ECHO: A Cognitively Inspired, Auditable Memory Plane for Long-Horizon Agents

Yu Qian*, Hong Miao*, Boyang Guo*, Tingyi Jiang, Shan Zhao, Tianxing Le, Lintian Li, and Meng Liu

***Abstract*—Long-horizon agents need memory that identifies relevant experience, resolves revisions, and exposes checkable provenance. We present ECHO (Embodied Context & History Orchestration), an auditable memory architecture and service prototype inspired by episodic encoding, consolidation, contextual reinstatement, reconsolidation, and executive control. This is functional inspiration, not neural equivalence; the empirical analysis focuses on retrieval and context construction.**

Development runs reach 96.29% Hit@10 and 73.64% turn Recall@5 on 1,536 LoCoMo category-1–4 questions, and 97.60% Hit@10, 88.84% turn Recall@5, and 88.71% session Recall@5 on all 500 LongMemEval-S questions. A five-history BEAM gate fails, and in a separate matched 91-question QA sample Mem0 OSS scores 64.84% versus ECHO's 41.76% (exact McNemar $p = .00107$), with a history-cluster interval crossing zero. A post-hoc audit found source-specific phrases in the query-expansion rules. Although no gold answer field entered the runtime, expansion-enabled retrieval scores are therefore descriptive development measurements, not independent confirmation.

***Index Terms*—cognitive memory, agent memory, bitemporal data, evidence retrieval, provenance, long-context evaluation, memory systems**

I. Introduction

Consider an assistant that remembers a user's former address, the correction that replaced it, and the conversation in which the correction occurred. A nearest-neighbor store may retrieve both addresses. A useful memory system must do more: preserve the original experience, represent the revision without rewriting history, decide which value is valid for the requested time, and return the evidence that justifies the decision. This is the difference between storing a past and maintaining a past that can safely constrain the present.

Persistent assistants and embodied agents therefore need a memory *plane*, not merely a vector index. Such a plane must coordinate four separable responsibilities: durable experience capture, evolution of structured state, context-sensitive discovery, and controlled realization into an answer or action. A failure at any boundary can survive a favorable headline score. A retriever may return one relevant turn while omitting the other events needed for a count; a semantically close stale value may outrank the current revision; or a reader may receive the right evidence and still answer incorrectly.

Existing long-memory benchmarks expose several parts of this problem. LoCoMo tests multi-session factual, temporal, and multi-hop memory [1]; LongMemEval tests information extraction, knowledge update, temporal reasoning, multi-session synthesis, preference, and abstention over timestamped histories [2]; BEAM extends coherent histories toward millions of tokens and broad memory operations [3]. These benchmarks are valuable, but headline scores often mix readers, prompts, retrieval cutoffs, data revisions, and judges. A single Hit@k additionally hides how much annotated evidence was recovered and whether a memory API exposes the provenance needed to measure that coverage. Consequently, it may be impossible to tell whether a failure arose in candidate discovery, temporal resolution, evidence packing, or answer realization.

We study memory as auditable cognitive infrastructure. ECHO follows a continual loop of *encoding, organization, recall, and evolution*: immutable events retain episodes; projection builds typed revisions; hybrid routes propose candidates; a bitemporal ledger alone determines currentness; provenance closure restores the dependencies required by an operation; and an executive boundary separates internal derivation from the visible answer. The cognitive mapping motivates this decomposition but makes no anatomical, biological, or clinical claim.

The empirical results deliberately include negative evidence. LongMemEval-S shows that question hits and evidence coverage can be simultaneously high on a full frozen run. The one-history BEAM development pilot instead shows that a perfect Hit@10 can coexist with incomplete evidence. Most importantly, a fresh-history BEAM gate does not reproduce the pilot result, and a matched 91-question QA sample favors Mem0 OSS. These outcomes narrow rather than weaken the contribution: they distinguish what the architecture guarantees from what the current ranker and answerer have demonstrated.

This paper makes five contributions:

- We translate episodic traces, semantic consolidation, reconsolidation, contextual recall, and executive control into explicit, testable systems commitments, with a strict neural-equivalence disclaimer.
- We specify and unit-test a typed bitemporal authority component in which valid time, transaction time, revision state, provenance, and conflicts are explicit, while semantic similarity cannot determine factual currentness.
- We implement a worker-oriented retrieval pipeline and evaluate retrieval, provenance-closure, and bounded-context behavior at explicit system boundaries.
- We report frozen descriptive retrieval artifacts for LoCoMo and LongMemEval-S, a scoped BEAM development pilot, and a preregistered five-history BEAM gate that fails, retaining question-level audits, failures, latency,

*Yu Qian, Hong Miao, and Boyang Guo contributed equally to this work. All authors are with XDream Robotics, Building T2, Moli Community, No. 1, Lane 188, Yuren Road, Pudong New Area, Shanghai 200120, China (e-mail: {ethan,harold,ray,poluz,chloe,letianxing,lynk,meng.liu}@xdreamrobo.com).

context size, and hashes while disclosing the query-expansion contamination risk.
- We provide a protocol-matched 91-question ECHO–Mem0 OSS QA comparison with paired statistics and an observability contract that forbids assigning provenance metrics when a baseline does not expose source lineage.

The evidence is stage-separated throughout. ECHO retrieval, matched sampled QA, and author-reported product scores use different estimands; none is silently converted into another. The result is an architecture-and-audit study whose claims are aligned with the corresponding empirical boundaries and evaluation protocols; reported results remain scoped to those settings.

II. Problem Formulation

A. Events, States, and Two Time Axes

Let an interaction history be an append-only event stream $E = (e_1, \dots, e_T)$. Each event contains immutable source identity, speaker or actor, observed text, source order, and available timestamps. Projection maps events to typed propositions and state revisions. A state record is

$$s = (k, v, \tau_v, \tau_t, \rho, \pi, g, \sigma), \tag{1}$$

where k is a typed entity–relation key, v a value, τ_v its valid-time interval, τ_t the transaction or known-time interval, ρ the revision relation, π immutable provenance, g a ledger generation, and σ a status such as active, superseded, revoked, or unresolved. Valid time denotes when a proposition is true in the modeled world; transaction time denotes when the system records it. This separation follows established bitemporal database semantics [4].

For a query view $t = (t_v, t_t)$, ledger resolution is

$$\mathcal{S}(k,t) = \{s : s.k = k,\ t_v \in s.\tau_v,\ t_t \in s.\tau_t,\ s.g \le g_{\text{visible}}\}. \tag{2}$$

If $\mathcal{S}$ contains one supported active revision, that state can be answered. If it is empty, evidence is missing. If it contains incompatible unresolved revisions, the system must not pick the most similar one; it either discloses the minimal conflict set when requested or abstains.

We make that authority boundary explicit with a three-valued resolver

$$\text{Resolve}(k,t) = \begin{cases} s, & \mathcal{S}_{\text{adm}}(k,t) = \{s\}, \\ \perp_{\text{miss}}, & \mathcal{S}_{\text{adm}}(k,t) = \varnothing, \\ \perp_{\text{conflict}}, & |\mathcal{S}_{\text{adm}}(k,t)| > 1, \end{cases} \tag{3}$$

where admissibility requires compatible valid time, known time, revision status, provenance, and visible ledger generation. Semantic similarity is deliberately absent from Eq. (3); it may discover a state but cannot make that state authoritative.

B. Evidence and Answer Operations

A retrieval hit is not necessarily sufficient evidence. Let u be an evidence unit and $\Pi(u)$ its source lineage. For a selected state or derived result, the evidence closure $\Gamma(u)$ contains the originating events, required revision links, and dependencies needed to verify the result. Candidate discovery first unions route-specific results,

$$\mathcal{H}(q) = \bigcup_{r \in \mathcal{R}} \text{Top}_{M_r}\big(\mathcal{H}_r(q)\big), \tag{4}$$

$$F(u \mid q) = \sum_{r \in \mathcal{R}} \alpha_r \phi_r(u,q) + \lambda\ \text{Support}(\text{session}(u)), \tag{5}$$

where routes include lexical, semantic, typed-state, and neighborhood discovery. The fusion score F orders inspection only; currentness remains the output of Eq. (3). An admissible context is then an atomic-closure packing problem,

$$\mathbf{x}^\star = \arg\max_{\mathbf{x} \in \{0,1\}^{|\mathcal{H}|}} \sum_{u \in \mathcal{H}} x_u\, U(u \mid q, o), \tag{6}$$

$$\text{s.t.} \quad \sum_{u \in \mathcal{H}} x_u\ \text{tokens}(\Gamma(u)) \le B, \qquad C = \bigcup_{u : x_u^\star = 1} \Gamma(u), \tag{7}$$

for utility-conditioned evidence units $\mathcal{H}$ and budget B. A packer may defer an atomic closure that does not fit, but it cannot silently include only the conclusion while dropping its required support.

The public question induces an operation such as FACT, COUNT, LIST, TEMPORALLOOKUP, or TEMPORALARITHMETIC. The operation determines which evidence must be complete and which answer surface is required. Importantly, internal completeness is not equivalent to visible completeness.

C. Objective and Claim Taxonomy

For question q, frozen context C, internal derivation z, and surface contract h, the reader computes

$$z = D(q, C, o), \tag{8}$$

$$a = R(q, z, h). \tag{9}$$

D may enumerate members or retain temporal endpoints. R emits only the answer requested by q. We optimize all-row strict answer correctness while also reporting support recall, abstention behavior, token use, and latency. Retrieval metrics diagnose C; they never substitute for the correctness of a.

The system objective is therefore constrained rather than a single retrieval score. For evaluation rows $i = 1, \dots, n$, we report

$$\begin{aligned} \mathcal{J} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}[a_i \equiv y_i] - \lambda_s L_{\text{stale}} - \lambda_c L_{\text{conflict}} \\ - \lambda_f L_{\text{false abstain}} - \lambda_b \overline{B}, \end{aligned} \tag{10}$$

subject to provenance closure and generation-fencing invariants. The first term is end-to-end answer correctness; the

TABLE I
FUNCTIONAL COGNITIVE INSPIRATION AND TESTABLE SYSTEM COMMITMENTS.

Cognitive principle	Engineering commitment
Episodic encoding	Immutable, source-ordered experience traces
Semantic consolidation	Typed state projected separately from raw episodes
Reconsolidation	Append revisions; never erase the prior trace
Contextual recall	Cue-driven candidates resolved in valid/known time
Executive control	Operation routing, selective expression, and abstention

remaining terms expose safety and cost failures that a high Hit@k can hide.

We distinguish four evidence levels throughout the paper: stage-level retrieval, matched visible development, frozen candidate evaluation, and fresh confirmatory evaluation. Only the last two can support final superiority language, and only when their preregistered gates pass.

III. THE ECHO MEMORY PLANE

A. Cognitive Inspiration and Engineering Commitments

Human memory is not a uniform store. The episodic–semantic distinction separates temporally situated experience from organized knowledge [5]; complementary learning systems explain why rapid experience capture and slower structured learning serve different roles [6], [7]; and reconsolidation shows that recall can reopen an established memory to revision [8]. Temporal-context models connect recall to a changing internal context [9], while cognitive-control accounts emphasize goal-dependent selection of what guides behavior [10].

These correspondences are design hypotheses, not anatomical claims. Their value is operational: each row creates an independently auditable invariant or ablation. In this sense, "brain-inspired" names the decomposition of memory functions, while bitemporal state and provenance provide the engineering semantics needed to test it.

B. Target Runtime and Evaluation Scope

Figure 2 specifies the intended online and asynchronous composition. Durable ingestion is decoupled from optional projection and index maintenance, and the target query path returns from derived-index discovery to the bitemporal ledger before packing evidence. The empirical study evaluates durable carrier, candidate discovery, and context-construction paths at service level, and assesses V25 ledger, provenance-closure, operation-planning, and surface-contract properties through domain-level tests. Results are attributed to the corresponding evaluation boundary.

C. Projection, Provenance, and Revision Integrity

Projection converts raw turns and events into typed units while preserving raw source pointers. Observed propositions remain distinct from derived summaries or states, following the database-provenance principle that derived results must remain traceable to contributing records [11]. Repeated textual mentions do not automatically become distinct countable entities; deduplication uses event and typed entity identity.

The ledger component appends revisions rather than overwriting prior state. A revision may supersede, revoke, or leave another revision unresolved, allowing late-arriving evidence to change transaction-time knowledge without rewriting historical valid time. Component tests cover ledger generations and stale-index fences. Resolver behavior is evaluated through component-level invariants, and benchmark results are reported as retrieval and context-construction measurements.

D. Candidate Discovery and Currentness Authority

The candidate union combines typed routes, lexical matching, semantic retrieval, and fixed neighborhood expansion. BM25 provides an exact-anchor floor [12]; semantic retrieval improves paraphrase recall; and relation or state routes expose connected evidence. Fusion produces a candidate set, not truth. The ledger subsequently applies the requested valid/known-time view and revision status.

This separation prevents a frequent failure: a semantically close stale value outranking a less similar current value. Semantic rank can affect which states are inspected, but not which inspected revision is declared current.

E. Provenance-Closed Evidence Packing

For each operation, the packer constructs atomic evidence sets. A count set groups the countable members; a temporal arithmetic set groups both endpoints; a current-state set groups the active revision with the revision evidence needed to exclude an older value. Sets are deduplicated, prioritized, and packed under budgets of 768, 2,048, or 7,000 estimated tokens. Within an admitted set, source order and timestamps are restored before reading.

The packer records context hash, source IDs, candidate and packed hit counts, truncation, admitted and deferred closure IDs, and the currentness owner. This trace makes an evidence intervention testable independently of the generated answer.

F. Selective Surface Realization

Evidence completeness and answer verbosity are different constraints. Let I_o be internal requirements and F_o the visible contract for operation o. The reader may enumerate members, compare revisions, or normalize temporal endpoints internally, but the public result is $R(D(q, C, I_o), F_o)$. This boundary is an engineering analogue of executive gating: the mapping is functional and testable, not neural.

Activation receives only the public question. It never receives benchmark, question ID, arm, retrieval fields, reference answer, rubric, or gold support. If evidence is missing, the reader emits `INSUFFICIENT_EVIDENCE`. If revisions are mutually inconsistent and no ledger field resolves them, the system abstains rather than selecting by similarity. This resembles selective prediction's rejection option [13], but the trigger is evidence and state integrity rather than classifier confidence.

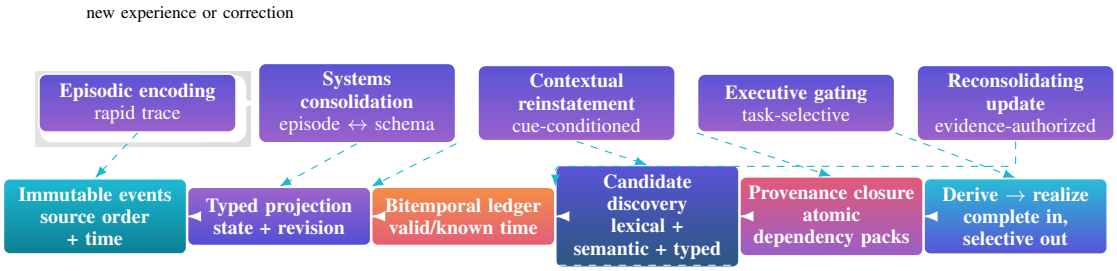

Fig. 1. ECHO at a glance. The upper lane translates a functional cognitive loop into the auditable memory plane below; colors distinguish experience, projected state, temporal authority, discovery, provenance, and realization. Solid arrows are runtime flow and dashed arrows are testable correspondences. The feedback loop admits only new experience or authorized correction, never an unverified retrieval result.

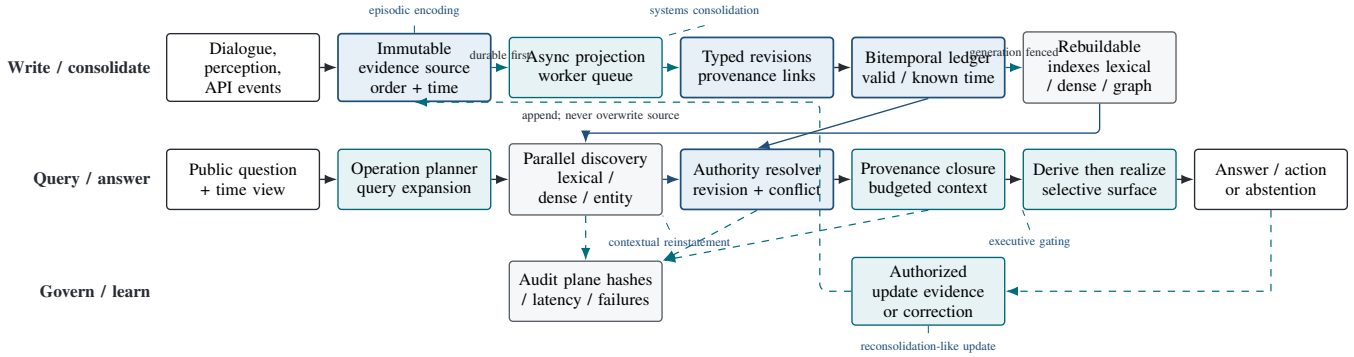

Fig. 2. Target ECHO runtime architecture and current integration boundary. Solid arrows denote the intended synchronous authority path; dashed arrows denote asynchronous projection, observability, and authorized updates. The empirical analysis focuses on retrieval and context construction. Cognitive labels are functional correspondences, not claims of anatomical or neural equivalence.

TABLE II
INTERNAL AND VISIBLE REALIZATION CONTRACTS.

Operation	Internal requirement	Final surface
Fact	Minimal sufficient support	Minimal supported answer
Count	Complete member set, deduplicated by identity, and verified count	Numeric count; unit only if needed
List	Complete deduplicated set; requested order retained	Requested items only
Temporal arithmetic	Both endpoints, normalized unit, verified calculation	Requested duration or date only
Temporal lookup	Valid/known-time view and supported target state	Time-scoped fact only

G. Failure Semantics

Every public row and every arm-budget cell is retained. Reader transport, parsing, and judge failures score zero. Exact abstentions are scored deterministically; remaining byte-distinct answers are judged once per question-response pair and reused across consuming cells. Prediction files are sealed before gold can be opened. These rules prevent retry, deduplication, or row filtering from improving reported accuracy after failures are observed.

IV. Implementation Status and Reproducibility

The evaluated retriever runs through the service carrier. The empirical study reports service-level retrieval results and domain-level typed-ledger and closure tests at their corresponding evaluation boundaries. Working bundles retain prompts, revisions, predictions, and scores. Reproducibility statements in this preprint refer to the archived artifacts and configurations described here. API credentials are never serialized.

The sampled Mem0 arm is pinned to a recorded OSS and benchmark revision, with its dependency, model, and protocol metadata archived alongside the artifact. The protocol-correct overlay maps benchmark timestamps to `metadata.created_at`, applies SDK entity filters and Top-10 limits, disables hidden Qwen thinking for wire compatibility, and enforces an exact extraction schema. The frozen stock adapter drops benchmark timestamps; this is retained as a protocol limitation, not repaired silently. Mem0 does not expose complete source-turn lineage across memory revisions, so ECHO-style Hit@10, turn/session recall, and provenance MRR are *not observable* for this arm.

Runs execute on Windows 11 with Docker 29.7.2 over WSL2, an Intel i9-13900H, 15.6 GiB host RAM, an RTX 4060 Laptop GPU with 8 GiB VRAM, and an 8 GiB Docker memory limit. Ollama 0.32.14 serves both Qwen models. These local measurements are distinct from managed Mem0 and from author-reported product results [14].

TABLE III
FROZEN ECHO RETRIEVAL SCOPES. REGISTERED/EVALUATED COUNTS ARE BOTH SHOWN TO PREVENT DENOMINATOR DRIFT.

Dataset	Reg./eval.	Hist.	Scope
LoCoMo	1,986/1,982	10	categories 1–4 primary; category 5 separate
LongMemEval-S	500/500	500	all six released question types
BEAM development	20/18	1	one 100K row; pilot only
BEAM fresh gate	91/89	5	preregistered histories; retrieval only

V. EVALUATION PROTOCOL

A. Research Questions

- **RQ1:** How much annotated evidence does the frozen ECHO read path recover on LoCoMo, LongMemEval-S, and BEAM histories?
- **RQ2:** Which benchmark types remain limited by session discovery, multi-session coverage, temporal filtering, or context packing?
- **RQ3:** Under a shared local answerer and judge, how does ECHO compare with pinned, protocol-correct Mem0 OSS on a fixed sampled QA set?
- **RQ4:** What latency, context, and construction costs accompany the retrieval gains on the fixed local stack?
- **RQ5:** Which ledger, closure, routing, and worker components remain load-bearing under ablation and concurrency tests?

B. Datasets

LoCoMo contains four questions with no registered evidence; the category-1–4 primary retrieval slice therefore has 1,536 evaluated rows, while category 5 has 446 and is reported separately. LongMemEval-S uses the cleaned 500-question release with 246,750 turns. The development BEAM result is one local 100K conversation with 188 turns, not the official BEAM-1M or BEAM-10M evaluation. The five-history gate uses rows fixed before retrieval outcomes were opened; five additional confirmation rows remain sealed. LoCoMo, LongMemEval-S, and the one-history BEAM pilot were visible during development.

C. Matched Arms and Budgets

The frozen ECHO runs use Top-10 retrieval, HNSW, eight workers, query expansion with at most three subqueries, and the local Qwen embedding endpoint. Candidate and context ceilings are dataset-specific: LoCoMo and BEAM use 256 candidates and 29.4 kB total context; LongMemEval-S uses 512 candidates and 60 kB total context.

A post-hoc source audit found that the frozen query-expansion rule set contains proper names and lexical phrases added while benchmark histories were visible. The public adapters still exclude answer, reference, rubric, and support fields, so this is not a direct gold-field path. Nevertheless, an expansion can inject an answer-bearing token that was absent from the question. We therefore classify every expansion-enabled retrieval and ablation number as a development-stage diagnostic. A source-neutral rewrite and clean rerun are required before these values can support confirmatory retrieval claims.

The Mem0 comparison is a sampled matched run rather than a full-dataset reproduction. It contains 91 questions from eight histories: 65 LoCoMo, six LongMemEval, and 20 from the local BEAM-100K history. Both systems use the same local `qwen3:4b` answerer and binary judge, while their saved Top-10 contexts remain system-specific. We retain every question and all transport, empty-answer, and judging failures. Because the pinned Mem0 revision does not retain complete source-turn lineage across memory revisions, ECHO-style Hit@10, turn recall, session recall, and provenance MRR have no Mem0 value in the matched table.

Author-reported systems are separated because their readers, judges, cutoffs, dataset releases, and managed components differ. A published LLM-judge score is not compared numerically with ECHO retrieval recall.

D. Metrics and Statistical Treatment

Let G_i^T and G_i^S be the unique gold turn and session sets for question i, let $\widehat{G}_{i,k}^T$ and $\widehat{G}_{i,k}^S$ be their recalled subsets in the first k hits, and let r_i be the rank of the first relevant hit (∞ if absent). We report

$$\mathrm{Hit@}\,k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[\widehat{G}_{i,k}^T \neq \varnothing], \qquad \mathrm{MRR} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{r_i}, \tag{11}$$

$$R_{\mathrm{turn}}@k = \frac{\sum_i |\widehat{G}_{i,k}^T|}{\sum_i |G_i^T|}, \qquad R_{\mathrm{session}}@k = \frac{\sum_i |\widehat{G}_{i,k}^S|}{\sum_i |G_i^S|}. \tag{12}$$

These retrieval estimands diagnose evidence delivery and are never substituted for strict answer accuracy $n^{-1}\sum_i \mathbb{1}[a_i \equiv y_i]$. We also retain mean query latency, index construction time, and context bytes. Latency includes the local query embedding call and is therefore not pure vector-index service time.

For descriptive uncertainty we pair x/n with the 95% Wilson interval

$$\frac{\widehat{p} + z^2/(2n) \;\pm\; z\sqrt{\widehat{p}(1-\widehat{p})/n + z^2/(4n^2)}}{1 + z^2/n}, \qquad z = 1.96, \tag{13}$$

and report the two-sided exact McNemar test for paired correctness. Let b be the number of ECHO-correct/Mem0-wrong questions and c the reverse:

$$p_{\mathrm{exact}} = \min\left(1, 2\sum_{j=0}^{\min(b,c)} \binom{b+c}{j} 2^{-(b+c)}\right). \tag{14}$$

Here $b = 9$ and $c = 30$. We complement the question-level test with stratified paired bootstrap intervals and a history-cluster sensitivity analysis. The sample covers eight histories, so the automated-judge analysis and its p-value are interpreted at the protocol level rather than as full-dataset population estimates.

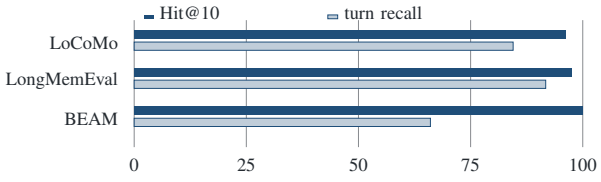

Fig. 3. Question-level hits conceal incomplete evidence coverage. The gap is largest on the local BEAM pilot. LoCoMo uses the category-1–4 primary slice.

E. Leakage and Invariance Audits

Public adapters reject answer, reference, rubric, and support fields. Prediction entry points have no gold path. LongMemEval requires special care because the released JSON co-locates `answer`, `has_answer`, and `answer_session_ids` with histories; the public adapter strips these labels before ingestion. LoCoMo likewise separates QA answer/evidence annotations from conversation payloads.

Before any selective-realization model call, we rebuild the complete task matrix and compare all context hashes, token counts, source IDs, hit counts, truncation fields, packing traces, retrieval latency, and provider identities against the frozen parent. The run is invalid if any context field differs.

VI. Results

A. Frozen ECHO Evidence Retrieval

Table IV replaces earlier presentations that mixed a LoCoMo Hit@10 and LongMemEval Recall@5 under an *accuracy* heading.

All three development runs have zero projection failure. The shape checks are 10 conversations/5,882 turns for LoCoMo, 500 conversations/246,750 turns for LongMemEval-S, and one local conversation/188 turns for BEAM. Index construction takes 13.9 minutes, 14.90 hours, and 71.2 seconds, respectively, on the fixed laptop stack. LongMemEval-S is dominated by 246,001 projected memories and local embedding generation; it should not be interpreted as a vector-index-only build time.

Figure 3 is the central diagnostic: Hit@10 answers whether at least one annotated turn appears, whereas turn recall measures how much required evidence appears. The two are not interchangeable.

B. Remaining Retrieval Errors

LoCoMo category 3 is the weakest slice (86.96% Hit@10, 66.83% turn recall, MRR 0.5210); category 1 also has a coverage gap at 71.38% turn recall. These misses are dominated by multi-turn and indirect support rather than simple absence of a relevant session. LongMemEval-S single-session preference is the weakest type (86.67% Hit@10 and session recall, MRR 0.6892), while the multi-session slice retains 86.24% turn recall and 90.70% session recall. The remaining LongMemEval errors therefore mix preference paraphrase and multi-session coverage rather than a global context ceiling.

The BEAM development pilot is a stronger warning. Event ordering and summarization each recover only 42.86% of annotated turns and 50% of sessions despite 100% Hit@10. Temporal reasoning has full turn/session recall but MRR 0.3214, indicating a ranking rather than coverage failure. These slices motivate separate ordering, summarization-coverage, and temporal-ranking workers; adding workers is justified only when a targeted smoke changes the implicated evidence and passes a stable-hit non-regression slice.

C. Fresh-History BEAM Generalization

The five-history run has zero projection failures but fails all three retrieval thresholds in Table V, falsifying a broad reading of the one-history pilot. Event ordering is weakest (40.0% Hit@10, 14.71% turn recall, and 21.57% session recall), followed by summarization (70.0%, 14.46%, and 26.0%). A post-gate Top-50 diagnostic places many gold turns at ranks 16–41, implicating candidate coverage and ranking rather than ingestion. A generic focus-rewrite smoke adds no Top-10 hits and is reverted. Under the stopping rule, cross-benchmark non-regression is skipped and the reserved confirmation histories remain sealed.

D. Protocol-Matched ECHO–Mem0 OSS QA

On this fixed 91-question sample, Mem0 is 23.08 percentage points higher in binary QA accuracy. The discordant counts are $b = 9$ and $c = 30$, yielding the exact test in Eq. 14. A stratified paired bootstrap gives a 95% interval of $[-35.16, -10.99]$ points. The questions come from only eight histories, however, and a paired history-cluster sensitivity interval is $[-25.5, +25.0]$ points. We therefore conclude only that Mem0 is stronger on this fixed question sample under the shared answer protocol; we do not infer a population advantage or a retrieval-quality ordering.

The protocol-correct Mem0 overlay maps benchmark timestamps to `metadata.created_at`, uses exact extraction schemas, and preserves all ingestion/search failures. The invalid pre-schema run and repaired replay remain in the artifact ledger. Mem0's revision objects do not preserve a complete mapping to every contributing benchmark turn, so its ECHO-style Hit@10, turn recall, session recall, and provenance MRR are *not observable*. The automated judge result is provisional until the frozen 30-question, 120-response packet is independently labeled by two humans.

E. Author-Reported Results Are a Separate Estimand

The managed numbers are taken from the Mem0 repository README (accessed 19 August 2026), which explicitly states that proprietary optimizations are not available in the open-source SDK. The LoCoMo paper numbers come from [14]. We therefore neither subtract them from ECHO retrieval metrics nor label the local matched sample as a reproduction of managed Mem0.

TABLE IV
CUTOFF-ALIGNED EVIDENCE RETRIEVAL, NOT QA ACCURACY. HIT@10 IS QUESTION-LEVEL; TURN/SESSION R.@5 AND R.@10 ARE MICRO ANNOTATED-UNIT COVERAGE RECOMPUTED FROM THE SAME FROZEN TOP-10 ROWS.

Dataset scope	Eval.	Hit@10 (%)	Turn R.@5 (%)	Turn R.@10 (%)	Sess. R.@5 (%)	Sess. R.@10 (%)
LoCoMo cat. 1–4	1,536	96.29	73.64	84.48	–	–
LongMemEval-S	500	97.60	88.84	91.74	88.71	93.46

TABLE V
PREREGISTERED FRESH-HISTORY BEAM GENERALIZATION FOR ECHO TOP-10 RETRIEVAL. THE PRESPECIFIED GATE REQUIRED AT LEAST 80 EVALUATED QUESTIONS, ZERO PROJECTION FAILURES, HIT@10 $\geq 85\%$, TURN RECALL $\geq 50\%$, AND SESSION RECALL $\geq 60\%$.

History row	Evaluated	Hit@10 (%)	Turn R. (%)	Sess. R. (%)	MRR
6	18	94.44	47.06	51.11	.6308
16	18	94.44	58.18	67.50	.7102
12	18	61.11	22.99	40.00	.4611
9	17	82.35	44.44	58.97	.6123
5	18	61.11	42.62	46.67	.4894
Aggregate	89	**78.65**	**40.80**	**52.34**	.5804
Gate	≥ 80	≥ 85	≥ 50	≥ 60	–

TABLE VI
MATCHED SAMPLED END-TO-END QA. BOTH ARMS USE THE SAME BINARY QWEN ANSWERER/JUDGE PROTOCOL. PARENTHESES CONTAIN WILSON 95% INTERVALS. MEM0 PROVENANCE RETRIEVAL METRICS REMAIN UNOBSERVABLE AND ARE NOT ASSIGNED ZERO.

Sample	n	ECHO correct (%)	Mem0 correct (%)	$\Delta_{\mathrm{ECHO-Mem0}}$ (pp)	exact p
LoCoMo	65	29 (44.62; 33.17–56.66)	46 (70.77; 58.80–80.42)	-26.15	.00333
LongMemEval	6	3 (50.00; 18.76–81.24)	2 (33.33; 9.68–70.00)	+16.67	1.00000
BEAM local 100K	20	6 (30.00; 14.55–51.90)	11 (55.00; 34.21–74.18)	-25.00	.12500
All sampled questions	91	38 (41.76; 32.16–52.02)	59 (64.84; 54.61–73.86)	-23.08	.00107

TABLE VII
PRIMARY-SOURCE MEM0 NUMBERS SHOWN FOR ORIENTATION ONLY. NONE IS METRIC-COMPATIBLE WITH ECHO RETRIEVAL IN TABLE IV OR THE MATCHED SAMPLED QA IN TABLE VI.

Source/system	Dataset	Reported metric	Score	Protocol distinction
Mem0 managed platform	LoCoMo	platform score	92.5	Top-200, production stack with proprietary optimizations
Mem0 managed platform	LongMemEval	platform score	94.4	same managed stack; not OSS and not ECHO retrieval recall
Mem0 managed platform	BEAM-1M / 10M	platform score	64.1 / 48.6	official larger scales, not the local one-row 100K pilot
Mem0 paper, base	LoCoMo	overall LLM-judge	66.88 ± 0.15	GPT-4o-mini-era paper protocol and averaged judge runs
Mem0 paper, graph	LoCoMo	overall LLM-judge	68.44 ± 0.17	graph memory; different storage, reader, judge, and cutoff
Zep / full context	LoCoMo	overall LLM-judge	$65.99/72.90$	author-reported baselines in the Mem0 paper

F. Evidence Scope

The frozen ECHO artifacts establish high evidence retrieval on two full conversational benchmarks. The one-history BEAM pilot is explicitly development evidence, and the preregistered five-history test demonstrates materially weaker generalization. The matched QA sample establishes a question-level disadvantage for ECHO on the fixed 91 questions, not a population ranking: only eight history clusters are represented and the judge is not human-calibrated. Author-reported managed results remain a third, non-comparable estimand.

VII. LOAD-BEARING ABLATIONS, ROBUSTNESS, AND SCALE

A. BEAM Retrieval Ablations

We use the complete local BEAM-100K row as a paired diagnostic because it is small enough to rerun exactly and exposes the largest hit–coverage gap. Every arm keeps the frozen v11 Top-10 budget, Qwen embedding model, candidate and context limits, 188 indexed turns, and the same 18 evidence-bearing questions. Only the named component changes. All four audits contain 20 unique question records (18 evaluated and two annotation-defined no-evidence questions), zero invalid records, and zero projection failures.

Table VIII rejects two shortcut explanations. Dense retrieval alone preserves a one-turn hit for every question but loses 9.44 turn-recall points, 13.16 session-recall points, and 0.226 MRR. Lexical retrieval alone misses three questions and loses 15.10 turn-recall points. Within this development configuration, query expansion is the largest load-bearing component: removing it halves mean query time, but Hit@10 falls by 16.67 points, turn recall by 26.42 points, and session recall by 18.42 points. Because the audited rules contain source-specific phrases, this drop measures dependence on that particular

TABLE VIII
LOAD-BEARING BEAM-100K RETRIEVAL ABLATIONS ($n = 18$ EVALUATED QUESTIONS). DENSE-ONLY DISABLES LEXICAL AND ENTITY ROUTES; LEXICAL-ONLY DISABLES DENSE AND ENTITY ROUTES. LATENCY INCLUDES LOCAL QUERY EMBEDDING. RESULTS ARE A PAIRED LOCAL DIAGNOSTIC, NOT A BEAM-1M/10M CLAIM.

Variant	Hit@10 (%)	Turn R. (%)	Sess. R. (%)	MRR	Mean query (ms)	Mean ctx. (KiB)
Full v11	**100.00**	**66.04**	**78.95**	**0.811**	927.8	30.6
Dense-only	100.00	56.60	65.79	0.585	796.4	28.3
Lexical-only	83.33	50.94	78.95	0.463	825.6	36.2
No query expansion	83.33	39.62	60.53	0.644	**427.5**	30.5
No session-support weight	100.00	62.26	78.95	0.784	994.3	30.5

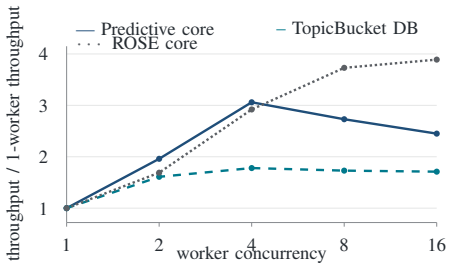

Fig. 4. Normalized worker throughput on the fixed Windows/WSL2 Docker host. Stateful TopicBucket saturates near four workers; pure C++ ROSE reranking continues scaling. These are worker-path measurements, not QA accuracy.

rule set, not the benefit of a generic reformulator. Session support remains a smaller diagnostic term, recovering two of 53 annotated turns and 0.027 MRR.

B. Robustness Boundary

The ablations validate component dependence, not universal robustness. Controlled tests still need to vary late arrival, correction, retraction, contradictory sources, missing support, repeated entity mentions, relative dates, and stale derived indexes. Required invariants are zero generation-stale leakage, no similarity-based currentness decision, preservation of historical as-of answers, and deterministic abstention when a required closure is absent. Prompt and paraphrase robustness must use source-neutral transformations fixed before answers are exposed.

C. Worker Concurrency Frontier

We fix 200 operations per point and 64 candidates and sweep concurrency $\{1, 2, 4, 8, 16\}$. The deterministic stress harness excludes remote LLM and embedding calls, so it isolates worker/state-path capacity rather than end-to-end answer latency. Across 40 configurations (8 paths per level) and 8,000 operations, no operation fails.

Figure 4 shows that worker count is not a monotone capacity knob. Predictive-core throughput peaks at 107.4k calls/s at four workers; TopicBucket/DuckDB peaks at 1,507 calls/s at four and remains near 1,447 calls/s at 16 while p95 grows from 2.07 to 20.43 ms. ROSE core reaches 15.9k reranks/s at 16 with 0.598 ms p95. In contrast, existing Ladybug graph paths stay near 40–44 calls/s and exhibit 1.28–2.15 s p95 at 16, locating the bottleneck in storage coordination rather than worker arithmetic. Four workers are therefore the defensible stateful operating point on this host; eight or more are useful only for CPU-local reranking. Production capacity still requires replication on target hardware with live model endpoints.

A deterministic post-merge carrier gate further separates repository overhead from model latency. With 2,000 active memories, 100 lexical Top-10 queries, and an in-memory DuckDB repository, Recall@1, Recall@10, MRR, and NDCG@10 are all 1.0; p50/p95/p99 query latency is 61.2/71.0/73.6 ms and peak RSS is 0.99 GiB. The gate excludes embedding and LLM calls and is therefore neither an end-to-end benchmark nor a cross-system comparison. Batched revision validation and recall-audit insertion preserve per-hit provenance while avoiding statement-per-hit preparation overhead.

Human judge calibration and reserved-history confirmation remain future work.

VIII. RELATED WORK

A. Persistent Agent Memory

Generative Agents stores observations, reflections, and plans [15]; MemGPT treats long-lived context as a virtual-memory management problem [16]; A-MEM organizes evolving notes with links [17]; Mem0 extracts and updates salient memories [14]. Zep represents evolving facts in a temporal knowledge graph [18], while HippoRAG 2 uses graph propagation for factual and associative retrieval [19]. These systems establish persistent, graph, and temporal memory as prior art. Our distinction is the explicit authority boundary among candidate discovery, bitemporal currentness, provenance closure, and selective answer realization under a matched audit protocol.

B. Temporal and Structured Memory

TReMu couples timeline memory with explicit temporal calculation [20]. LightMem combines filtering, topic grouping, and offline consolidation [21]. MemoryAgentBench broadens evaluation to retrieval, test-time learning, long-range understanding, and selective forgetting [22]; HaluMem localizes hallucinations to memory lifecycle stages [23]. ECHO builds on temporal database and provenance semantics rather than claiming that timestamps, graphs, or consolidation are new by themselves.

C. Long-Context Evaluation and Answer Judging

Lost in the Middle shows that readers can underuse evidence depending on its position [24]; full context is therefore a

control, not an assumed oracle. LoCoMo, LongMemEval, and BEAM progressively test longer and more dynamic conversational memory. Because LLM judges can vary with prompt and model, our protocol pins the judge, stores its exact outputs, and reports strict deterministic metrics alongside it. External headline scores remain author-reported unless rerun in the matched harness.

IX. Discussion

A. From an Archive to a Governed Memory Plane

The central architectural choice is to separate *discovery* from *authority*. Lexical and dense similarity are effective ways to propose what should be inspected, but they are poor mechanisms for deciding which revision is current or whether two states conflict. ECHO therefore places a bitemporal ledger and provenance closure between retrieval and realization. This boundary makes a stale answer, an unresolved conflict, and missing support different observable failures rather than different forms of "low relevance."

The cognitive framing is useful precisely where it produces such engineering commitments. Episodic encoding motivates immutable experience; consolidation motivates typed projections that remain linked to episodes; reconsolidation motivates append-only revision; contextual reinstatement motivates multi-route recall; and executive control motivates selective realization and abstention. These are functional correspondences that can be tested and ablated. They do not imply that the modules implement, simulate, or localize biological memory.

B. What the Evidence Supports—and Falsifies

The full LoCoMo and LongMemEval-S runs support a bounded claim: on the frozen local files and stack, ECHO delivers high Top-10 evidence coverage with zero projection failures. The development BEAM pilot then demonstrates that Hit@10 alone is insufficient, because one relevant item can coexist with missing multi-event support. The fresh-history gate goes further: its 78.65% Hit@10 and 40.80% turn recall falsify any claim that the 100% pilot hit rate generalizes across BEAM histories. Top-50 diagnostics localize much of the gap to ranking and coverage, especially for event ordering and summarization.

The matched QA sample draws a second boundary. Mem0 OSS is significantly better at the question level on the 91 fixed items, while the eight-history cluster sensitivity interval remains inconclusive. This does not negate ECHO's retrieval results, because retrieval coverage and answer accuracy are different estimands; it shows that an auditable evidence plane is not by itself a stronger reader. Conversely, Mem0's lack of complete source-turn lineage prevents its QA result from being interpreted as a provenance-retrieval score.

C. Implications for Memory-System Evaluation

Three reporting practices follow. First, report question hits together with micro turn/session coverage and first-relevant rank. Second, separate frozen development, fresh generalization, and end-to-end QA; a gain at one stage must not be relabeled as another. Third, treat observability as part of the protocol: a metric that a baseline cannot expose is missing, not zero. These practices make negative results actionable and reduce the temptation to compare managed product scores, open-source retrieval metrics, and judge-based QA as though they were interchangeable.

X. Limitations and Ethical Considerations

LoCoMo and LongMemEval-S were visible during development, and a post-hoc audit found source-specific query-expansion phrases; those retrieval scores are provisional despite exclusion of gold QA fields. The five-history BEAM gate fails and does not characterize BEAM-1M/10M; its reserved confirmation set remains sealed. Accordingly, the authority and reproducibility claims in this paper are tied to the evaluated components and archived artifacts.

The matched Mem0 comparison contains 91 questions but only eight history clusters. Question-level paired inference favors Mem0, whereas the history-cluster sensitivity interval crosses zero; neither should be extrapolated to complete datasets. Mem0 memory revision also does not expose complete source-turn provenance, so retrieval Hit@10, turn/session recall, and provenance MRR are neither zero nor inferable. Managed Mem0 scores use proprietary components and different protocols.

The current study evaluates evidence delivery, not autonomous behavior or human-like cognition. The cognitive mapping is a functional design hypothesis, not a claim that software modules are anatomically, neurally, or clinically equivalent to biological memory. End-to-end QA is sensitive to answerer and judge choice. A 30-question, 120-response blinded calibration packet is frozen, and judge-dependent conclusions are therefore reported as protocol-level results rather than population estimates.

Persistent personal memory raises privacy, consent, deletion, and access-control risks. Raw evidence and derived state should be scoped per user, encrypted, auditable, and deletable through provenance-aware revocation. Benchmark artifacts must exclude credentials, private hostnames, machine identifiers, and personally identifiable user logs.

XI. Conclusion

ECHO is an auditable memory architecture and service prototype separating immutable experience, bitemporal authority, candidate discovery, provenance closure, and realization. Its visible-dataset retrieval scores are descriptive because source-specific expansion rules were used; it also fails the fresh BEAM gate, and Mem0 OSS is more accurate on the fixed 91-question QA sample. The contribution is therefore architectural and methodological: make currentness, support, conflict, and failure observable, while aligning each claim with its corresponding evidence boundary.

Appendix A
Evidence Contract

All reported ECHO retrieval rows must be bound to dataset, source, model, prompt, and executable hashes; interrupted

audits may be resumed only by preserving valid records and retaining truncated lines as failure evidence. Retrieval and QA metrics are separate estimands. A baseline field that is not exposed by its API is reported as *not observable*, never as zero. Sampled baseline QA must retain n, failures, and uncertainty and must not be extrapolated to a full benchmark.

A future superiority claim requires a frozen answerer and human-calibrated judge, a sufficiently powered protocol-matched baseline, a fresh source that passes its preregistered gate, zero denominator drift, and non-regression gates fixed before answers are opened. Accordingly, the paper claims an auditable architecture, component-level invariants, and development-stage protocol findings, with confirmatory claims reserved for protocols satisfying the preregistered requirements above.

Appendix B
Artifact Ledger

Frozen ECHO artifacts include the LoCoMo v7, LongMemEval v5, BEAM v11 development pilot, and five-history fresh BEAM gate summaries and question-level audits. The evidence ledger also includes four complete BEAM ablation pairs, their analyses, and the 40-configuration worker-stress report bundle. A post-merge 2,000-memory/100-query carrier gate binds the Release source revision. Mem0 artifacts retain the invalid pre-schema run, exact failed-pair replay, protocol-correct ingestion, saved retrievals, all 91 paired predictions and scores, manifests, and completion records. The artifact ledger is organized around the associated paths, datasets, prompts, images, container revisions, Qwen model IDs, environment metadata, and failures by SHA-256.

References

[1] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of LLM agents," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024, pp. 13 851–13 870.

[2] D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu, "LongMemEval: Benchmarking chat assistants on long-term interactive memory," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=pZiykaVOWL

[3] M. Tavakoli, A. Salemi, C. Ye, M. Abdalla, H. Zamani, and J. R. Mitchell, "Beyond a million tokens: Benchmarking and enhancing long-term memory in LLMs," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://arxiv.org/abs/2510.27246

[4] K. Torp, C. S. Jensen, and R. T. Snodgrass, "Effective timestamping in databases," *The VLDB Journal*, vol. 8, no. 3–4, pp. 267–288, 2000.

[5] E. Tulving, "Episodic and semantic memory," in *Organization of Memory*, E. Tulving and W. Donaldson, Eds. New York: Academic Press, 1972, pp. 381–403.

[6] J. L. McClelland, B. L. McNaughton, and R. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory," *Psychological Review*, vol. 102, no. 3, pp. 419–457, 1995.

[7] D. Kumaran, D. Hassabis, and J. L. McClelland, "What learning systems do intelligent agents need? complementary learning systems theory updated," *Trends in Cognitive Sciences*, vol. 20, no. 7, pp. 512–534, 2016.

[8] K. Nader, G. E. Schafe, and J. E. LeDoux, "Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval," *Nature*, vol. 406, no. 6797, pp. 722–726, 2000.

[9] M. W. Howard and M. J. Kahana, "A distributed representation of temporal context," *Journal of Mathematical Psychology*, vol. 46, no. 3, pp. 269–299, 2002.

[10] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," *Annual Review of Neuroscience*, vol. 24, pp. 167–202, 2001.

[11] J. Cheney, L. Chiticariu, and W.-C. Tan, "Provenance in databases: Why, how, and where," *Foundations and Trends in Databases*, vol. 1, no. 4, pp. 379–474, 2009.

[12] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[13] Y. Geifman and R. El-Yaniv, "Selective classification for deep neural networks," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[14] P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav, "Mem0: Building production-ready AI agents with scalable long-term memory," in *European Conference on Artificial Intelligence*, 2025, pp. 2993–3000.

[15] J. S. Park, J. C. O'Brien, C. J. Cai, M. R. Morris, P. Liang, and M. S. Bernstein, "Generative agents: Interactive simulacra of human behavior," in *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 2023.

[16] C. Packer, V. Fang, S. G. Patil, K. Lin, S. Wooders, and J. E. Gonzalez, "MemGPT: Towards LLMs as operating systems," *arXiv preprint arXiv:2310.08560*, 2023. [Online]. Available: https://arxiv.org/abs/2310.08560

[17] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang, "A-MEM: Agentic memory for LLM agents," *arXiv preprint arXiv:2502.12110*, 2025. [Online]. Available: https://arxiv.org/abs/2502.12110

[18] P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef, "Zep: A temporal knowledge graph architecture for agent memory," *arXiv preprint arXiv:2501.13956*, 2025. [Online]. Available: https://arxiv.org/abs/2501.13956

[19] B. Jiménez Gutiérrez, Y. Shu, W. Qi, S. Zhou, and Y. Su, "From RAG to memory: Non-parametric continual learning for large language models," in *International Conference on Machine Learning*, 2025. [Online]. Available: https://arxiv.org/abs/2502.14802

[20] Y. Ge, S. Romeo, J. Cai, R. Shu, Y. Benajiba, M. Sunkara, and Y. Zhang, "TReMu: Towards neuro-symbolic temporal reasoning for LLM-agents with memory in multi-session dialogues," in *Findings of the Association for Computational Linguistics: ACL 2025*, 2025, pp. 18 974–18 988.

[21] J. Fang, X. Deng, H. Xu, Z. Jiang, Y. Tang, Z. Xu, S. Deng, Y. Yao, M. Wang, S. Qiao, H. Chen, and N. Zhang, "LightMem: Lightweight and efficient memory-augmented generation," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://arxiv.org/abs/2510.18866

[22] Y. Hu, Y. Wang, and J. McAuley, "Evaluating memory in LLM agents via incremental multi-turn interactions," in *International Conference on Learning Representations*, 2026. [Online]. Available: https://arxiv.org/abs/2507.05257

[23] D. Chen, S. Niu, K. Li, P. Liu, X. Zheng, B. Tang, X. Li, F. Xiong, and Z. Li, "HaluMem: Evaluating hallucinations in memory systems of agents," *arXiv preprint arXiv:2511.03506*, 2025. [Online]. Available: https://arxiv.org/abs/2511.03506

[24] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang, "Lost in the middle: How language models use long contexts," *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024.