Hashing Is Not Retrieval: A Three-Arm Study of Context Accumulation in Small-Model Code Cartography

Anonymous Author(s)
Affiliation
Address
email

Abstract

Multi-agent language-model systems repeatedly retransmit state, but a shorter prompt can mean either useful state externalization or simple information deletion. We present a prospectively frozen three-arm evaluation that separates those mechanisms in a two-worker code- cartography harness. This is deliberately a bounded-context state-passing study, not a test of a model's advertised million-token context capacity. FULLHISTORY includes all prior worker outputs; OPAQUEREF replaces the previous output with an unretrievable SHA-256 digest; and RETRIEVED round-trips the previous output through a tenant-scoped content-addressed store before injecting the resolved bytes. All arms receive identical source-derived evidence. Every request logs the exact transmitted message-content token count, attempted and resolved model/provider identity, generation identifier, usage, cost, retry disposition, and retrieval proof. A five-task gate preceded a fixed 50-task controlled benchmark and a five-scale Medusa case study. On the controlled benchmark, RETRIEVED reduces paired peak context by 38.3% (task bootstrap 95% interval 35.2–41.4%; domain-cluster sensitivity 36.0–40.6%) relative to FULLHISTORY. Its structural-score advantage is an output-contract effect, not superior reasoning. The nested Medusa scales show a descriptive 38.1% context reduction, but strict structural accuracy is zero in every arm. The complete gated study used 871 paid attempts at an accounted v2 cost of $0.1729, with no model substitutions or accepted identity mismatches. These results support a bounded-context claim for orchestrator-resolved state, not a universal claim about autonomous memory: the real-code evidence comes from one repository and retrieval is automatic rather than worker-initiated.

After the primary study, we prospectively froze a separate state-required follow-up: four non-overlapping evidence packets must be accumulated to construct the final eight-module map. Across 24 deterministic fixtures, opaque references collapse to 0.0% strict exact maps (25.0% structural rate), whereas verified retrieval reaches 25.0% (78.6%), below full history (83.3%; 96.9%) but far above deletion. Thus the follow-up establishes state necessity while also showing that one-step retrieval is not automatically quality-equivalent to full history.

1 Introduction

Orchestrator–worker systems make language-model computation compositional, but their communication can grow with every turn. Frameworks such as AutoGen make multi-agent conversation easy to express [Wu et al., 2023]; they do not make accumulated context free. Repeated history increases input cost, approaches finite context windows, and may even reduce effective use of evidence placed

Submitted to NeurIPS 2026 Workshop on Long Context Foundation Models. Do not distribute.

in the middle of a prompt [Liu et al., 2024]. Communication pruning [Zhang et al., 2024a], prompt compression [Jiang et al., 2023], and external memory [Packer et al., 2023, Chhikara et al., 2025] address related problems, but they intervene at different points. A communication edge can be removed, a prompt can be compressed, or state can be stored and later recovered. Conflating the three makes an experiment hard to interpret.

Long-context benchmarks test whether models can reason over much longer inputs [Bai et al., 2023, Hsieh et al., 2024, Zhang et al., 2024b]. We do not treat our roughly 3.5K-token peak prompts as a proxy for 128K–1M-token capability: the ceiling intentionally isolates the state interface as a trace grows. Large-window scaling is outside the estimand.

This distinction matters for content-addressed state. Replacing a message with a digest will almost always shrink a prompt. If the receiver cannot resolve that digest, however, the operation is deletion rather than retrieval. A small prompt in such a condition does not show that an artifact architecture preserved information. Conversely, simply resending resolved content may preserve information without reproducing the quadratic growth of complete history. A clean evaluation needs both controls.

We study this question in MERIDIAN-RESEARCH, a local Express/Node.js harness that alternates two open-weight workers on structured repository-mapping tasks. The study was motivated by a forensic audit of earlier internal runs: curves had included unsent states, one batch silently substituted a worker after provider failures, source evidence was asymmetric, and hashes had been described as artifacts although workers could not dereference them. Those historical outputs are not pooled into the present estimates. Instead, the current study freezes a new protocol before paid dispatch, makes model and provider identity an acceptance condition, and distinguishes an opaque-reference ablation from actual byte recovery.

Our contributions are:

- a three-arm design that identifies prompt accumulation, information removal, and orchestrator-resolved content-addressed state separately;
- an executable measurement contract in which $C(t)$ is computed from the exact system and user strings sent, with one curve point and one immutable ledger entry per upstream request;
- a source-grounded benchmark comprising 50 controlled tasks and a pinned Medusa case study, with path-copying separated from stricter structural evaluation; and
- a fully gated live run whose protocol, prices, source hashes, attempted calls, outputs, and analysis are released together; and
- a separately frozen state-required follow-up that tests the mechanism under non-overlapping evidence rather than inferring continuity from repeated evidence.

The key boundary is equally important: this is evidence for automatic, orchestrator- resolved one-step state passing. It is not evidence that a worker can decide when or what to retrieve, and five nested Medusa scales are not five independent repositories.

2 Related Work

Multi-agent communication. Multi-agent LLM systems coordinate specialized roles through message histories and communication graphs [Wu et al., 2023]. AgentPrune formalizes redundant spatial and temporal edges and reports substantial token reductions while retaining task performance on its evaluated benchmarks [Zhang et al., 2024a]. Our question is narrower: rather than learn a graph, we hold the worker schedule and task evidence fixed and manipulate only how prior outputs enter the next request. This yields a directly measurable state- passing contrast but does not optimize the overall graph.

Prompt compression and long context. LLMLingua removes low-importance prompt tokens to accelerate inference [Jiang et al., 2023]. Work on long-context use shows that nominal context capacity does not guarantee robust access to information at every position [Liu et al., 2024], and benchmarks such as LongBench, RULER, and ∞Bench test substantially larger inputs [Bai et al., 2023, Hsieh et al., 2024, Zhang et al., 2024b]. MERIDIAN-RESEARCH performs neither learned token compression nor truncation. It compares full history against two explicit state interfaces, and

reports both the message-content proxy and provider usage so peak context is not mistaken for total cost or for a long-window capability measurement.

External memory. MemGPT treats context as a hierarchy whose contents move between memory tiers [Packer et al., 2023]; Mem0 extracts and retrieves salient state rather than retaining an entire dialogue [Chhikara et al., 2025]. Recent benchmarks evaluate memory effectiveness, efficiency, and capacity across richer interactive settings [Tan et al., 2025], while MemInsight and Memory-R1 respectively augment or learn memory management [Salama et al., 2025, Yan et al., 2026]. Mem2ActBench further tests whether an agent proactively applies long-term memory to tool-mediated tasks [Shen et al., 2026]. These systems and evaluations are more capable than our mechanism. Our content-addressed store uses a complete SHA-256 digest [National Institute of Standards and Technology, 2015]; the orchestrator immediately resolves the prior output and places the byte-identical content in the next prompt. There is no learned salience model, persistent cross-session memory, or worker-issued tool call. CoALA taxonomizes agent memory across working, episodic, semantic, and procedural types over a broad design space [Sumers et al., 2023]; our three conditions instead isolate a single axis within that space—whether a reference the agent already holds is actually resolvable. Retrieval- augmented generation retrieves passages from an external corpus to ground generation [Lewis et al., 2020]; our retrieval arm instead resolves a content-addressed digest of the agent's own immediately prior output in one deterministic orchestrator-side step, with no learned salience and no worker-issued tool call. Thus this paper complements, rather than competes with, active-memory work: it supplies a minimal control for determining whether an observed prompt reduction represents recoverable state or information removal.

Reliability of multi-agent evaluation. MAST categorizes multi-agent failures across design, inter-agent alignment, and verification [Cemri et al., 2025]. The present study treats several normally incidental properties—provider routing, retry behavior, parsing, and prompt assembly—as part of the scientific treatment. This is especially important when hosted routing can otherwise serve a different endpoint. We set provider order explicitly and disable fallbacks using the hosted API's routing contract [OpenRouter, 2026].

3 Research Questions and Conditions

For task i, condition k, and upstream request t, let s_i be the system objective and $u_{i,k,t}$ the assembled user string. With one fixed proxy tokenizer T,

$$\begin{aligned} C_{i,k}(t) &= T(s_i) + T(u_{i,k,t}), \\ C_{i,k}^{\max} &= \max_t C_{i,k}(t). \end{aligned} \tag{1}$$

This message-content measure excludes provider-specific chat-template framing. The same tokenizer is used in the gateway span and experiment driver, while provider-reported input and completion tokens are retained as a separate operational measure. For treatment k relative to FULLHISTORY, the paired peak reduction is

$$r_{i,k} = 1 - \frac{C_{i,k}^{\max}}{C_{i,\text{full}}^{\max}}. \tag{2}$$

We ask: **RQ1**, how much do the two bounded-state interfaces reduce peak transmitted context and total provider tokens? **RQ2**, does verified retrieval retain source-grounded structural quality relative to full history? **RQ3**, what does the opaque arm reveal about gains caused by information removal rather than recovery?

Every arm receives the complete, identical source-evidence packet at every step. They differ only in prior-output state:

FULLHISTORY Evidence followed by every previous worker output, labeled by step.
OPAQUEREF Evidence followed by the complete SHA-256 digest of the previous output. The prompt explicitly states that retrieval is unavailable.

RETRIEVED Evidence followed by the previous output after it has been stored, addressed, and resolved in the same run-scoped store. The digest and exact resolved bytes are both transmitted.

At step one no prior output exists; each arm therefore contains evidence plus its own minimal first-step framing. In later steps, OPAQUEREF approximates deletion of the prior answer, whereas RETRIEVED retains only the immediately preceding answer. Neither arm receives gold labels through routing.

4 Apparatus and Protocol

4.1 Workers and dispatch

Odd steps use Qwen2.5-7B-Instruct and even steps use Llama-3.1-8B-Instruct [Qwen Team, 2025, AI at Meta, 2024]. The models are server-pinned; the request cannot choose a model slug. Live preflight verified Qwen on Together and Llama on DeepInfra. For every call, provider fallbacks were disabled, temperature was zero, the requested seed was 20,260,720, and completion length was capped at 800 tokens. A returned output was accepted only when both the resolved model slug and provider string exactly matched the request. Protocol v2 permits at most two retries of the *same* prompt on the same worker/model/provider, and only after HTTP 429, HTTP 5xx, network, or timeout failures. Fixed backoffs are 5 and 10 seconds, with at least 500 ms between attempt starts; every failed attempt remains in the ledger. There are no whole-trial retries. Exhausted transport failure, empty output, identity mismatch, unverified retrieval, missing source packet, or incorrect fixed-step count aborts the study rather than substituting a model or selectively replacing a completed observation.

An initial zero-retry feasibility execution stopped, as intended, when DeepInfra returned HTTP 429 on upstream attempt 176. It is retained but excluded from all estimates. Before restarting from task one, v2 froze the same-worker transient retry rule above; no task, condition, decoding, metric, or quality threshold changed. Reporting this history avoids silently omitting an inconvenient operational failure.

The full study alternates workers for four to six fixed steps per task. Fixed length prevents a formatting-dependent `TASK_COMPLETE` marker from changing exposure across conditions. Three condition orders are rotated deterministically across tasks, and tasks are ordered by a SHA-256 ranking fixed with the same study seed. Because the hosted provider may not guarantee bitwise determinism even when a seed is requested, the generation identifier and complete output are retained for every call.

4.2 Content-addressed state and isolation

Artifacts are partitioned by run identifier in an in-process store. Storing content computes its full 64-hex-character SHA-256 digest. RETRIEVED resolves that digest in the same tenant and refuses a miss; the trace records whether resolved content equals the previous output. Cross-tenant lookup is impossible through the store interface. This is functional state recovery from the downstream model's perspective because the resolved bytes reach its prompt. It is nevertheless *orchestrator-resolved*: the worker does not issue a retrieval action, and the store is not durable across process restarts.

4.3 Telemetry and budget guard

The driver constructs the exact two message strings and computes Eq. 1 immediately before dispatch. The gateway independently counts those same strings. Each attempt records the run and experiment IDs, logical step, attempt index, requested and resolved identities, generation and span IDs, context hash, evidence hash and occurrence count, artifact digest, retrieval result, usage, termination reason, and cumulative budget. A retried prompt creates another curve point with the same $C(t)$ and a distinct attempt index. Curves contain no pre-dispatch rejection or post-final phantom point.

Current endpoint prices were fetched from the provider listing before the protocol was frozen: Qwen/Together cost \$0.30 per million input or output tokens; Llama/DeepInfra cost \$0.02 input and \$0.03 output. The aggregate ledger stopped a request before its worst-case completion could

exceed either 1.5 million v2 study tokens or $0.45. The smaller dollar ceiling reserves headroom for the excluded $0.0344 feasibility execution, keeping total authorized spend below $0.50. The guard charged against the greater of locally estimated and provider-reported cumulative cost.

5 Tasks, Outcomes, and Analysis

5.1 Controlled benchmark

The controlled set contains ten software domains—including web, game-engine, data- pipeline, commerce, chat, IoT, ML-serving, and build-system templates—at five nested scales of three to seven modules, for 50 tasks. Each module fixture contains deterministic imports and exported declarations; gold structure is generated mechanically from the same template specification, without using an LLM. This design makes source-grounded exports and within-set dependencies answerable, but the fixtures are deliberately simple and the five scales within a domain are correlated.

5.2 Pinned Medusa case study

The real-code case uses Medusa [MedusaJS, 2026] commit `dfdcdd7467`; the full commit and file hashes are frozen in the released protocol. Ten TypeScript module-service files were copied from the clean snapshot and independently compared byte-for-byte; all ten SHA-256 hashes match. Five tasks expose nested sets of 4, 5, 6, 8, and 10 files. The evidence packet is mechanically extracted from source and includes exact paths, import specifiers, exported declarations, and public method-signature names. A hand-checked gold map scores exports and direct relationships within the selected set. These are five scales of one repository, not independent repository samples.

5.3 Quality and uncertainty

Only the final worker output is parsed for analysis. A tolerant JSON extractor removes the requested completion marker and accepts either an array or a `modules` array; it does not scan an earlier turn to rescue the final response. We report direct-parse rate, path coverage, strict exact rate, purpose-token overlap, and set F1 for exports, inbound dependencies, and outbound dependencies. The primary quality measure is mean module-level *structural exact rate*: exports and both dependency sets must exactly match gold. Paths alone cannot constitute success.

The frozen protocol prospectively declared a descriptive non-inferiority margin of -0.10 for the paired RETRIEVED minus FULLHISTORY structural rate. We report means and deterministic 10,000-resample paired percentile intervals [Efron and Tibshirani, 1986]. Two-sided paired sign-flip tests use exact enumeration for at most 20 units and 100,000 fixed-seed draws otherwise. The task is the prospectively fixed pairing unit. Because controlled scales are nested, we also report a sensitivity interval that resamples the ten domains as clusters. Medusa intervals summarize its five fixed scales only and are not population intervals over repositories. The five-task pilot was an apparatus gate and is excluded from every effect estimate.

6 Results

The complete run contains 792 full-study logical requests across 165 topology runs, in addition to the 72-request pilot and two identity probes. Integrity checks require exact source evidence once per prompt, unique accepted generation IDs, the requested model/provider on every accepted response, only prospectively allowed transient reasons before a retry, and verified byte recovery in every applicable RETRIEVED step. The v2 aggregate ledger reports 1,033,568 tokens and $0.1729 accounted cost for all v2 gated attempts. The excluded feasibility execution cost $0.0344, making the combined accounted spend $0.2073. Within v2, 4 logical requests required recovery, contributing 5 failed transient attempts before an accepted response.

6.1 Context accumulation

On the controlled set, RETRIEVED reduces paired peak message content by 38.3% relative to FULLHISTORY (task bootstrap 95% interval 35.2–41.4%, sign-flip $p = 0.00001$). The domain-cluster

Table 1: Full-study results. Means are over 50 controlled tasks or five nested Medusa scales. Cost is per topology run. Medusa uncertainty is descriptive because all scales come from one repository.

Data	Condition	$C_{\max}$	Structural	Path coverage	Provider tokens	Cost/run
Controlled ($n = 50$)	FULLHISTORY	1370.0	3.5%	10.0%	5656	$0.00094
	OPAQUEREF	537.0	39.9%	92.0%	4050	$0.00068
	RETRIEVED	808.2	16.8%	54.0%	4927	$0.00082
Medusa (one repo, five scales)	FULLHISTORY	3562.2	0.0%	20.0%	15112	$0.00252
	OPAQUEREF	1625.2	0.0%	60.0%	10534	$0.00182
	RETRIEVED	2156.0	0.0%	80.0%	12528	$0.00210

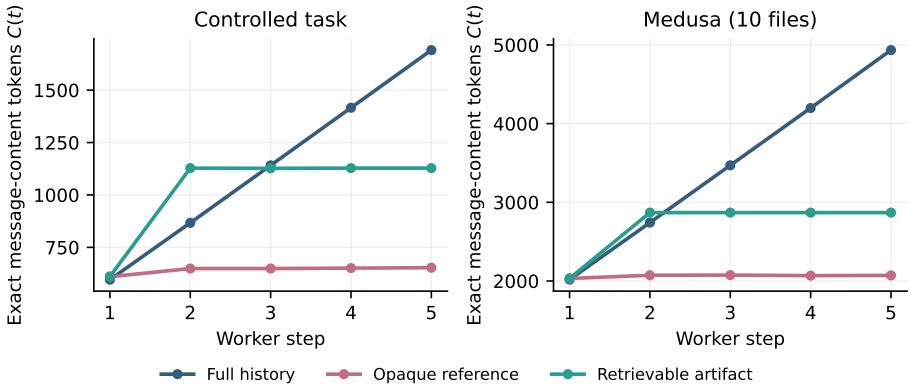

Figure 1: Exact sent-message content curves for one maximum-scale controlled task and the ten-file Medusa scale. Each point corresponds to one accepted upstream request; no post-final state is plotted.

sensitivity interval is 36.0–40.6% with $p = 0.00195$, so the result is not an artifact of treating five scales in one domain as unrelated. OPAQUEREF is smaller still or comparable, with a paired 58.5% reduction [56.2–60.8%]. This difference is expected: OPAQUEREF sends a digest but no previous answer, whereas RETRIEVED pays to transmit one resolved answer.

The Medusa scales show a descriptive 38.1% retrieved-state reduction [31.5–44.0%] and 52.9% opaque-reference reduction [47.3–58.0%]. Figure 1 illustrates the mechanism. Full history grows as answers accumulate; opaque references remain close to the source-packet floor; retrieved state is bounded by the packet plus one prior answer. Peak context is not identical to billable tokens, so Table 1 reports provider totals and cost separately.

6.2 Source-grounded quality

Controlled structural rates are 3.5%, 39.9%, and 16.8% for full, opaque, and retrieved conditions; their path coverages are 10.0%, 92.0%, and 54.0%. RETRIEVED minus FULLHISTORY has a paired structural difference of 0.133 [0.063–0.212]. On Medusa the corresponding rates are 0.0%, 0.0%, and 0.0%, with a retrieved-minus-full difference of 0.000 [0.000–0.000]. These direct, field-level outcomes determine whether the predeclared -0.10 descriptive margin is met; a short prompt alone never does.

The controlled quality difference should not be read as improved semantic reasoning. It is driven mainly by final-output contract adherence: direct parse succeeds for 10.0% of full-history, 92.0% of opaque, and 54.0% of retrieved controlled tasks. Among directly parsable controlled outputs, the post-hoc mean structural rates are 34.7%, 43.4%, and 31.1%, respectively. Thus the primary end-to-end score—which correctly penalizes unusable final output—favors retrieval over full history, while the parsable-only diagnostic does not show an intrinsic reasoning advantage. All three Medusa arms have zero strict structural accuracy, a floor that makes its nominal non-inferiority result uninformative. Medusa path coverage and parse rates remain useful diagnostics, but do not rescue the strict quality endpoint.

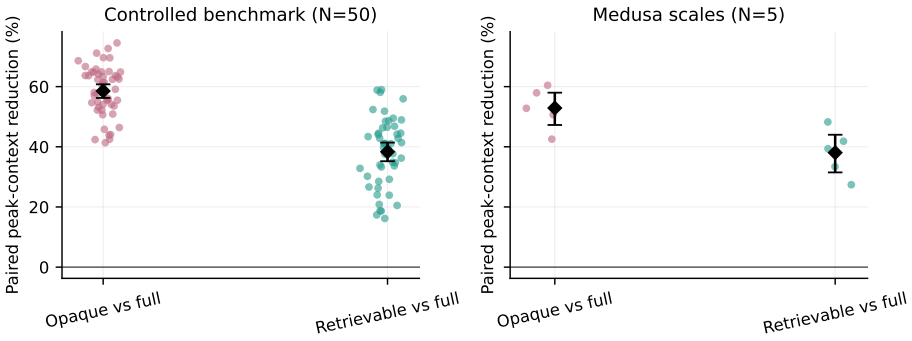

Figure 2: Task-level paired peak-context reductions. Diamonds show means and bootstrap intervals. Medusa points are nested scales of one repository and are descriptive.

The 800-token completion ceiling is also audited rather than treated as an invisible provider detail. Accepted controlled responses hit the ceiling on 0.0% of full-history, 0.0% of opaque, and 0.0% of retrieved steps; the corresponding Medusa rates are 0.0%, 26.0%, and 20.0%. These rates help distinguish state-condition effects from malformed final JSON caused by generation truncation. Direct final-output parse rates are listed above for controlled tasks; on Medusa scales they are 20.0%, 60.0%, and 80.0% (full/opaque/retrieved).

The opaque arm must be interpreted carefully. Identical source evidence appears at every step, so a worker can reconstruct the requested map without any prior answer. Strong opaque quality therefore means that prior conversational state was redundant for these tasks, not that a digest carried its semantics. Retrieved-state quality is the relevant test of bounded continuity, while the gap between opaque and retrieved arms estimates the cost of retaining one answer.

6.3 Post-primary state-required follow-up

The primary tasks repeat the source packet, so their opaque arm is an information-removal control but not a test of whether state is needed. We therefore froze a new, separate protocol after the primary run. Each of 24 deterministic fixtures exposes two of eight modules in each of four non-overlapping packets; only the final turn is scored against the complete map. The final sample excludes two prespecified apparatus pilots. Full history reaches 83.3% strict exact and 96.9% structural accuracy. Removing state by an opaque hash drops these outcomes to 0.0% and 25.0%, whereas byte-verified retrieval reaches 25.0% and 78.6%. Retrieval's structural difference from full history is -18.2% [-24.5% to -11.5%]. This descriptive, one-model fixture result eliminates the central alternative explanation: a hash-only baseline is not retrieval, and deletion is harmful when earlier evidence is actually necessary. It also tempers the original bounded-context message: retrieving only the previous map retained much of the required state, not all of full-history quality.

7 Discussion

What the three arms identify. A two-arm full-versus-hash experiment cannot distinguish state externalization from state deletion. Here, OPAQUEREF exposes the smallest prompt attainable when prior outputs are discarded, while RETRIEVED tests whether keeping one byte-identical answer bounds growth. The result supports an implementation claim: for this alternating-worker mapping protocol, the next prompt need not contain the complete transcript to retain the immediate state represented by the prior answer. It also exposes a task-design result: when the complete source packet is repeated, prior prose can interfere with the final serialization contract. The opaque arm's high score is evidence of redundancy and interference in this benchmark, not evidence that a digest encodes useful state.

Peak context is not total cost. $C_{\max}$ characterizes capacity pressure at the largest request. Provider totals add every input and completion, and the two workers have different endpoint prices. Reporting all three prevents a large peak reduction from being presented as the same percentage

cost saving. The hard budget also makes failed and successful requests part of one experiment ledger rather than an untracked operational detail.

Reliability is part of treatment fidelity. Model substitution, fallback routing, retry filtering, and post-hoc parse rescue can all change the observed sample. The present run makes these properties falsifiable. Every accepted output joins to a generation ID and a model/provider pair; every attempted request appears once; every retrieval joins to a full digest and equality check. This does not eliminate hosted-provider variance, but it prevents an unnoticed endpoint change from masquerading as a topology effect.

Implications for agent design. Automatic last-answer retrieval is useful when a worker produces a self-contained state artifact each turn. More open-ended agents will need selective retrieval, durable stores, schema evolution, access control, and recovery from stale or conflicting artifacts. Learned pruning and memory systems remain complementary: they decide what to retain; content addressing supplies identity and integrity for retained bytes.

8 Limitations and Threats to Validity

First, the 50 controlled tasks are deterministic templates. Their evidence closely matches the scored structure by design, and five scales within each domain are nested. Task-level intervals describe this benchmark; domain-cluster sensitivity is more conservative but still covers only ten templates. Second, Medusa contributes one repository and ten related service files. Its five scales measure within-repository growth and cannot establish cross-project generality.

Third, the Medusa packet exposes extracted signatures and imports rather than complete source bodies. The study evaluates bounded code cartography, not arbitrary program comprehension, execution, or hallucination resistance. Fourth, only two small instruction- tuned models and two hosted providers are used. Worker position, model, and provider are balanced across conditions but cannot be disentangled from one another. A requested seed does not guarantee deterministic hosted inference. The largest transmitted prompt is only roughly 3.5K tokens; this is intentional for isolating state-passing pressure, but it cannot support claims about 128K–1M-token long-context capability or scaling.

Fifth, RETRIEVED is orchestrator-triggered and one-step. It demonstrates byte recovery and bounded prompt assembly, not autonomous retrieval policy, long-term memory, or crash durability. The in-process store disappears on restart. Sixth, $C(t)$ uses a fixed `gpt-4o-mini` tokenizer over message content for cross-model comparability; it omits native tokenization and hidden chat-template tokens. Provider usage is therefore reported alongside it. Finally, the non-inferiority margin is descriptive and the study was not powered from prior variance. Passing it on the controlled tasks motivates broader replication, not universal equivalence; the all-zero Medusa structural endpoint cannot validate non-inferiority despite mechanically satisfying the margin.

The fixed-step protocol also asks for a complete map at every turn rather than assigning different modules or tools to specialized agents. This deliberately isolates the cost of retransmitting state, but it makes earlier outputs potentially redundant and should not be read as a benchmark of sophisticated collaborative planning. Future work should combine the same three-arm measurement contract with genuinely interdependent subtasks. The post-primary state-required fixtures improve that continuity diagnosis but remain deterministic and use the same two hosted worker roles; they do not provide independent repository or model-family generalization.

Conclusion. Content hashes identify state but do not make it retrievable. Our three-arm design separates full history, deletion-by-reference, and verified recovery. In MERIDIAN-RESEARCH, automatic recovery cuts peak context by 38.3% and total provider tokens by 10.1% on controlled tasks; its apparent structural gain reflects JSON-contract adherence, not superior reasoning. The Medusa case supports only bounded context, while the state-required follow-up confirms that deletion fails when prior evidence is necessary but retrieval remains below full history. Broader repositories, model families, and worker-initiated retrieval policies are the next test.

A Reproducibility, Ethics, and Artifact Statement

The frozen protocol records the task files, source hashes, apparatus hashes, prices, provider pins, condition order, stop rule, seed, and analysis plan before the first paid dispatch. Incremental checkpoints preserve partial work, while the completed `runs.json` contains full outputs and request ledgers. `analysis.json` is the sole numerical source for the generated LaTeX macros and figures. The gateway suite has 57 of 57 passing tests. The eight Meridian core-CS evaluation suites – `ablation`, `audit_existing_results`, `drift_validation`, `pipeline_e2e`, `pipeline_e2e_real_task`, `run_experiment`, `task_schema`, and `verify_chaos` – have 60 of 60 passing tests, including regression tests for exact message counting, evidence parity, no cross-model substitution, complete SHA-256 digests, tenant isolation, strict identity, budget accumulation, one-step retrieval, and direct structural scoring.

The study uses MIT-licensed open-source code and deterministic synthetic fixtures. It contains no personal data. API credentials remain environment-only and are absent from traces. The primary ethical risk is overclaiming from a convenient benchmark. We therefore expose the pilot, attempt ledger, nested-task structure, automatic nature of retrieval, and all failed gates rather than presenting only favorable aggregate numbers.

References

AI at Meta. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL `https://arxiv.org/abs/2407.21783`.

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

Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, Kurt Keutzer, Aditya Parameswaran, Dan Klein, Kannan Ramchandran, Matei Zaharia, Joseph E. Gonzalez, and Ion Stoica. Why do multi-agent LLM systems fail? *arXiv preprint arXiv:2503.13657*, 2025. URL `https://arxiv.org/abs/2503.13657`.

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. URL `https://arxiv.org/abs/2504.19413`.

Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986. doi: 10.1214/ss/1177013815.

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

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. LLMLingua: Compressing prompts for accelerated inference of large language models. *arXiv preprint arXiv:2310.05736*, 2023. URL `https://arxiv.org/abs/2310.05736`.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. *arXiv preprint arXiv:2005.11401*, 2020. URL `https://arxiv.org/abs/2005.11401`.

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

MedusaJS. Medusa: Building blocks for digital commerce. Software, commit dfdcdd7467ede40e1bb80ce866bfe2c256b8ff86, 2026. URL `https://github.com/medusajs/medusa`.

National Institute of Standards and Technology. Secure hash standard (SHS), FIPS PUB 180-4. Technical report, U.S. Department of Commerce, 2015.

OpenRouter. Provider routing. Documentation, 2026. URL https://openrouter.ai/docs/features/provider-routing. Accessed 2026-07-20.

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

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2025. URL https://arxiv.org/abs/2412.15115.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. MemInsight: Autonomous memory augmentation for LLM agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33136–33152, 2025. doi: 10.18653/v1/2025.emnlp-main.1683. URL https://aclanthology.org/2025.emnlp-main.1683/.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. Mem2ActBench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8173–8190, 2026. doi: 10.18653/v1/2026.acl-long.370. URL https://aclanthology.org/2026.acl-long.370/.

Theodore R. Sumers, Shunyu Yao, Karthik Narasimhan, and Thomas L. Griffiths. Cognitive architectures for language agents. *arXiv preprint arXiv:2309.02427*, 2023. URL https://arxiv.org/abs/2309.02427.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 19336–19352, 2025. doi: 10.18653/v1/2025.findings-acl.989. URL https://aclanthology.org/2025.findings-acl.989/.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023. URL https://arxiv.org/abs/2308.08155.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z. Pan, Hinrich Schuetze, 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 (Volume 1: Long Papers)*, pages 12805–12825, 2026. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

Guibin Zhang, Yao Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, Dawei Cheng, Jeffrey Xu Yu, and Tianlong Chen. Cut the crap: An economical communication pipeline for llm-based multi-agent systems. *arXiv preprint arXiv:2410.02506*, 2024a. URL https://arxiv.org/abs/2410.02506.

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