode reconstruction is therefore more accurate and compact than scaling the native context window alone.

6.5 Missing vs. Misusing Evidence

On Gemma2:9b, Standard RAG retrieves expected evidence on only 7.7% of examples and reaches 24.4% accuracy, while TSIM reaches 70.7% CL Hit and 69.6% accuracy. Baseline failure is therefore dominated by missing the decisive episode,

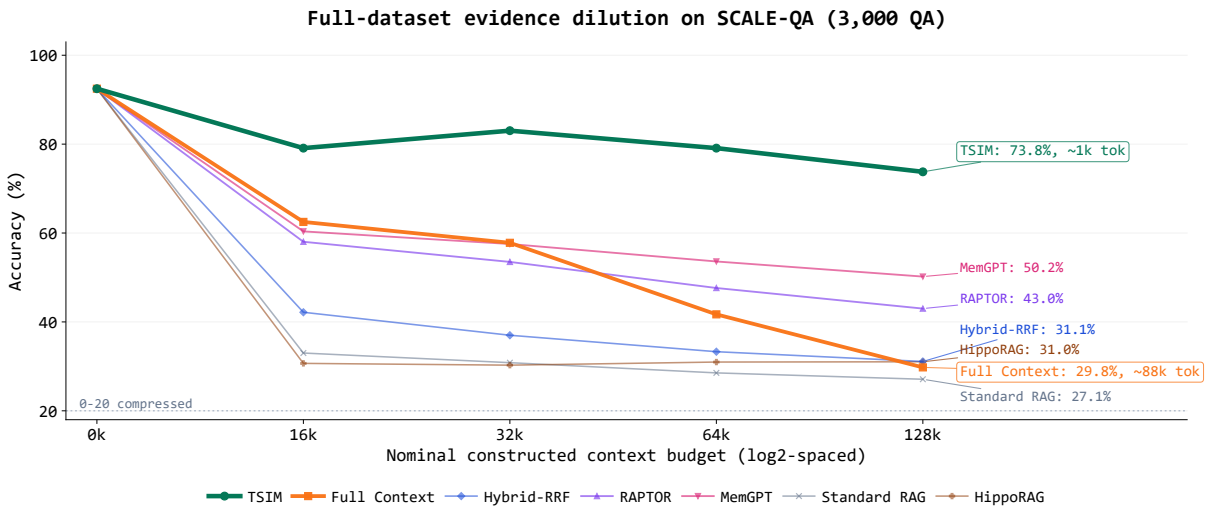

Figure 4: Context scaling on all 3,000 SCALE-QA questions under GPT-4o-mini. The shared 0k point is an evidence-only prompt; 16k–128k points use length-controlled mixed-session packages, with Full Context receiving the constructed context directly. The x-axis is log2-spaced; 128k callouts report measured packed prompt/context estimates from Appendix F.

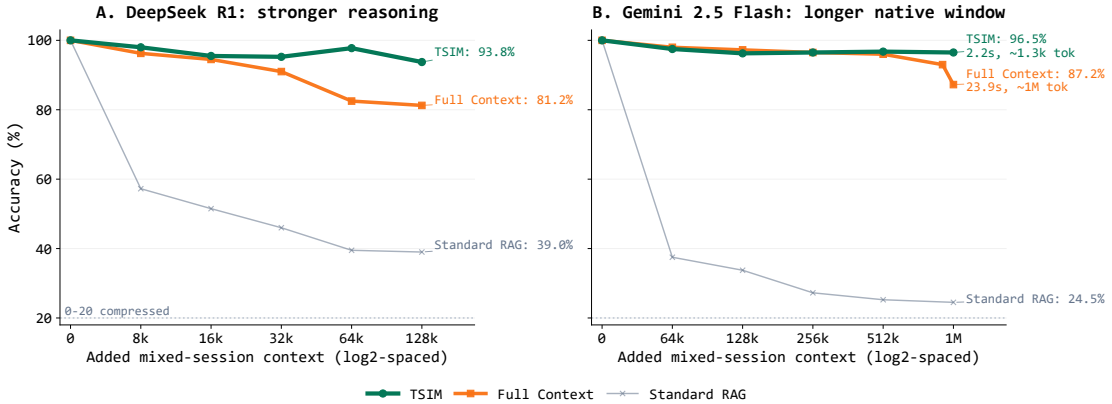

Figure 5: Context-scaling stress test on a stratified 400-question SCALE-QA subset. The log2-spaced x-axis shows added mixed-session context; 0k is evidence-only. Plot A uses DeepSeek R1; Plot B uses Gemini 2.5 Flash at 1M. Standard RAG is a lower-anchor control; latency uses serial averages.

whereas residual TSIM errors reflect a different bottleneck: verbose or conflicting memory can still lead the answer model to misuse local constraints, especially in Social and Biz-Ops examples where local overrides contradict plausible public defaults. Appendix Table 20 confirms this pattern across all three stress views.

7 Conclusion

We introduced SCALE-QA and TSIM to study episode integrity failure, a regime topic-isolated benchmarks largely miss: recovering dormant local constraints from long mixed-topic conversations. The bottleneck is not whether evidence fits inside the context window, but whether the memory system reconstructs the episode that makes it operative.

On the 3,000-question SCALE-QA dataset, TSIM outperforms Standard RAG, Hybrid-RRF Chunk RAG, RAPTOR, MEMGPT, and HIPPORAG, while SCALE-QA remains oracle-answerable and zero-shot hard. The ablation shows why: semantic-drift episodes outperform fixed-token chunks, and the multi-granularity memory stack makes those episodes more usable without simply inflating the prompt. The 1M-token diagnostic makes the implication concrete: a long-window model can see the evidence and still pay 1.05M tokens and 23.87s for 87.2% accuracy, while TSIM answers from about 1.3k retrieved tokens at 96.5%. Future long-context agent evaluation should therefore move beyond needles in static haystacks toward dynamic conversational reconstruction: deciding which episode is still operative and which local exception overrides the generic rule. SCALE-QA and the TSIM reference implementation accompany this submission as supplementary material.

8 Limitations

Two limitations are important to keep in view. First, SCALE-QA is counterfactually constructed rather than sampled from naturally occurring assistant logs, so it cannot fully capture the distributional, stylistic, or privacy constraints of deployed systems. To support reliability, we include exact evidence audits, oracle/zero-shot calibration, and a 300-example blind human audit with three anonymous annotators, but real-log validation remains important future work.

Second, SCALE-QA uses four-way multiple-choice questions to make episode integrity failure reproducible and evidence-auditable. This improves auditability but does not cover partial answers, hedged responses, tool-use follow-up, or long-form explanation quality. We therefore view SCALE-QA as a targeted diagnostic for constraint-grounded task QA, with open-ended assistant-memory evaluation left as complementary future work.

9 Ethics Statement

The benchmark includes scenarios inspired by medicine, law, finance, and operations. These are used to evaluate context-grounded memory and evidence use rather than to provide professional advice. The use of counterfactually privatized synthetic scenarios is deliberate: it reduces privacy risks and benchmark leakage while still allowing realistic decision structures to be modeled.

The distractor/noise material used by the runtime builder is drawn from publicly released ShareGPT (Apache 2.0) and UltraChat (MIT) chat data, following prior long-term memory benchmark practice (Wu et al., 2025). Three human reviewers performed construction-stage quality control, and three separate anonymous human auditors conducted the realism audit; all six were unpaid research-group members familiar with the task.

Because scenario seeds and counterfactual constraints are LLM-assisted, they may inherit biases from generation models or selected domains; deterministic gates and human audit reduce but do not eliminate this risk. SCALE-QA evaluates memory and evidence-use capability and is not intended for deployed decision systems in clinical, legal, or financial settings. We therefore recommend that deployment-oriented follow-up treat this benchmark as an evaluation resource, not as a substitute for domain-qualified human expertise.

References

Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2022. TopiOCQA: Open-domain conversational question answering with topic switching. *Transactions of the Association for Computational Linguistics*, 10:468–483.

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. 2025. Memorybench: A benchmark for memory and continual learning in llm systems. *arXiv preprint arXiv:2510.17281*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336.

Yiruo Cheng, Kelong Mao, Ziliang Zhao, Guanting Dong, Hongjin Qian, Yongkang Wu, Tetsuya Sakai, Ji-Rong Wen, and Zhicheng Dou. 2025. CORAL: Benchmarking multi-turn conversational retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1308–1330, Albuquerque, New Mexico. Association for Computational Linguistics.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Anand Kannappan, Rebecca Qian, and Peng Wang. 2025. Memtrack: Evaluating long-term memory and state tracking in multi-platform dynamic agent environments. *arXiv preprint arXiv:2510.01353*.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. 2023. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1762–1777, Toronto, Canada. Association for Computational Linguistics.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems*, volume 37.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? COLM 2024; arXiv preprint arXiv:2404.06654.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Sangyeop Kim, Yohan Lee, Sanghwa Kim, Hyunjong Kim, and Sungzoon Cho. 2025. Pre-storage reasoning for episodic memory: Shifting inference burden to memory for personalized dialogue. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 22096–22113, Suzhou, China. Association for Computational Linguistics.

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2026. LLMs get lost in multi-turn conversation. In *International Conference on Learning Representations*.

Jinhyuk Lee, Anthony Chen, Zhuyun Dai, Dheeru Dua, Devendra Singh Sachan, Michael Boratko, Yi Luan, Sébastien M. R. Arnold, Vincent Perot, Siddharth Dalmia, Hexiang Hu, Xudong Lin, Panupong Pasupat, Aida Amini, Jeremy R. Cole, Sebastian Riedel, Iftekhar Naim, Ming-Wei Chang, and Kelvin Guu. 2025. Loft: Scalable and more realistic long-context evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 6713–6738, Albuquerque, New Mexico. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474.

Mufei Li, Dongqi Fu, Limei Wang, Si Zhang, Hanqing Zeng, Kaan Sancak, Ruizhong Qiu, Haoyu Wang, Xiaoxin He, Xavier Bresson, Yinglong Xia, Chonglin Sun, and Pan Li. 2025. Haystack engineering: Context engineering for heterogeneous and agentic long-context evaluation. *arXiv preprint arXiv:2510.07414*.

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

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.

Endel Tulving. 1972. Episodic and semantic memory. In Endel Tulving and Wayne Donaldson, editors, *Organization of Memory*, pages 381–403. Academic Press, New York.

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

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-Pack: Packed Resources for General Chinese Embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. ∞Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277, Bangkok, Thailand. Association for Computational Linguistics.

A SCALE-QA Dataset Construction and Audits

The SCALE-QA dataset includes the strict-valid records, machine-readable validation reports, a dataset card, and the deterministic runtime builder. The split labels, `codex` and `claude-code`, are balanced dataset partitions rather than method baselines. Tables 5, 6, and 7 summarize the domain composition, dataset validation, and source-level calibration.

Each accepted record passes the same construction and filtering loop used in the main paper: scenario seed generation, chat and multiple-choice realization, deterministic normalization, adversarial refinement, exact evidence alignment, oracle answerability, and zero-shot hardness checks. The construction-log aggregates report 28.8% acceptance after deterministic machine filtering and 84.3% acceptance after review by 3 human reviewers. The validation report confirms the final dataset properties rather than these intermediate construction-log rates: answer labels are globally balanced and nearly balanced within each topic, with each domain containing 300 examples and each split contributing 150 examples per domain.

Construction and validation details. Machine filters check schema validity, option parseability, answer-label consistency, exact evidence alignment, oracle answerability, and zero-shot hardness. Three human reviewers then assess answerability, uniqueness, evidence grounding, and distractor ambiguity. Per-stage acceptance rates and final dataset counts are reported in the tables above.

Multiple-choice evaluation protocol. SCALE-QA uses deterministic four-way MCQ to make evidence-grounded grading auditable. This design isolates memory recovery from generation-style variation, which otherwise conflates surface style, answer length, and evaluator behavior with memory ability. In each item, distractors are plausible under generic priors but invalidated by local evidence, so a fluent answer that misses the operative constraint should fall into the plausible-distractor trap. Conversely, a system that retrieves the correct evidence span but produces awkward prose should still receive credit for selecting the correct option. Open-ended assistant-memory evaluation remains complementary; SCALE-QA focuses on reproducible episode-integrity diagnosis rather than long-form explanation quality.

A.1 Blind Human Realism Audit

The final human audit uses 300 stratified examples and three anonymous annotators. All three annotators completed all examples, producing 900 valid annotation rows. Table 8 reports item-level means on a 1–5 Likert scale; lower is better for ambiguity risk. The majority answer agrees with gold in 296 of 299 majority-valid examples (99.0%); the remaining audit set contains one no-majority case and three wrong-majority cases. Annotator-level response balance, pairwise agreement, stress-label distribution, and per-topic realism are reported in Tables 9, 10, 11, and 12.

The single no-majority case is a Game-Novel item (`S300-204`) where annotators split across three options under overlapping long-range-bridge and constraint-trap cues. The three wrong-majority cases are Network-Hardware (`S300-084`), Engineering (`S300-213`), and Social-Personal (`S300-247`) items; all involve overlapping stress cues, and two have all three cue labels active. These cases are retained in the audit accounting rather than removed.

B Length-Controlled Runtime Evaluation Protocol

SCALE-QA evaluates systems under a user-specified target constructed context length. This value is a benchmark-side packing target, not a model-window budget, a benchmark-internal upper bound, or a claim that every provider tokenizer assigns exactly the same number of native tokens. For each experiment, the builder creates a runtime package with the same selected records, noise, seeds, batch mapping, and executable files for every compared method. We therefore use the benchmark-side estimate for deterministic packing and use stored prompt/context text for any backend-specific tokenizer audit. Table 13 summarizes the length-controlled packing protocol.

For the full 3,000-question dataset, the truth corpus is approximately 393,245 benchmark-side tokens. Thus 128k and 256k experiments naturally in-

Axis	LongMemEval	SCALE-QA
History format	Timestamped chat histories with session structure	Flat mixed-topic threads with no boundary metadata
Question type	Flexible personal-memory QA	Constraint-grounded task QA
Domain ontology	Personal-life ontology (health, hobbies, work-life, etc.)	Ten task-oriented operational domains
Target failure mode	Long-term memory over sessions and updates	Episode integrity failure in unsegmented threads
Evaluation protocol	Judged open-ended responses	Deterministic four-way MCQ with exact evidence traces

Table 4: Axis-by-axis distinction between LongMemEval and SCALE-QA. SCALE-QA builds on scalable long-memory evaluation but isolates boundary-free episode reconstruction in flat task-oriented threads.

Domain	QA	Typical evidence forms	Common retrieval stress cues
CS-Software	300	code/version/deployment rules	stale defaults; tool exceptions
Network/Hardware	300	port, routing, hardware policies	constraint traps; local compliance
Finance	300	credit, reimbursement, risk rules	overwritten eligibility; denials
Legal	300	contracts, filings, exceptions	exception resolution; operative clauses
Biomed	300	protocols, safety notes, exclusions	dormant safety constraints
Engineering	300	materials, tests, device specs	incompatibilities; hidden failures
Biz-Ops	300	HR, procurement, process memos	state overwrite; approval chains
Social/Personal	300	preferences and personal context	pragmatic overrides; false defaults
Game/Novel	300	world rules and quest states	long-range bridges; state validity
Daily Life	300	household, travel, schedule rules	local exceptions; stale plans

Table 5: Supplementary SCALE-QA domain overview. Every domain contributes exactly 300 questions; correct labels are globally balanced A/B/C/D = 750/750/750/750. Stress cues are representative rather than mutually exclusive, since many examples combine stale state, long-range bridges, and local exception traps.

Release property	Value
Total QA records	3,000
Public split labels	$2 \times 1{,}500$
Topics	10×300
Split-topic cells	20×150
Correct labels	A/B/C/D = 750/750/750/750
Audited evidence snippets	4,346
Full-turn exact records	3,000/3,000
Critical validation issues	0

Table 6: SCALE-QA dataset audit. Counts are taken from the validation report and expected-document audit.

Split	QA	Dom.	Evid.	Zero	Oracle
Full dataset	3,000	10×300	3,000/3,000	7.63/18.10	100.00/100.00
Codex	1,500	10×150	1,500/1,500	7.40/19.27	100.00/100.00
Claude	1,500	10×150	1,500/1,500	7.87/16.93	100.00/100.00

Table 7: Supplementary source-level calibration. Zero columns report `gemma2:9b`/Gemini 2.5 Flash zero-shot solvability without supporting chat evidence; oracle columns give the model the relevant context.

Metric	Mean	Median	Std.	Direction
Naturalness	3.80	3.67	0.25	higher better
Answerability	4.91	5.00	0.17	higher better
Ambiguity risk	1.45	1.33	0.32	lower better
Constraint plausibility	3.99	4.00	0.29	higher better

Table 8: Overall blind human realism audit on 300 stratified examples with three anonymous annotators.

Ann.	Gold Acc.	Max ans.	Natural	Answerable	Ambig.	Plausible
Ann. 1	99.0	25.7	4.17	4.90	1.19	3.98
Ann. 2	91.0	30.3	4.04	4.87	1.92	4.08
Ann. 3	96.3	25.3	3.20	4.95	1.25	3.90

Table 9: Annotator-level descriptive checks for the three-annotator human audit. Gold Acc. is agreement with the benchmark answer; Max ans. is the largest selected-answer share, used as a response-balance check.

Pair	Ans. agr.	Ans. κ	Primary agr.	State	Long	Trap	Cons. gold
Ann. 1–2	91.3	0.884	56.7	43.3	83.0	84.7	272/274
Ann. 1–3	96.0	0.947	16.3	45.3	83.0	100.0	287/288
Ann. 2–3	89.0	0.853	32.7	57.3	72.0	84.7	265/267

Table 10: Pairwise agreement in the final human audit. Answer-choice agreement is high, including chance-corrected pairwise Cohen's κ; primary-stress agreement is lower because many examples contain overlapping stress cues. Cons. gold reports gold agreement among pairwise majority-resolved cases.

stantiate length-batched evaluation, whereas 512k and larger targets can enter the full-corpus regime. The main 128k comparison uses four deterministic batches with the same mix seed and truth-cap policy for every retrieval system. Beyond the reported settings, longer packages can be generated by drawing additional distractor turns; future longer-window models can be evaluated without changing the benchmark logic.

View	Label	Count	Rate
Multi-label cue	State overwrite	196	65.3
Multi-label cue	Long-range bridge	291	97.0
Multi-label cue	Constraint trap	300	100.0
Multi-label cue	Any multi-label	298	99.3
Multi-label cue	All three cues	189	63.0
Primary cue	Constraint trap	158	52.7
Primary cue	State overwrite	80	26.7
Primary cue	Long-range bridge	9	3.0
Primary cue	Needs adjudication	53	17.7

Table 11: Human stress-label distribution on the same 300 examples. Stress cues are intentionally multi-label; primary labels summarize the dominant cue only. Needs adjudication indicates cases where annotators did not form a stable dominant-stress label, not invalid examples.

Topic	Natural	Answerable	Ambig.	Plausible
Biomed	3.66	4.86	1.49	4.00
Biz-Ops	3.79	4.92	1.31	4.13
CS-Software	3.74	4.94	1.34	4.09
Daily-Life	4.09	4.91	1.38	4.04
Engineering	3.72	4.90	1.42	4.13
Finance	3.79	4.88	1.41	4.06
Game-Novel	3.71	4.93	1.61	3.63
Legal	3.67	4.92	1.38	3.94
Network	3.72	4.90	1.33	3.94
Social	4.13	4.90	1.86	3.88

Table 12: Per-topic realism means in the 300-example human audit. Ambig. denotes ambiguity risk, where lower is better.

Protocol item	Definition
Token accounting	Benchmark-side estimate tokens $\approx 1.3\times$ whitespace word count, used for deterministic packing and constructed-length reporting, not for asserting exact provider-native token parity.
Length scaling	No built-in length cap; practical limits are distractor-pool size and computational budget. This paper reports settings through 1M to match evaluated native-context backends.
Full-corpus regime	Used when selected truth tokens fit inside the target context length; the full selected truth background is retained and noise fills remaining space.
Length-batched regime	Used when selected truth tokens exceed the target length; records are deterministically partitioned into budget-controlled batches.
Packing rule	Deterministic capacity-constrained best fit: records are ordered by descending truth-token count and then by question name.
Truth cap	In length-batched mode, the default truth-cap ratio is 0.82, leaving room for noise and prompt/interface overhead.
Noise fill	Noise blocks are deterministically shuffled with the noise seed and inserted identically for all methods evaluated on the package.
Runtime outputs	Each package contains a manifest, stats, selected IDs, `GROUND_TRUTH_HISTORY`, `EVALUATION_QUERIES`, and `NOISE`.

Table 13: Length-controlled evaluation protocol used by the deterministic runtime-package builder.

C Method and Baseline Implementation Details

All methods are evaluated under the same persistent writeback setting. None of the retrieval baselines receives dataset-provided gold blocks, future turns, or method-specific noise. The key implementation differences are summarized in Table 14.

Tuned Hybrid-Rerank control. The Tuned Hybrid-Rerank Chunk RAG control combines BM25 sparse retrieval (Robertson and Zaragoza, 2009), BGE dense retrieval and reranking (Xiao et al., 2024), HyDE query rewriting (Gao et al., 2023), reciprocal-rank fusion (Cormack et al., 2009), parent-window expansion, and MMR packing (Carbonell and Goldstein, 1998). It is non-episodic: it does not use semantic episode segmentation or multi-granularity TSIM memory.

All dense TSIM memory levels use `BAAI/bge-large-en-v1.5` through Sentence-Transformers. Raw turns, episode summaries, and L2 centroids are stored in separate Chroma HNSW indices with cosine distance. The implementation keeps `all-MiniLM-L6-v2` only as an offline portability fallback, not for reported runs. Table 15 gives compact pseudocode for the streaming semantic-drift segmenter.

M1 formal definitions. Let $x_i \in \mathbb{R}^d$ be the normalized embedding of turn i, R_i the recent turns still inside the current episode, and L_i the current episode length. The segmenter uses:

$$
\begin{aligned}
c_i &= \operatorname{norm}\left(\frac{1}{|R_i|}\sum_{j\in R_i} x_j\right), \\
\Delta_i &= 1 - \cos(x_i, c_i), \\
\operatorname{cut}(i) &= \mathbb{I}[\Delta_i > \theta \wedge L_i \geq L_{\min}] \\
&\quad \vee \mathbb{I}[L_i \geq L_{\max}] \,.
\end{aligned}
$$

The first guard opens a boundary only after the current episode reaches the minimum length, while the second prevents oversized episodes.

The ranking terms abbreviated in the main-text

System	Retrieval unit	Context assembly	Purpose in comparison
Standard RAG top-5	Individual retrieved chunks/turns	Top five retrieved units are passed to the answer backend.	Tests whether short semantic retrieval alone can recover the decisive evidence.
Hybrid-RRF Chunk RAG	Dense chunks + BM25 sparse hits + neighboring turns	Dense and sparse hits are fused with reciprocal-rank fusion, expanded with local neighbors, and packed under a matched context cap.	Tests whether lightweight hybrid chunk retrieval closes the episode-reconstruction gap.
Tuned Hybrid-Rerank Chunk RAG	Dense/sparse candidates + HyDE + cross-encoder reranking	Adds query expansion, RRF fusion, BGE reranking, parent-window expansion, and MMR packing under the same GPT-4o-mini 128k task.	Tests whether a strongly tuned but non-episodic chunk pipeline can close the gap, and exposes its reranking cost.
RAPTOR strict no-block	Hierarchical summaries built without gold blocks	Retrieved hierarchy outputs are mapped into the same no-block evaluation regime.	Tests whether hierarchical abstraction alone solves interleaved conversational memory.
MEMGPT paper-default	Explicit memory-management substrate	Retrieved memory material is inserted under the same answer and writeback protocol.	Tests whether higher recall from a memory-style substrate converts into final accuracy.
HIPPORAG paper-default	Graph-style retrieval substrate	Retrieved graph/memory evidence is evaluated with the same question set and scoring protocol.	Tests transfer of graph-centric memory retrieval to flat-thread writeback evaluation.
Official TSIM	Raw turns, reconstructed episodes, and L2 semantic clusters	Raw and summary hits contribute to episode scores; L2 routes to related episodes; final top episodes form the prompt.	Tests block-free episode reconstruction and multi-granularity memory organization.

Table 14: High-level implementation distinctions for the main retrieval and memory systems.

Streaming semantic-drift segmentation in Official TSIM
1. Flatten the mixed conversation into a turn stream and encode each turn as a normalized dense vector.
2. Maintain the current segment start s and token count L.
3. For incoming turn i, average only the recent turns still inside the current segment to form local center c_i.
4. Score the newest turn by cosine similarity to c_i, plus a small bonus for a `model` $\rightarrow$ `user` transition.
5. If $L \geq \tau_{\min}$ and the score falls below threshold θ, close the segment before turn i.
6. If $L \geq \tau_{\max}$, force a boundary even if semantic drift is weak.
7. Continue streaming without revisiting earlier boundaries.

Table 15: Compact pseudocode view of the semantic-drift segmenter used by Official TSIM. The final system uses a simple local-window cosine rule rather than a heavier offline clustering procedure for boundary decisions.

scoring equation are:

$$R_r(e,q) = \sum_{r \in H_r(e,q)} \mathrm{sim}(q,r), \tag{5}$$

$$R_s(e,q) = \sum_{s \in H_s(e,q)} \mathrm{sim}(q,s), \tag{6}$$

$$R_l(e,q) = \sum_{c \in H_l(q)} \mathbb{I}[e \in \mathrm{Expand}(c)] \cdot \lambda_l^{\mathrm{rank}(e;c)} \mathrm{sim}(q,c), \tag{7}$$

$$R_{\mathrm{sem}}(e,q) = \sum_{e' \in \mathrm{Seeds}(q)} \lambda_{\mathrm{sem}} \, \mathrm{sim}(q,e') \cdot \mathrm{sim}(e',e). \tag{8}$$

Here $H_r(e,q)$ and $H_s(e,q)$ are raw-turn and summary hits attached to episode e, while $H_l(q)$ is the set of retrieved L2 clusters. $\mathrm{Expand}(c)$ returns the small set of episodes attached to cluster c, and $\mathrm{Seeds}(q)$ are the top local episode candidates used for semantic expansion.

Stage	Data	Purpose
Search	300 QA	Fast screening of candidate retrieval configurations.
Confirm	600 QA	Check stability across 64k, 128k, 256k, and 512k targets.
Answer validation	200 QA	Confirm that retrieval gains transfer to answer accuracy.
Full retrieval check	3,000 QA	Verify evidence coverage and context stability at 128k.

Table 16: Frozen-configuration selection protocol for TSIM.

D TSIM Configuration Selection and Ablation

All reported TSIM results use one frozen configuration rather than backend-specific retuning. Because TSIM is an architecture rather than an end-to-end trained model, its scalar coefficients are calibrated constants rather than learned parameters. The selection protocol emphasizes evidence-reconstruction stability over small-sample answer accuracy alone. Table 16 summarizes the frozen-configuration selection stages, and Table 18 lists the exact constants used in reported runs.

Table 17 summarizes representative neighborhood stability from the bounded development sweep. The table is not meant to present every tried configuration or to claim that the reported constants are uniquely optimal. Instead, it shows that several nearby settings around the same semantic-drift threshold, retrieval breadth, and L2 routing weight remain strong on disjoint search, confirmation, and answer-validation splits. The reported setting is frozen because it gives the best balance of evidence coverage, answer transfer, and compact context; search-only peaks are not selected unless they also confirm under the longer-context stability check.

Candidate	Representative variation	Search Hit	Confirm Hit	Answer Hit	Answer Acc	Answer Ctx
Official calibrated	$\theta = .70, k_f = 5, k_r/k_s = 28/20, w_l = .75$	86.33	**79.33**	**87.50**	78.50	1011.6
High-threshold compact	$\theta = .74, k_r = 44, k_s = 16, w_l = .85$	85.67	77.62	86.00	**79.00**	757.7
Lean memory stack	$k_f = 4, k_r/k_s = 36/12, w_l = .65$	84.00	76.58	**87.50**	77.50	818.2
Lean L2-exp = 1	Same as lean, one L2-to-episode expansion	83.67	76.85	**87.50**	77.50	818.1

Table 17: Representative neighborhood stability from the development sweep. Search300 and confirm600 report CL Hit; answer200 reports CL Hit, Accuracy, and average retrieved context tokens. The answer-validation prompt constrains outputs to valid answer choices, so the reported answer accuracy follows the paper's main Accuracy convention.

Component	Frozen reported constants
Segmenter	`min_tokens`=120, `max_tokens`=320, `recent_window`=4, $\theta = 0.70$, speaker bonus $b = 0.03$, minimum two-message buffer.
Retrieval breadth	$k_r = 28, k_s = 20, k_l = 2, k_{\text{final}} = 5$.
Ranking weights	$w_r = 1.15, w_s = 1.20, w_l = 0.75, \lambda_{\text{sem}} = 0.55, \lambda_l = 0.70$.
Expansion	Cluster expansion $k = 2$, L2-to-episode expansion $k = 2$, cluster threshold 0.42, soft margin 0.08, temporal expansion hops 0.
Prompt assembly	Reconstructed episodes capped at eight messages with two-message overlap; L2 cluster summaries route and boost candidates but are not inserted directly into the prompt.

Table 18: Exact TSIM constants used in reported runs. These values are moved out of the main text to avoid visually overloading the method narrative with implementation constants.

E Additional Experimental Results

Table 19 expands the aggregate cross-backend comparison from Table 2 into all ten domains. It is an audit table rather than an additional leaderboard.

F Statistical and Token Accounting Audit

The main text reports compact aggregate scores, but the raw result files are example-aligned. We therefore audit the final main-table comparisons with paired bootstrap tests over the same 3,000 questions. The Hybrid-RRF Chunk RAG runs passed validation with 3,000 rows, zero run errors, and no duplicate IDs; the separate Tuned Hybrid-Rerank Chunk RAG control also passed validation with 3,000 rows and zero duplicate IDs. The backend-output split, paired bootstrap comparisons, and variance/token audit are reported in Tables 23, 24, and 25. Tables 21 and 22 provide Full Context evidence-containment and prompt-template diagnostics.

Metric and backend notes. Accuracy is the fraction of examples for which the selected option matches the gold label under forced-choice four-way MCQ grading. The three main answer backends provide complementary checks: Gemma2:9b for local reproducibility, Gemini 2.5 Flash for long-

Gemma2:9b Domain	TSIM	Std. RAG	Hybrid	RAPTOR	MemGPT	HippoRAG
CS	79.0	29.3	33.7	48.0	66.7	25.3
Net	77.3	26.7	33.7	49.3	67.0	27.0
Fin	54.3	22.7	29.3	32.7	46.7	23.3
Legal	60.0	22.3	28.0	36.3	56.3	26.7
Bio	74.7	25.3	36.7	39.7	66.7	21.3
Eng	69.7	23.0	30.3	35.3	55.0	22.0
Biz	64.3	25.3	32.3	36.3	54.0	29.7
Social	68.0	20.7	25.3	36.0	56.3	18.7
Game	68.3	18.0	26.3	36.0	58.3	33.0
Daily	80.7	31.0	35.3	57.3	73.7	25.7
All	**69.6**	24.4	31.1	40.6	60.1	25.2
Gemini 2.5 Flash						
CS	85.3	37.7	35.7	66.0	91.7	34.7
Net	93.0	30.7	33.3	61.3	87.0	34.3
Fin	67.0	26.3	28.0	37.7	58.3	31.3
Legal	79.0	27.7	30.3	45.7	67.3	38.0
Bio	86.3	34.3	43.7	62.3	84.0	35.0
Eng	79.3	28.0	32.0	50.3	73.3	28.0
Biz	72.0	28.0	30.7	45.7	67.0	39.0
Social	75.0	25.7	28.7	43.3	64.0	29.0
Game	83.3	29.0	34.7	50.3	69.7	47.0
Daily	81.3	31.3	31.3	63.0	83.7	28.7
All	**80.2**	29.8	32.8	52.6	74.6	34.4
GPT-4o-mini						
CS	85.0	37.3	34.7	56.0	60.3	33.7
Net	80.3	27.7	31.7	50.3	59.7	29.3
Fin	66.3	25.0	27.7	35.3	39.7	31.3
Legal	72.3	22.3	29.3	39.0	45.7	35.7
Bio	77.0	29.0	37.7	43.3	54.7	31.0
Eng	72.3	26.7	30.7	41.3	46.3	26.0
Biz	70.3	23.7	29.7	38.0	46.0	36.0
Social	66.3	22.0	25.0	33.3	46.0	25.7
Game	64.0	25.3	34.0	37.0	40.3	34.0
Daily	83.7	32.3	30.7	56.7	63.7	28.3
All	**73.8**	27.1	31.1	43.0	50.2	31.0

Table 19: Supplementary per-domain Accuracy for the 128k cross-backend comparison. Hybrid denotes Hybrid-RRF Chunk RAG. Domain abbreviations: CS = software, Net = network/hardware, Fin = finance, Bio = biomedicine, Eng = engineering, Biz = business operations.

Stress view	N	Std. RAG	Hybrid-RRF	RAPTOR	MemGPT	HippoRAG	TSIM
All audited	300	26.7/5.7	29.0/9.3	41.7/29.7	52.7/58.3	29.3/8.7	**73.7/78.7**
State overwrite	196	24.0/3.6	27.0/8.2	38.3/26.5	49.5/56.6	28.6/7.1	**70.9/76.0**
Long-range bridge	291	26.8/5.8	28.9/8.9	42.3/29.9	52.9/58.1	29.6/8.6	**73.9/78.4**
Constraint trap	300	26.7/5.7	29.0/9.3	41.7/29.7	52.7/58.3	29.3/8.7	**73.7/78.7**

Table 20: Full stress-type method audit on the 300-example three-annotator human-audited subset under GPT-4o-mini at 128k. Cells report Accuracy / CL Hit. Stress views are multi-label, so counts are not expected to sum to 300. Full Context is omitted because only 11 audited items have matched rows in this diagnostic slice.

Budget	Hit	Acc	HitRank	Lat.	CtxTok
16k	100.00	62.5	335.4	4.32	29,938
32k	100.00	57.8	669.2	4.57	60,145
64k	100.00	41.7	836.3	10.03	86,315
128k	100.00	29.8	711.7	14.61	88,425

Table 21: Full Context sanity audit for Figure 4. Hit is evidence containment in the constructed prompt. The collapse is therefore not explained by missing gold evidence alone.

context commercial evaluation, and GPT-4o-mini for high-throughput closed-loop comparison and ablation.

Prompt-robust Full Context diagnostic. Because Full Context can depend on prompt structure, we ran an independent no-writeback GPT-4o-mini diagnostic on a stratified subset. A 20-example development split selected an evidence-first prompt from vanilla, evidence-first, and question-first variants using a fixed accuracy-maximization rule; the selected prompt was then applied unchanged to a disjoint 100-example evaluation split, with all three systems evaluated on the same 100 examples. All Full Context variants use the same visible-history window for each example, so differences reflect prompt structure rather than context selection. We focus on three structurally distinct prompt templates—vanilla baseline, evidence-first instruction, and question-first instruction—rather than exhaustively varying decoding strategies, because the diagnostic targets whether prompt structure alone can recover Full Context performance.

Token accounting separates constructed context length, a benchmark-side packing target estimated as $\text{tokens} \approx 1.3\times$ whitespace words, from retrieved-context tokens, the compact prompt fragments passed after memory selection. We use constructed tokens only to build identical text packages, not to claim provider-native tokenizer parity across Gemma, OpenAI, Gemini, and DeepSeek.

System	Prompt	Rows	Acc.	Prompt tok.	Lat.
Full Context	Vanilla	100	49.0	98,339	5.52
Full Context	Dev-selected	100	49.0	98,298	2.80
TSIM	–	100	77.0	1,389	2.40

Table 22: Prompt-robust Full Context diagnostic under GPT-4o-mini in an independent no-writeback setting. The dev-selected prompt is evidence-first, chosen on a disjoint 20-example development split and frozen before evaluation. Rows are the same stratified 100-example evaluation subset with balanced domains, answer labels, and logical-context quartiles. Prompt tokens and latency are provider-reported averages; TSIM uses its standard memory pipeline rather than a Full Context prompt template. Numbers are not directly comparable to the main stateful results in Table 2 because this diagnostic uses a 100-example subset under a no-writeback evaluation regime; it isolates prompt-template effects on Full Context rather than re-evaluating system rankings.

Backend	Method	Acc	Hit-C	Hit-W	Miss-C	Miss-W
Gemma2:9b	Standard RAG	24.4	5.2	2.1	10.6	47.7
	Hybrid-RRF Chunk RAG	31.1	8.4	2.1	9.5	27.2
	RAPTOR	40.6	25.2	10.0	14.6	47.0
	MEMGPT	60.1	48.6	14.0	11.4	25.9
	HIPPORAG	25.2	8.7	3.7	15.6	68.4
	TSIM	69.6	58.2	12.4	11.3	17.7
Gemini 2.5 Flash	Standard RAG	29.8	6.1	0.2	0.5	0.3
	Hybrid-RRF Chunk RAG	32.8	10.0	0.2	0.7	0.6
	RAPTOR	52.6	32.1	1.5	6.2	3.2
	MEMGPT	74.6	61.5	0.5	5.2	1.0
	HIPPORAG	34.4	11.4	0.3	1.8	1.5
	TSIM	80.2	66.5	1.2	8.4	2.8
GPT-4o-mini	Standard RAG	27.1	5.0	1.4	5.3	21.0
	Hybrid-RRF Chunk RAG	31.1	8.7	1.7	1.8	5.5
	RAPTOR	43.0	26.0	8.9	14.6	40.7
	MEMGPT	50.2	36.4	9.4	1.5	3.3
	HIPPORAG	31.0	7.7	2.5	3.6	7.3
	TSIM	73.8	63.5	10.4	8.4	10.2

Table 23: Backend-output audit for the 128k main comparison. Hit-C/W and Miss-C/W split closed-loop evidence hit by correct and wrong answers, all in percentages.

Backend	Baseline	Base	TSIM	Δ	95% CI
Gemma2:9b	Standard RAG	24.4	69.6	+45.21	[43.48, 47.03]
	Hybrid-RRF Chunk RAG	31.1	69.6	+38.52	[36.72, 40.24]
	RAPTOR	40.6	69.6	+28.98	[27.13, 30.76]
	MEMGPT	60.1	69.6	+9.53	[7.90, 11.22]
	HIPPORAG	25.2	69.6	+44.42	[42.41, 46.36]
Gemini 2.5 Flash	Standard RAG	29.8	80.2	+50.33	[48.98, 51.63]
	Hybrid-RRF Chunk RAG	32.8	80.2	+47.34	[45.88, 48.73]
	RAPTOR	52.6	80.2	+27.59	[26.02, 29.19]
	MEMGPT	74.6	80.2	+5.58	[4.15, 6.97]
	HIPPORAG	34.4	80.2	+45.72	[44.24, 47.23]
GPT-4o-mini	Standard RAG	27.1	73.8	+46.67	[45.04, 48.32]
	Hybrid-RRF Chunk RAG	31.1	73.8	+42.67	[41.03, 44.23]
	RAPTOR	43.0	73.8	+30.77	[29.04, 32.53]
	MEMGPT	50.2	73.8	+23.57	[22.00, 25.12]
	HIPPORAG	31.0	73.8	+42.73	[41.16, 44.34]

Table 24: Paired bootstrap audit for the final 3,000-question main-table runs. The closest final API-backed comparison is Gemini 2.5 Flash with MEMGPT; the improvement remains significant under paired resampling.

Method	Domain Acc Std.	Batch Acc Range	Ctx Mean	Ctx Median	Ctx P95	Ctx Max
Standard RAG	3.73	20.4–30.0	929.5	937	1076	1259
Hybrid-RRF Chunk RAG	3.65	25.0–43.2	1193.6	1194	1200	1200
RAPTOR	7.62	37.4–45.5	790.5	780	996	1219
MEMGPT	7.70	52.6–72.7	2301.9	2317	2608	3359
HIPPORAG	3.97	20.8–29.7	748.9	757	891	1043
TSIM	8.08	64.5–76.9	1060.8	1037	1408	1950

Table 25: Variance and retrieved-context token audit for the final Gemma2:9b main-table run. Context-token columns summarize selected answer context, not benchmark-side constructed length.

Rationale similarity diagnostic. While Accuracy and CL Hit are primary, Table 2 also reports rationale similarity as a secondary diagnostic. The evaluator parses each output into an answer choice and rationale, embeds the predicted rationale and gold $q[\texttt{reasoning}]$ with Sentence-Transformers (Reimers and Gurevych, 2019), and computes cosine similarity. Reported runs use `BAAI/bge-large-en-v1.5`; `all-MiniLM-L6-v2` is retained only as a portability fallback. Missing rationales or evaluator failures receive 0, and table values average per-example scores over the completed full-dataset run. This column diagnoses semantic alignment with the gold evidence rationale only; it is not used for ranking or calibration.

Reproducibility artifacts. The supplementary material includes the dataset card, protocol, validation report, exact-evidence audit, deterministic runtime builder, per-run manifests, and raw JSON/CSV outputs with predictions, evidence hits, context-token estimates, and latency. Run manifests record the local Gemma2:9b 128k configuration and the `openai/gpt-4o-mini` full-dataset ablation settings.