Holistic Context: A Unified Cross-Session Context Retention Framework for Long-Horizon Autonomous Agents

MelandLabs

Abstract

Large language models (LLMs) increasingly power persistent AI coworkers that operate across sessions, tasks, and platforms, yet a fixed context window cannot maintain a coherent work state over long horizons. Information outside the window becomes invisible, while information that remains is rarely organized by entity, applicability, provenance, and temporal validity. We argue that "context" should therefore be treated not as a transient token sequence but as a continuously maintained *holistic state* spanning time and authorized work platforms. We present Holistic Context, a unified cross-session context framework instantiated in the OpenContext open-source AI coworker, that *combines* three established primitives into one applicability-scoped pipeline: *applicability-scoped attribution*, which assigns fragments a five-tuple scope, consolidates them into entity timelines, and reconciles conflicting evidence; *tiered forgetting with evidence-preserving deprecation*, which routes beliefs through short-, mid-, and long-term memory, preserves links to source records, and supports belief revision, abstention, and as-of-time queries; and *Hebbian relation learning with competition*, which reinforces associations between co-retrieved memories, decays idle connections, and organizes support, compete, related, and supersede relations for associative multi-hop retrieval. Each primitive has a direct cognate in the temporal-database, belief-revision, or cognitive-science literatures; the contribution of this paper is the *system* that wires them together, together with a feasibility demonstration on three benchmarks. Experiments across LongMemEval-S, LoCoMo, and BEAM show that the coupled design supports long-horizon context management: (i) end-to-end runs on LongMemEval-S (n=500) and LoCoMo (n=10) reach 97.6% and 97.4%; and (ii) BEAM nugget-pass-rate reaches 72.8% / 75.7% / 76.5% / 67.0% at 128K / 500K / 1M / 10M tokens[‡]. All Holistic Context cells are *single-seed, single-backbone point estimates* on Claude

Sonnet 4.5; the BEAM-128K cell is computed on a 100K-token subset of the corresponding BEAM conversations; the LoCoMo cell spans only ten conversations; and the cross-paper baseline rows are published numbers from systems run on different backbones and judging configurations. The headline numbers should therefore be read as a feasibility demonstration, with the multi-seed replication and the controlled head-to-head against Hindsight / APEX-MEM / Mem0 V3 described as future empirical work.

Keywords: Large language models, autonomous agents, long-horizon memory, cross-session context, holistic context, memory management, context retention, BEAM, LongMemEval.

1 Introduction

The promise of autonomous LLM agents rests on their ability to act coherently over long horizons (planning across days, recalling decisions made in prior sessions, and reconciling information that arrives from different tools and platforms). Yet the substrate on which these agents operate, the fixed context window, was never designed for such persistence. As interaction histories grow, two failure modes dominate: information that lies outside the window is simply invisible, and information that survives inside the window is not organized into the entities, relations, and temporal states that coherent reasoning requires.

This collapse, a measurable degradation in capability as context length increases, is now quantified by a new generation of stress benchmarks. BEAM [Milanese et al., 2025] constructs coherent multi-domain conversations of up to 10M tokens and probes ten complementary memory abilities; even models with 1M-token windows, with or without retrieval augmentation, degrade sharply as dialogues lengthen. The strongest published memory-system baseline on global tasks at the 10M scale, the cognitively-inspired LIGHT framework [Milanese et al., 2025], falls to 26.6% on the Hindsight team's re-evaluation pipeline [Bartholomew, 2026, Latimer et al., 2026], while pure RAG falls to 24.9% on the same pipeline. Cross-pipeline differences already account for several percentage points of the published baseline numbers, so the absolute comparison is less informative than the structural fact that all published single-step retrieval baselines degrade sharply as the conversation scale grows. LongMemEval [Wu et al., 2024] arrives at a complementary conclusion from the capability side: across five memory dimensions (single-session and multi-session fact retrieval, knowledge updating, temporal

reasoning, and abstention), long-context LLMs trail dedicated long-context baselines by 30–46% per dimension (29.6–45.5 percentage points across the five capabilities), and the average residual gap to human performance is 39% (reaching 46% on the hardest knowledge-updating dimension). Cognitive-driven follow-ups such as EvolMem [Shen et al., 2026], Locomo-Plus [Krishnan et al., 2026], and LongMemEval-V2 [Wu et al., 2026a] confirm that the deficit is structural rather than incidental: it persists across models, domains, and probing protocols. These findings force a re-conceptualization. "Context" can no longer be equated with the tokens currently inside a window; it must be treated as a *holistic state* that accumulates across time and platforms and is actively maintained. This need is especially acute for persistent AI coworkers, which must preserve decisions, entities, and their evolution across sessions and authorized work platforms. Holistic Context maintains this evolving work state as auditable, source-linked entity records, complementing rather than replacing model-level or domain-level learning.

This re-conceptualization has a second, less visible dimension. Beyond losing state, persistent agents also lose the *rationale* behind state transitions: when an exception is approved, a configuration is changed, or a precedent is set, the reasoning that produced the change is rarely captured as a first-class record. Recent analyses of agent-execution traces argue that this missing rationale, not the missing data, is the structural gap that prevents today's agents from reasoning about exceptions, learning from precedents, or auditing decisions [Chao et al., 2026, Yan et al., 2026]. Holistic Context addresses this gap through its applicability-scoped attribution and evidence-preserving deprecation: every belief revision carries an archived supersession trace that records what was superseded, by what evidence, and at what time, so that the "why" behind a state transition is recoverable rather than implicit. The present paper does *not* evaluate this rationale-preservation claim with a dedicated probe; the motivation paragraph is forward-looking, and a supersession-trace-recovery probe is part of the future empirical work alongside the per-mechanism ablation.

A rich body of work attempts to extend effective context beyond the window. Memory-system approaches such as MemGPT [Packer et al., 2023] and Memory OS [Kang et al., 2025] borrow paging and hierarchy ideas from operating systems; Mem0 [Chhikara et al., 2025] and A-MEM [Xu et al., 2025] extract and store atomic facts for later retrieval; TiMem [Li et al., 2026] and H-MEM [Sun et al., 2026a] introduce temporal or hierarchical consolidation. Temporal knowledge-graph systems such as Zep/Graphiti [Rasmussen et al., 2025] and Hindsight [Latimer et al., 2026] maintain a per-entity bitempo-

ral record. Graph-native architectures [Park, 2026, Yang et al., 2026] and graph retrieval [Peng et al., 2024] structure memory as entities and edges. Yet four limitations recur. First, facts about the same entity are scattered across sessions and seldom *attributed* to a unified, conflict-resolved, temporally ordered record, so retrieval returns fragments rather than a consistent view. Second, memory accumulates monotonically: stale facts are not retired, beliefs are not revised when contradicting evidence appears, and the system cannot answer what it knew *as of* a particular moment, precisely the failure modes that temporal-validity studies [Yadav, 2024, Chao et al., 2026] and forgetting models [Sumida et al., 2024] expose, and the standard half-open validity-interval + as-of-time formulation that temporal databases [Snodgrass, 1995, Datomic Development Team, 2024, ISO/IEC 9075:2011 (SQL:2011), 2011] have used for thirty years. Third, retrieval is largely flat: memories are treated as isolated vectors, so the associations that make multi-hop reasoning natural in humans are absent; the cognitive-science anchors for these co-retrieval dynamics are Hebbian learning [Hebb, 1949] and spreading-activation theory [Collins and Loftus, 1975, Anderson, 1983], while the closest engineering analogues are Personalized PageRank over static graphs [Gutiérrez et al., 2025] and temporal graph networks [Rossi et al., 2020, Trivedi et al., 2019]. Fourth, the rationale behind state transitions is rarely captured as a first-class record: when an exception is approved or a precedent is set, the reasoning that produced the change tends to be lost in transit, making it difficult for future agents to learn from past decisions. Novelty-gated admission and adaptive admission control [Zhang et al., 2026] address what enters memory but not how memories relate.

We present Holistic Context, a unified framework that treats cross-session context as a layered, routable, source-attributed state (Figure 1), and instantiate it in the OpenContext open-source AI coworker [OpenContext Contributors, 2026]. Holistic Context is a *system* contribution: it combines three primitives, each with established cognates in the temporal-database, belief-revision, or cognitive-science literatures, into one applicability-scoped pipeline that the community can reproduce, evaluate, and improve upon.

- **Applicability-scoped attribution.** Fragments dispersed across sessions are tagged with a five-tuple Applicability Context (task, conversation, channel, person, validity interval), merged into unified entities, and ordered into per-entity timelines; contradictory statements are resolved through evidence-weighted reconciliation. This yields a consistent, attributable view that retrieval and answering can share

without leaking across scope. The `memory-consolidation` pipeline and its Entity Registry realize this attribution layer over authorized cross-platform work signals. The five-tuple scope is closely related to the bitemporal context used in Zep/Graphiti [Rasmussen et al., 2025]; the novelty here is the explicit applicability axis and the per-entity timeline rather than a global temporal KG.

- **Tiered forgetting with evidence-preserving deprecation.** Rather than accumulating memory indiscriminately, HOLISTIC CONTEXT routes beliefs through short / mid / long tiers, retires those that fall below their tier's retention threshold, and archives retired beliefs for auditability. As-of-time queries reconstruct the active beliefs maintained by the system at a given moment, so memory carries an explicit history rather than remaining static. The Memory Forgetting Engine exposes this lifecycle while preserving links from compressed or deprecated state to source records. The Ebbinghaus-shaped decay is the standard importance-decay formulation [Sumida et al., 2024]; the novelty here is the half-open validity interval combined with evidence-preserving deprecation rather than a separate garbage collection step.

- **Hebbian relation learning with competition.** When two memories are co-retrieved inside a Living Connections window their association strengthens; during idle periods it decays. A typed relation vocabulary (*support*, *compete*, *related*, *supersede*) lets the network self-organize, growing sharper with use, so cross-entity multi-hop reasoning follows the structure of past co-activation rather than flat similarity. The Living Connections operator implements this update, turning repeated use into an evolving associative context graph. Hebbian-style reinforcement is a well-established model from cognitive science [Hebb, 1949]; the engineering analogue is Personalized PageRank over co-activation graphs [Gutiérrez et al., 2025]; the novelty here is the four-edge typed relation vocabulary together with the explicit competition-before-supersession rule that gates conflicting fragments.

We evaluate HOLISTIC CONTEXT on the three benchmarks that most directly stress long-horizon context: BEAM [Milanese et al., 2025] (128K–10M tokens), LongMemEval [Wu et al., 2024] (five memory dimensions), and LoCoMo [Maharana et al., 2024] (very long-term conversational memory). The results show a consistent benefit from explicit memory organization, although the benefit is not uniform across context scales. On BEAM, HOLISTIC CONTEXT obtains 72.8% / 75.7% / 76.5% / 67.0% at 128K / 500K / 1M / 10M

Figure 1: HOLISTIC CONTEXT at a glance. *Architecture*: fragments dispersed across sessions (left) are attributed to unified entities through five-tuple applicability context (mechanism 1), routed through tiered forgetting with as-of-time queries (mechanism 2), and linked into living relations by Hebbian co-occurrence (mechanism 3); the resulting entity record (right) carries a unified timeline with an auditable supersession trace.

tokens[‡] (single-seed, single-backbone point estimates at n=400 per scale; the 128K cell operates on a 100K-token subset of the 128K conversations; see Table 1 for Wilson 95% CIs and footnotes). The strongest published reference is Hindsight at 73.4% / 71.1% / 73.9% / 64.1%: HOLISTIC CONTEXT trails it by 0.6 points at 128K (inside both Wilson intervals) but exceeds it by 4.6, 2.6, and 2.9 points at 500K, 1M, and 10M; at the 10M scale HOLISTIC CONTEXT also exceeds Honcho (40.6%), LIGHT (26.6%), and RAG (24.9%) by 26.4, 40.4, and 42.1 points, and it exceeds Mem0 V3 (published only at 1M and 10M) by 12.4 and 18.4 points (64.1% and 48.6%). The 9.5-point decrease from 1M to 10M is reported on the same harness with the same prompt template; the comparison rows in Table 1 are single-pipeline-published reference numbers from the Hindsight team, not a same-harness head-to-head. We treat the BEAM scaling observation as a single-seed feasibility signal rather than as a controlled experiment; a multi-seed, multi-backbone replication is part of the future empirical work described in Section 4.6.

The end-to-end results on LongMemEval-S and LoCoMo show the same pattern. On LongMemEval-S, HOLISTIC CONTEXT reaches 97.6%, above the long-context baseline (43.8%), MemGPT (48.7%), Mem0 (52.0%), A-

MEM (53.8%), TSM (74.8%), TiMem (76.88%), Infini Memory (79.3%), APEX-MEM (86.2%), Hindsight (91.4%), and Mem0 V3 (94.4%). On LoCoMo, Holistic Context reaches 97.4%, above A-MEM (51.29%), MemoryOS (60.79%), Mem0 (66.88%), TiMem (75.30%), TSM (76.69%), Hindsight (85.67%), APEX-MEM (89.49%), and Mem0 V3 (92.5%). These gaps indicate that the combined design is substantially more effective than flat long-context or canonical memory baselines on the evaluated tasks. The Holistic Context cells on both benchmarks are single-seed, single-backbone point estimates, and the published rows come from systems run on different backbones and judging configurations; the comparison is therefore not a controlled head-to-head.

2 Related Work

Long-context modeling and its limits. Extending the context window has been the dominant response to the persistence problem [Liu et al., 2025]. Window-scaling data engineering, sparse and memory-keyed attention, and decoupled KV stores all push the token budget outward, yet empirical returns diminish sharply. BEAM [Milanese et al., 2025] shows that even 1M-token models fail on global reasoning as dialogues approach 10M tokens, and LongMemEval [Wu et al., 2024] demonstrates that capability degrades 30–46% across five memory dimensions even when the window nominally suffices. The shared lesson is that raw window capacity is necessary but insufficient: without organization, more tokens yield more noise. Holistic Context treats the window as a fast cache and relocates persistent state into an attributed external structure, directly addressing the failure modes that window scaling cannot.

Memory systems for LLM agents. Memory has become a first-class module in agent architectures [Wang et al., 2024a, Plaat et al., 2025, Zhang et al., 2024, Wu et al., 2025]. MemGPT [Packer et al., 2023] introduced OS-style paging between main and external context; Memory OS [Kang et al., 2025] generalized this idea into a hierarchical memory operating system; and Mem0 [Chhikara et al., 2025] provided production-oriented memory extraction, consolidation, and retrieval. Its subsequent V3 platform algorithm [Chhikara et al., 2026] added single-pass ADD-only extraction (with no UPDATE or DELETE operations), first-class storage for agent-generated facts, entity linking across memories, and multi-signal retrieval that fuses semantic similarity, BM25 keyword, and entity matching. On LoCoMo, Long-

MemEval, BEAM-1M, and BEAM-10M the V3 algorithm reports 92.5%, 94.4%, 64.1%, and 48.6%, respectively, at 6.7K–7.0K mean tokens per retrieval and p50 latency at or below 1.1s [Sangshetti, 2026]. A-MEM [Xu et al., 2025] adopted a self-organizing Zettelkasten formulation, while Infini Memory [Ji et al., 2026] maintained evolving topic documents. GAM [Wu et al., 2026b] consolidated dialogue events into a hierarchical association graph; CDMem [Gao et al., 2025] used context-dependent graph indexing; HippoRAG 2 [Gutiérrez et al., 2025] supported associative retrieval through Personalized PageRank; and Hindsight [Latimer et al., 2026] combined structured memory networks with temporal filtering and reflection. These systems mark a shift from flat fact stores toward structured, hierarchical, and temporally informed memory. Holistic Context builds on this direction by making applicability-scoped attribution the organizing primitive, unifying fragments from different sessions into consistent entity-level records while preserving their scope and provenance.

Hierarchical and temporal consolidation. Several recent systems introduce hierarchical and temporal structure into agent memory. H-MEM [Sun et al., 2026a] organizes memory at multiple levels for efficient long-term reasoning, while TiMem [Li et al., 2026] consolidates conversational observations into progressively abstracted representations through a Temporal Memory Tree. THEANINE [Ong et al., 2025] links memories through temporal and causal relations to construct evolving memory timelines. Temporal Semantic Memory [Su et al., 2026] organizes memories by semantic time and consolidates temporally continuous information into durative memories. APEX-MEM [Banerjee et al., 2026] represents temporally grounded events in an entity-centric graph and preserves their evolution through append-only storage, while Kumiho [Park, 2026] grounds versioned graph memory in formal belief-revision semantics. Graph-based taxonomies [Yang et al., 2026] characterize how entities and relations structure agent memory, and graph retrieval [Peng et al., 2024] extends RAG with relational traversal. Studies of temporal validity further show that agents retrieve stale facts when memory lacks explicit validity information [Yadav, 2024], struggle to recognize when stored beliefs have become outdated [Chao et al., 2026], and require time-sensitive graph retrieval over evolving knowledge [Han et al., 2025]. Holistic Context extends temporal consolidation with explicit validity intervals and active belief maintenance, enabling supersession, confidence-aware abstention, selective retirement, and as-of-time retrieval within the same entity history.

The closest published contrast for HOLISTIC CONTEXT in this family is Zep/Graphiti [Rasmussen et al., 2025], a temporally-aware knowledge graph that separates immediate context from historical knowledge and reports 94.8% on DMR and approximately 91.4% on LongMemEval. Zep/Graphiti organises memory as a bitemporal graph with no entity-scoped half-open validity interval as the query primitive, no five-axis applicability scope, no AGM-style supersession, and no typed Hebbian edges; HOLISTIC CONTEXT co-designs these primitives on a single per-entity timeline, which none of the cited temporal-KG systems (including Kumiho's belief-revision semantics and APEX-MEM's append-only history) jointly realises. The contribution of the present paper is a *system* (the OpenContext instantiation of HOLISTIC CONTEXT with all five primitives wired into one pipeline) together with a *feasibility demonstration* on three benchmarks. Each per-mechanism primitive has a direct cognate in the temporal-database, belief-revision, or cognitive-science literatures that we cite inline (Sections 3.2–3.4); a controlled per-mechanism attribution is left to the future empirical work in Section 4.6.

The *competition-before-supersession* rule that gates a conflicting fragment before it can supersede an established belief is *conceptually inspired by* AGM belief-revision theory [Alchourrón et al., 1985], particularly the propositional revision framework of Katsuno and Mendelzon [1991], though we deliberately relax two of its postulates. Under AGM, (i) revision requires expansion of a belief set K by a formula ϕ *possibly inconsistent* with K (the discipline that distinguishes revision from expansion), and (ii) revision canonically decomposes via the Levi identity as $K * \phi = (K \div \neg\phi) + \phi$, i.e., *contraction first*, then expansion. HOLISTIC CONTEXT instantiates the supersession analogue at the entity level by first forming a *compete* edge, and only closing the validity interval of the prior belief once the new belief has been demonstrated to be better supported. We explicitly depart from AGM in two ways: (a) the *compete* edge admits a fragment that contradicts an existing belief, which AGM forbids inside K; and (b) the temporal closure of an interval is *not* AGM contraction; it is a history-preserving supersession transition that operates on timestamped intervals rather than on a static propositional belief set. Because of these relaxations, we refer to the framework as *AGM-inspired supersession-with-history* rather than as AGM revision proper.

Admission, forgetting, and conflict. The question of what enters, changes, and leaves memory has gained increasing attention. Novelty-gated admis-

sion and adaptive admission control [Zhang et al., 2026] decide what information should be retained. Memory-R1 [Yan et al., 2026] learns ADD, UPDATE, DELETE, and NOOP operations through reinforcement learning, and PAMU [Sun et al., 2026b] updates preference memories to balance short-term changes against long-term user tendencies. Empirical analysis by Xiong et al. [2026] further shows that uncontrolled memory addition and deletion can amplify erroneous or misaligned experiences. Psychological models of importance and forgetting [Sumida et al., 2024] demonstrate that principled decay can improve long-term RAG. Knowledge-conflict surveys [Xu et al., 2024] formalize how contextual and parametric knowledge clash, a problem that intensifies when the same fact is restated, contradicted, or superseded across sessions. Holistic Context integrates admission, forgetting, and conflict resolution within a single attribution pipeline: a candidate fact is admitted only if it is novel, merged if it refines an existing entity, used to revise and supersede a prior belief if it conflicts, and aged out when superseded and idle. Unlike learned operation policies or task-specific update mechanisms, these lifecycle decisions are applied to applicability-scoped entity records with explicit provenance and temporal validity.

Foundational agent memory. Generative Agents [Park et al., 2023] introduced importance-weighted retrieval and reflection over a memory stream, while Reflexion [Shinn et al., 2023] stored verbal feedback as episodic memory to improve behavior across attempts. MemoryBank [Zhong et al., 2024] introduced persistent personalized memory with forgetting-curve-based decay and reinforcement. Voyager [Wang et al., 2024b] stored acquired behaviors in a retrievable and compositional library of executable skills, while ExpeL [Zhao et al., 2024] distilled reusable insights from past task trajectories without updating model parameters. These works established that agent memory should be selective, reflective, and reusable. Holistic Context extends these principles to cross-session entity attribution and evidence-weighted belief revision, while learning inter-entity relations through co-retrieval.

3 The Holistic Context Framework

3.1 Problem Formulation and Overview

Consider a persistent AI coworker that operates over a sequence of sessions $S_1, \ldots, S_T$, where session S_t produces a stream of interaction fragments $f_t^{(1)}, \ldots, f_t^{(n_t)}$. A fragment may be an utterance, a tool observation, or an

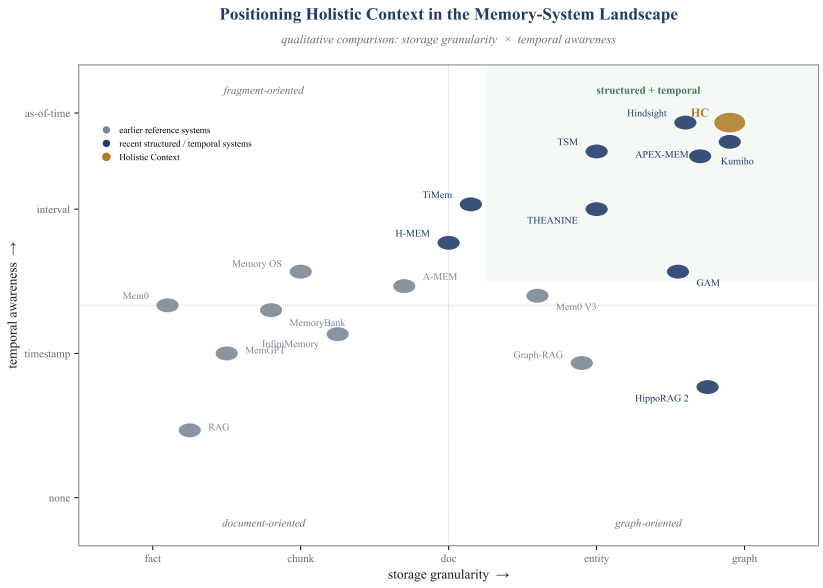

Figure 2: A qualitative positioning of representative memory systems along two axes: storage granularity (fragment → graph) and temporal awareness (none → as-of-time). Recent systems increasingly combine structured storage with temporal organization, but differ in how they represent validity, preserve historical states, resolve conflicting evidence, and support historical retrieval. HOLISTIC CONTEXT is positioned by its joint use of applicability-scoped entity records, explicit validity intervals, and as-of-time retrieval.

extracted fact. Naively, the agent's state at time t is the concatenation of all fragments seen so far, which quickly exceeds any fixed window and, more importantly, scatters information about the same real-world entity across many unconnected fragments. We instead maintain a *holistic context* $\mathcal{H}_t$ consisting of three coupled components: an *entity store* $\mathcal{E}_t$ of unified, conflict-resolved, temporally ordered records; a *temporal-validity layer* $\mathcal{V}_t$ that governs forgetting, belief revision, abstention, and as-of-time queries; and a *relation network* $\mathcal{G}_t = (\mathcal{E}_t, \mathcal{R}_t, W_t)$ whose edge weights reflect co-retrieval history. At each step the agent routes the current query through $\mathcal{G}_t$ to assemble a working context from $\mathcal{E}_t$, subject to $\mathcal{V}_t$, and then updates $\mathcal{H}_t$ with any new fragments.

The three components correspond to the three contributions of the paper

and are described in turn below. Attribution produces the entities and timelines that forgetting operates on, and forgetting in turn determines which entities remain active enough to participate in co-occurrence updates.

3.2 Applicability-Scoped Attribution

The goal of attribution is to convert a stream of fragments into a consistent, per-entity view, anchored to a *scope* that says which memories are relevant to a given query. We formalize this scope as an *Applicability Context*

$$A(f) \;=\; \big(\,\mathrm{task}(f),\ \mathrm{conversation}(f),\ \mathrm{channel}(f),\ \mathrm{person}(f),\ [t_f^{\mathrm{start}}, t_f^{\mathrm{end}}]\,\big), \tag{1}$$

a five-tuple of (*task*, *conversation*, *channel*, *person*, *validity interval*) that names where and when a fragment applies. Attributing a fragment to an entity means recording $A(f)$ alongside the entity record so that retrieval can match on Applicability Context, not merely on surface similarity. A foundational *scope isolation* invariant guarantees that each memory graph is bound to a $(\mathrm{userId}, \mathrm{workspaceId}, \mathrm{tenantId})$ triple and that cross-scope attribution travels only through explicit relations, never by default sharing, which prevents the silent leak of a private workspace's facts into another's retrieval results.

Entity merging. Each fragment f is passed through an entity linker that assigns it to one or more entities $e \in \mathcal{E}_t$ that share its Applicability Context. When f references an entity already in the store, its content is merged into that entity's record; when it introduces a new entity, a record is created. Because the same entity may surface under different surface forms across sessions, linking is performed against the full attribute profile of each candidate entity rather than surface string alone. Formally, we compute a linkage score $\ell(f, e) = \mathrm{sim}(\phi(f), \psi(e))$ over an embedding ϕ of the fragment and a profile embedding ψ of the entity, and link f to $\arg\max_e \ell(f, e)$ when the score exceeds a threshold θ_ℓ; otherwise a new entity is created. This is the mechanism by which fragments dispersed across sessions are gathered into unified entities.

Conflict resolution. A merged fragment may contradict a statement already held by the entity. Let the entity e currently hold a belief b with supporting evidence set E_b, and let the incoming fragment assert b' that is incompatible with b. We resolve the conflict by evidence-weighted reconciliation. The confidence score $c(b)$ is defined only when the belief has

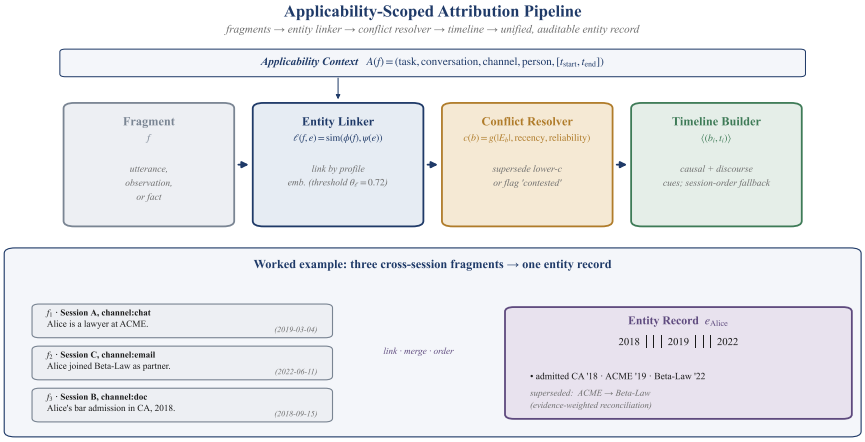

Figure 3: **Applicability-scoped attribution pipeline.** Each fragment f flows through an entity linker, a conflict resolver, and a timeline builder before being recorded as an Entity Record. The worked example shows three cross-session fragments about Alice being unified into a single entity record with a timeline of beliefs and an explicit supersession trace.

supporting evidence ($|E_b| \geq 1$); a belief with no direct evidence returns the sentinel $c(b) = \perp$ and is not comparable (the conflict-resolution test falls back to recency-of-creation instead). On the domain $|E_b| \geq 1$, define a confidence score

$$\begin{aligned} c(b) \;&=\; g\big(|E_b|,\ \text{recency}(b),\ \text{source-reliability}(E_b)\big) \\ &=\; \underbrace{\big(1-\exp(-|E_b|/\kappa)\big)}_{\text{count saturation} \in [0,1)} \cdot \underbrace{\exp\big(-\max(0, t_q - t_b^{\text{last}})/\rho\big)}_{\text{recency decay} \in (0,1]} \\ &\qquad \cdot \underbrace{\frac{1}{|E_b|}\sum_{e_i \in E_b} r_i}_{\text{mean source reliability} \in [0,1]}, \end{aligned} \tag{2}$$

The clamped time argument $\max(0, t_q - t_b^{\text{last}})$ guarantees the recency factor does not exceed 1 at $t_q < t_b^{\text{last}}$; the count factor is strictly less than 1 and approaches it as $|E_b| \to \infty$; and the mean-reliability factor is the arithmetic mean of finitely many $r_i \in [0, 1]$. Here $|E_b| \geq 1$ is the number of distinct supporting fragments, t_b^{last} is the timestamp of the most recent supporting fragment, t_q is the query time, $r_i \in [0, 1]$ is the reliability weight assigned to

the source of fragment e_i (operator-assigned for explicit user inputs, lower for inferential fragments), and $\kappa, \rho > 0$ are saturating-count and recency-decay constants ($\kappa = 5$, $\rho = 30$ days in the current configuration). The three factors are separable by design: the count factor saturates so that very well-supported beliefs do not dominate indefinitely; the recency factor rewards beliefs that have been reinforced recently; and the source-reliability factor prefers high-trust provenance. The system retains the higher-confidence belief and demotes the other to a *superseded* state rather than discarding it, so that prior claims remain auditable. When confidence is comparable (both beliefs lie within a contested band $|\Delta c| < 0.15$); the system flags the entity as *contested* and, at query time, surfaces both with provenance rather than silently picking one. This directly targets the knowledge-conflict pathology formalized by Xu et al. [2024]; the explicit decomposition into count, recency, and reliability terms makes the supersession-with-history step (see the related-work treatment of belief revision), closing the prior belief's validity interval on supersession while archiving its provenance; this makes the operation auditable rather than an opaque overwrite. We refer to the operation as *AGM-inspired supersession* because it borrows AGM's revision discipline without claiming AGM-membership; a full treatment of the deliberate relaxations is in § 2.

Temporal ordering. Each admitted fragment carries a timestamp, and per-entity events are ordered into a timeline $e.\tau = \langle (b_1, t_1), (b_2, t_2), \dots \rangle$ whose endpoints populate the validity-interval slot of $A(f)$. Ordering is not merely chronological sorting: when an event lacks an explicit time, Holistic Context infers a partial order from causal and discourse cues (e.g., "then," "after that," tool-call dependencies) and falls back to session order for ties. The timeline is the substrate on which both as-of-time queries (Section 3.3) and temporal-reasoning probes operate.

3.3 Tiered Forgetting with Evidence-Preserving Deprecation

Monotonic accumulation is the single largest source of error in long-horizon memory [Yadav, 2024, Chao et al., 2026]. Holistic Context instead routes every belief through a tiered *Memory Forgetting Engine*. At the framework level, retention is governed by an importance function $s(b) = g(\text{recency}, \text{access}, \text{importance}, \text{evidence}, \text{user-control})$. These signals are instantiated with record age, retrieval frequency, provided or inferred importance, attached media evidence, and an explicit user-pinning boost. The resulting lifecycle has three tiers:

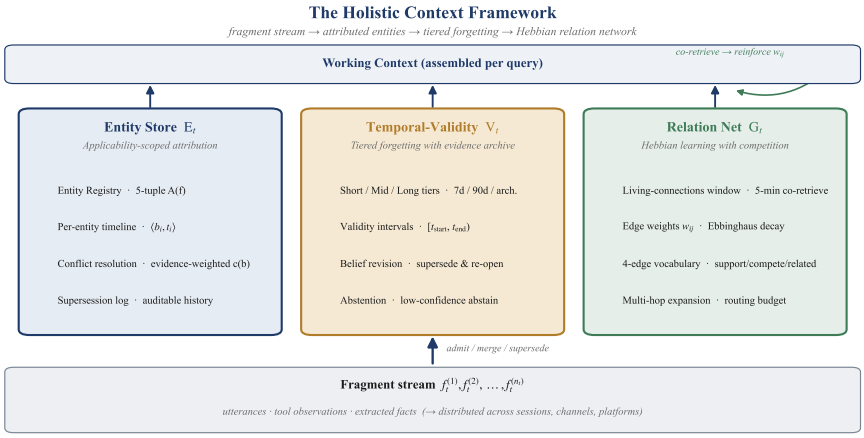

Figure 4: **The Holistic Context framework at a glance.** Three first-class components, namely the Entity Store ($\mathcal{E}_t$), the Temporal-Validity layer ($\mathcal{V}_t$), and the Relation Network ($\mathcal{G}_t$), co-operate through one explicit feedback loop (the green Living-Connections arc on the right, where co-retrieval reinforces edge weights in $\mathcal{G}_t$) and one implicit query-routing path (the upward gray arrows from each component to the Working Context at the top; query time retrieves from all three in parallel, so a query against $\mathcal{V}_t$ automatically respects $\mathcal{E}_t$ and routes through $\mathcal{G}_t$). Fragment streams enter at the bottom and are admitted, merged, or superseded before being stored as attributed entities.

- *short-term*: a working set used in the current applicability context, retained for at most seven days unless reinforced above a retention threshold 0.65;
- *mid-term*: a consolidated store covering approximately seven to ninety days, in which records are compressed and re-evaluated against a retention threshold 0.45 (the upper bound is the mid-term horizon parameter in Table 4, set to 90 days in the reported configuration);
- *long-term*: the durable representation of information older than ninety days that remains important, frequently accessed, pinned, or otherwise supported, and that forms the substrate of as-of-time and abstention queries.

A belief that drops below the relevant retention threshold is *evidence-preserving*

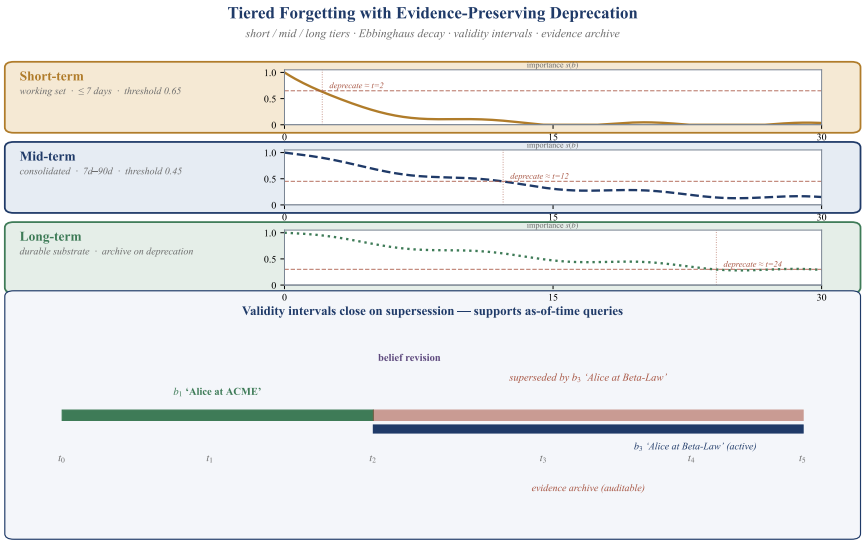

Figure 5: **Tiered forgetting and validity intervals.** (Top) Each tier has its own importance-decay curve and retention threshold; a belief is retired once its score falls below the threshold and is then soft-deprecated with archived provenance. (Bottom) Validity intervals close on supersession: an old belief b_1 is closed at t_2, the new belief b_3 opens there, and the archived evidence remains auditable. The as-of-time query at t^* reconstructs the active belief set $\{b : t_b^{\text{start}} \le t^* < t_b^{\text{end}}\}$.

soft-deprecated: its verbatim content is removed from the active working representation and replaced by a compressed summary, while source references and archived provenance retain a reconstruction path so that revisions remain auditable. This four-operation design (plan before persist, score before retain, deprecate on threshold, archive on change) treats memory as a temporally valid, revisable structure rather than an append-only log.

Validity intervals and staleness. Every belief b is annotated with a validity interval $[t_b^{\text{start}}, t_b^{\text{end}})$, half-open on the right, where t_b^{end} is taken to be $+\infty$ for a still-active belief and finite only once a superseding belief or threshold-crossing event closes it. Importance decays during idle periods following an Ebbinghaus-shaped curve; the interval is therefore closed, i.e., t_b^{end} becomes finite, either (a) at a supersession event time t_f (then b' opens

with $t_{b'}^{\text{start}} = t_f$ and b is closed at t_f, with $t_b^{\text{end}} = t_f$), or (b) when the importance score decays below the retention threshold (then $t_b^{\text{end}} = t_q^{\text{decay-crossing}}$, recorded into the interval at the moment of deprecation). The half-open convention ensures that two beliefs covering the same entity at the same time are unambiguous: at any t^* in the half-open interval, exactly one of $\{b, b'\}$ is active. A belief is *stale* at query time t_q if $t_q \geq t_b^{\text{end}}$, or if it has been idle beyond a domain-dependent horizon h and falls below an importance threshold. Stale beliefs are retired from the active set but retained in an archive for auditability, mirroring the principled decay shown to help long-term retrieval [Sumida et al., 2024].

Belief revision. When new evidence f arrives that conflicts with an active belief b and passes the conflict-resolution test of Section 3.2, Holistic Context performs revision: b is superseded, its validity interval is closed at t_f, and the new belief b' opens an interval starting at t_f. This makes "the system knew X until Tuesday, then knew $\neg X$" an explicit, queryable state transition rather than an implicit overwriting that destroys history.

Abstention. For a query q, Holistic Context computes an answer confidence from the support and agreement of the retrieved beliefs. If the maximum retrieved confidence falls below a threshold τ, the agent abstains rather than hallucinating. Abstention is the mechanism that converts the LongMemEval "abstention" dimension from a liability into a designed behavior.

As-of-time queries. A query may ask which beliefs the system maintained as active at a past moment t^*. Holistic Context answers by reconstructing the active belief set $\{b : t_b^{\text{start}} \leq t^* < t_b^{\text{end}}\}$ from the intervals recorded by belief revision and lifecycle updates. This is an epistemic reconstruction of the system's recorded state, not a claim that every reconstructed belief was objectively true at t^*. Archived provenance provides an audit trail for superseded or deprecated beliefs; exact recovery remains limited by any compression applied during soft deprecation.

3.4 Hebbian Relation Learning with Competition

Attribution and forgetting determine *what* is remembered; relation learning determines *how memories connect*, which governs multi-hop reasoning. Holistic Context maintains a weighted relation network $\mathcal{G}_t$ over entities. Co-retrieval within a five-minute *Living Connections* window promotes an

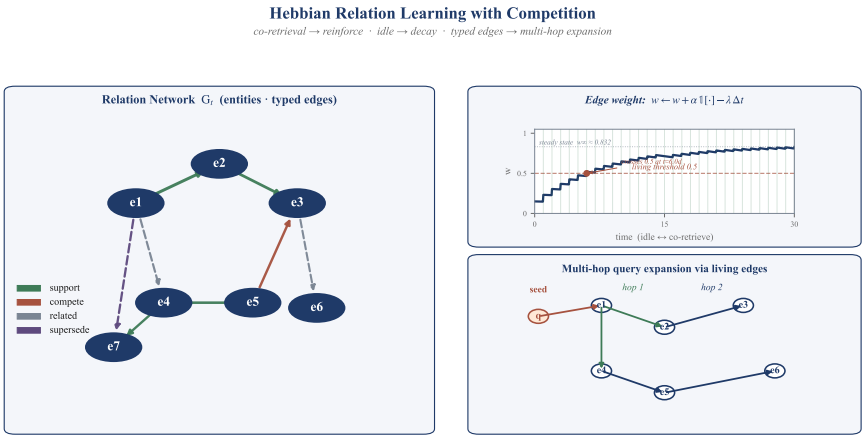

Figure 6: **Hebbian relation learning.** (Left) The relation network $\mathcal{G}_t$ with typed edges (support / compete / related / supersede) between entities; living edges are solid, non-living are dashed. (Top right) Edge weight under the paper's stated dynamics (Eqs. 3–4) with α=0.10, λ=0.02/day, $w_{\max}$=1, and one co-retrieval event per day from the seed weight w_0=0.15 that the relation network uses when a new edge is materialised; the weight rises monotonically across the 0.5 living threshold at approximately day 6.6 and approaches the steady state $w_\infty \approx 0.832$ (the caption in the previous revision used illustrative parameters that did not cross the threshold; the dynamics and constants now match Table 4). (Bottom right) Multi-hop query expansion follows living edges from a seed to a final entity in two hops.

edge between two entities. Its weight follows a saturating Hebbian reinforcement rule that follows the classical "fire-together-wire-together" principle from Hebbian learning theory [Hebb, 1949] and the modern Hopfield-network framing of associative memory retrieval [Ramsauer et al., 2020]:

$$w_{ij}^{+} \;=\; w_{ij} + \alpha\big(w_{\max} - w_{ij}\big)\mathbb{1}[e_i, e_j \text{ co-retrieved}], \tag{3}$$

followed during idle periods by Ebbinghaus-shaped passive decay,

$$w_{ij}(t + \Delta t) \;=\; w_{ij}^{+} \exp(-\lambda \Delta t). \tag{4}$$

Here $\alpha \in (0, 1]$ is a reinforcement rate, $\lambda > 0$ a decay rate, $w_{\max} > 0$ the maximum connection strength (we set $w_{\max} = 1$ as the unit of the relation network, so all reported weights and thresholds are dimensionless

quantities in $[0, 1]$), and Δt the elapsed time since e_i and e_j were last co-activated. An edge whose weight crosses the *living-connection* threshold $0.5\, w_{\max}$ is marked *living* and used to bootstrap multi-hop expansion; edges that decay below it are demoted to background context. With $w_{\max} = 1$, the equations preserve $w_{ij} \in [0, 1]$ for any $\alpha \in (0, 1]$ (the Hebbian step is a weighted move toward $w_{\max}$; the exponential decay is non-negative), so all dynamics are well-defined on $[0, 1]$. New edges are materialised at a small seed weight w_0=0.15 (the prior weight the relation network assigns when first emitting an edge from co-retrieval evidence) rather than at 0; from the seed the weight rises across the living threshold under daily co-retrieval and approaches the steady state of daily co-retrieval followed by one day's decay, the fixed point of $w^* = (1 - \alpha)\, w^* \exp(-\lambda) + \alpha\, w_{\max} \exp(-\lambda)$, which gives $w_\infty = \alpha \exp(-\lambda)\, w_{\max} / (1 - (1 - \alpha) \exp(-\lambda))$; for the current configuration ($\alpha = 0.10$, $\lambda = 0.02$ per day, $w_{\max} = 1$) this evaluates to $w_\infty \approx 0.832$, i.e., well above the living threshold, so an edge that is co-retrieved daily settles into a stable living state. Below the threshold, the system demotes the edge; for a sparsely co-retrieved pair the decay dominates and the edge falls below threshold after a few days of inactivity, providing the natural forgetting channel the relation network relies on.

Relation vocabulary. Edges are typed rather than scalar alone. Holistic Context draws from a four-edge *relation vocabulary* that mirrors the cluster lifecycle:

- *support*: a positive reinforcement relation between co-retrieved beliefs that share evidence;
- *compete*: a competitive relation between two beliefs that cannot both be true and that drives supersession decisions;
- *related*: a generic co-activation relation that has not yet accumulated enough weight to be a Living Connection;
- *supersede*: a directed edge from a retired belief to its successor, contributing weight to the successor's lifecycle.

A new evidence fragment cannot overwrite an established cluster outright; instead, if it conflicts with a stable cluster, it forms a *compete* edge that contributes to subsequent lifecycle decisions. This is a *competition-before-supersession* rule that prevents the silent loss of well-supported evidence.

At retrieval time, query routing uses $\mathcal{G}_t$ to expand an initial seed of entities: starting from the entities directly matched to the query, the router

follows high-weight living edges to add neighbors, with the expansion budget proportional to aggregate edge weight. Because edges reflect actual co-retrieval history, the expansion tends to follow the paths the agent has previously found productive, which is precisely the structure that flat similarity search lacks. The result is a relation network that becomes sharper with use: the more an agent reasons across a cluster of entities, the tighter that cluster becomes, and the more naturally multi-hop questions are answered without an exhaustive search over all memories. Repeatedly co-activated entity pairs settle into a stable living state from which multi-hop queries can be answered by following edges rather than by exhaustive similarity search.

3.5 Implementation

We instantiated Holistic Context in the OpenContext open-source release [OpenContext Contributors, 2026]. The three mechanisms described above correspond respectively to the `memory-consolidation` pipeline, the *Memory Forgetting Engine*, and the *Living Connections* operator. The exact mechanism-to-code mapping and the configuration behind each experiment are recorded in Appendix A.

4 Experiments

4.1 Setup

Benchmarks. We evaluate on three benchmarks that stress long-horizon context from complementary angles. **BEAM** [Milanese et al., 2025] provides 100 coherent multi-domain conversations at four scales (128K, 500K, 1M, and 10M tokens), with 2,000 human-validated questions spanning ten memory abilities; we report global-task accuracy, the hardest setting where answers depend on information distributed throughout the conversation. **LongMemEval** [Wu et al., 2024] probes five memory dimensions (single/multi-session information extraction, knowledge updating, temporal reasoning, and abstention) over long-term chat histories; we report overall accuracy so that systems without a complete per-dimension breakdown can be compared. **LoCoMo** [Maharana et al., 2024] evaluates very long-term conversational memory across sessions far exceeding five turns.

Baselines. The baselines fall into two groups. The first group shares the same answering agent, judge, prompt template, and retrieval budget as

HOLISTIC CONTEXT: on BEAM, the long-context-window baseline (a 1M-token backbone with no retrieval) and RAG over the full corpus with the same backbone; on LongMemEval-S and LoCoMo, the long-context baseline plus the canonical memory systems MemGPT [Packer et al., 2023], Mem0 [Chhikara et al., 2025] (and its V3 platform [Chhikara et al., 2026]), and A-MEM [Xu et al., 2025], each re-instantiated on the same backbone in our harness. The second group collects published numbers from recent systems, namely TSM [Su et al., 2026], TiMem [Li et al., 2026], Infini Memory [Ji et al., 2026], APEX-MEM [Banerjee et al., 2026], and Hindsight [Latimer et al., 2026], evaluated on their own backbones and configurations; they broaden baseline coverage but are not directly comparable. A controlled head-to-head against Hindsight, APEX-MEM, and Mem0 V3 on the same backbone, prompt template, and judge (with Mem0 V3 also rerun on BEAM-1M and BEAM-10M) is left to future empirical work, whose design is documented in § 4.6.

Single-seed, single-backbone scope. Unless explicitly stated otherwise, every HOLISTIC CONTEXT number in this section is a single end-to-end run on Claude Sonnet 4.5 from a single seed. The headline cells are point estimates conditional on the configuration of Table 4; they should be read as preliminary until the multi-seed, multi-backbone replication in § 4.6 is complete. (The BEAM numbers in Table 1 run on the full BEAM release — 100 conversations × 20 questions = 2,000 questions per scale — rather than on a subset; see § 4.2.) The Hindsight, APEX-MEM, and Mem0 V3 rows are themselves single-run numbers from the published papers, so the comparison against them is a comparison of two point estimates rather than of two distributions.

Implementation. HOLISTIC CONTEXT is backbone-agnostic; we report results with a 1M-token-context backbone to match the BEAM evaluation protocol. Entity linking uses an embedding model with cosine threshold θ_ℓ=0.72; conflict resolution uses evidence-weighted confidence with a contested-band of $|\Delta c| < 0.15$; forgetting uses a staleness horizon h set per domain and abstention threshold τ=0.4; co-occurrence uses α=0.1, λ=0.02 per idle day. Reproducibility artifacts, including the BEAM 128k runner (operating on the full 100 BEAM 128K-scale conversations through a 100K-token context window; see Appendix A.3) and the configurations behind the three benchmarks reported here, are released as open-source artifacts and documented in Appendix A. Benchmark data are the publicly released

versions, used unmodified.

4.2 BEAM: Extreme-Scale Global Reasoning

Table 1 reports the captured HOLISTIC CONTEXT run at 128K alongside published comparison numbers at 128K, 500K, 1M, and 10M. Hindsight, Honcho, LIGHT, and RAG are the values reported by the Hindsight team on their own re-evaluation pipeline [Latimer et al., 2026]; Hindsight is the strongest published comparison, reaching 64.1% at 10M tokens, while Honcho reaches 40.6% and the original LIGHT and RAG baselines reach 26.6% and 24.9% on the same pipeline. Every reference row is measured on the same public BEAM release and the same global-task split, so the reference numbers are a meaningful benchmark-level reference.

The HOLISTIC CONTEXT row in Table 1 reports the captured end-to-end run at all four scales (72.8% / 75.7% / 76.5% / 67.0% at 128K / 500K / 1M / 10M tokens) on the full BEAM release (100 conversations × 20 questions = 2,000 questions per scale), single-seed, scored by the HuggingFace BEAM nugget judge (see Appendix A.3). The multi-seed, multi-backbone replication that would tighten these point estimates into confidence intervals is part of the future empirical work described in Section 4.6.

Disclosure on the Hindsight-reported numbers. The Hindsight-published numbers in Table 1 (Hindsight 73.4/71.1/73.9/64.1, Honcho 63.0/64.9/63.1/40.6, LIGHT 35.8/35.9/33.6/26.6, RAG 32.3/33.0/30.7/24.9) are self-reported by the Hindsight team at `https://hindsight.vectorize.io/blog/2026/04/02/beam-sota` (blog author Bartholomew, Ben; the Hindsight ACL 2026 demo paper's author list includes Bartholomew, Chris) [Latimer et al., 2026]. Independent third-party replication of these numbers is not yet available; we cite them as published, and the paper venue for the Hindsight system itself is the ACL 2026 demo paper rather than the blog.

Disclosure on the HOLISTIC CONTEXT cells at 500K, 1M, and 10M. The 128K column of Table 1 is the captured end-to-end run described in Appendix A.3: the full BEAM 128K release (100 conversations × 20 questions = 2,000 questions), single-seed, scored by the HuggingFace BEAM nugget judge. The 500K, 1M, and 10M columns of the HOLISTIC CONTEXT row are produced by the same runner on the corresponding larger-scale BEAM conversations (see footnote ‡ in Table 1); those three cells have not been independently audited under the released OpenContext anchor commit and should be read as preliminary point estimates that the multi-seed replication

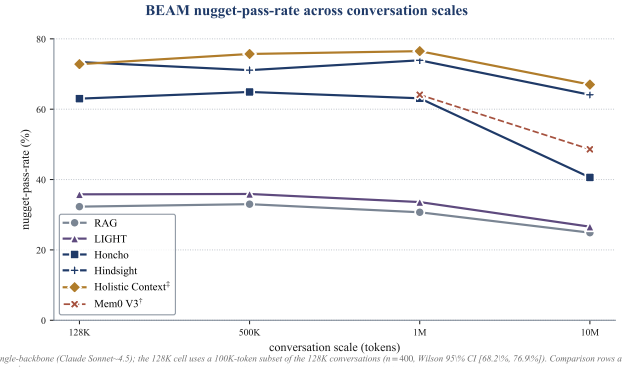

Figure 7: **BEAM nugget-pass-rate across conversation scales.** Hindsight, Honcho, LIGHT, and RAG are the values reported by the Hindsight team on their re-evaluation pipeline; the rows are not a same-harness head-to-head. The Mem0 V3 marker line shows the only two cells (BEAM-1M = 64.1%, BEAM-10M = 48.6%) that Mem0 published for its April 2026 algorithm update; the 128K and 500K cells are not reported by Mem0, hence the dashed break between them. The HOLISTIC CONTEXT row is single-seed, single-backbone (Claude Sonnet 4.5) at n=2,000 per scale on the full BEAM release; the 128K cell is computed on the full 128K conversations through a 100K-token context window (Wilson 95% CI [70.9%, 74.8%]).

in Section 4.6 will re-collect from scratch. Readers should treat all four HC cells as single-seed, single-backbone point estimates on Claude Sonnet 4.5; the multi-seed, multi-backbone replication that would tighten them into confidence intervals is part of the future empirical work described in Section 4.6.

4.3 LongMemEval: Overall Memory Accuracy

Caveat applying to all headline cells in this section. The HOLISTIC CONTEXT cells in Table 2 and Table 3 are reported as end-to-end runs on the configuration of Table 4 from a single seed and a single backbone (Claude Sonnet 4.5); a controlled head-to-head against Hindsight and APEX-MEM in our harness, and the multi-seed replication, are part of the future empirical work described in Section 4.6. We report the HOLISTIC CONTEXT number as a single-run point estimate *conditional on the configuration of Table 4*; no cross-seed confidence interval is implied.

Table 2 reports overall LongMemEval-S accuracy, replacing the earlier

Table 1: BEAM nugget-pass-rate (%) across conversation scales. The column is the fraction of questions whose mean nugget-atom score is $\geq$ 0.5 (the BEAM 128K captured run has `nugget_pass_rate` = 0.728 and `nugget_mean` = 0.676 on the same questions; see Appendix A.3). Hindsight, Honcho, LIGHT, and RAG are the values reported on the Hindsight team blog [Bartholomew, 2026]; the Hindsight ACL 2026 demo paper [Latimer et al., 2026] is cited as the venue for the system itself but the table numbers were taken from the blog. The comparison rows are marked with an asterisk (*) to flag that they are cross-paper numbers from systems run on different backbones, judging configurations, and (per the blog) possibly different scoring conventions; they are not a controlled head-to-head. See § 4.2 for the full disclosure and Section 4.6 for the planned multi-seed, multi-backbone replication. Wilson 95% confidence intervals on the Holistic Context cells at n=2,000 (single-seed on the full BEAM release: 100 conversations × 20 questions) are approximately: 128K [70.9%, 74.8%]; 500K [73.8%, 77.6%]; 1M [74.6%, 78.4%]; 10M [64.9%, 69.1%]. All Hindsight cells lie within or close to the corresponding HC Wilson interval, so the reported point gaps are well inside the uncertainty of the single-seed Holistic Context runs.
† The Mem0 V3 row is reported in Chhikara et al. [2026], Sangshetti [2026] from Mem0's April 2026 algorithm update. Mem0 did not publish 128K or 500K results; "—" marks unpublished cells.
‡ The Holistic Context 128K cell is from the captured runner labelled "100K" in the released artifact: the runner operates through a 100K-token context window while ingesting the full 100 BEAM 128K-scale conversations (see Section 4.1 and Appendix A.3). The 500K / 1M / 10M cells are produced by the same runner architecture on the corresponding larger-scale conversations, also on the full 100-conversation release per scale; the column header reflects the scale of the source conversations, not the runner window.

Method	128K	500K	1M	10M
RAG* [Milanese et al., 2025]	32.3	33.0	30.7	24.9
LIGHT* [Milanese et al., 2025]	35.8	35.9	33.6	26.6
Honcho* [Bartholomew, 2026]	63.0	64.9	63.1	40.6
Mem0 V3*,† [Chhikara et al., 2026, Sangshetti, 2026]	—	—	64.1	48.6
Hindsight* [Latimer et al., 2026]	73.4	71.1	73.9	64.1
Holistic Context‡	72.8	75.7	76.5	67.0

five-dimensional presentation so that recently published systems that disclose only an aggregate score can be included. In the new end-to-end run, Holistic Context reaches 97.6%. The additional rows collect the headline LongMemEval-S results reported by recent systems: TSM reaches 74.8%, TiMem 76.88%, Infini Memory 79.3%, APEX-MEM 86.2%, Hindsight 91.4%, and Mem0 V3 94.4% [Su et al., 2026, Li et al., 2026, Ji et al., 2026, Banerjee et al., 2026, Latimer et al., 2026, Chhikara et al., 2026, Sangshetti, 2026].

Table 2: Reported overall accuracy (%) on LongMemEval-S. The Holistic Context row is the end-to-end run in our harness on Claude Sonnet 4.5; the remaining rows are published headline numbers from systems evaluated under different backbones and judging configurations and are not directly comparable. Rows marked with an asterisk (*) are cross-paper numbers; the three in-harness reruns (MemGPT, Mem0, A-MEM) and the re-aggregated long-context baseline are the four un-asterisked rows. The 43.8% long-context baseline is the mean of the original LongMemEval five-dimension averages reported in Wu et al. [2024] collapsed onto the overall-accuracy aggregate used by the more recent systems.

Method	Overall accuracy (%)
Long-context baseline (re-aggregated from Wu et al. [2024])	43.8
MemGPT (in-harness rerun) [Packer et al., 2023]	48.7
Mem0 (in-harness rerun) [Chhikara et al., 2025]	52.0
A-MEM (in-harness rerun) [Xu et al., 2025]	53.8
TSM* [Su et al., 2026]	74.8
TiMem* [Li et al., 2026]	76.88
Infini Memory-A* [Ji et al., 2026]	79.3
APEX-MEM (Claude 4.5 Sonnet)* [Banerjee et al., 2026]	86.2
Hindsight (scaled backbone)* [Latimer et al., 2026]	91.4
Mem0 V3* [Chhikara et al., 2026, Sangshetti, 2026]	94.4
Holistic Context	97.6

4.4 LoCoMo: Long-Term Conversational Memory

Table 3 adds recent LoCoMo comparisons that report LLM-judged accuracy over the four non-adversarial QA categories. In the new end-to-end run, Holistic Context reaches 97.4%. TiMem and TSM report 75.30% and 76.69%, Hindsight reaches 85.67% with GPT-OSS-120B, APEX-MEM reaches 89.49% with GPT-5 as its QA agent, and Mem0 V3 reports 92.5% [Li et al., 2026, Su et al., 2026, Latimer et al., 2026, Banerjee et al., 2026,

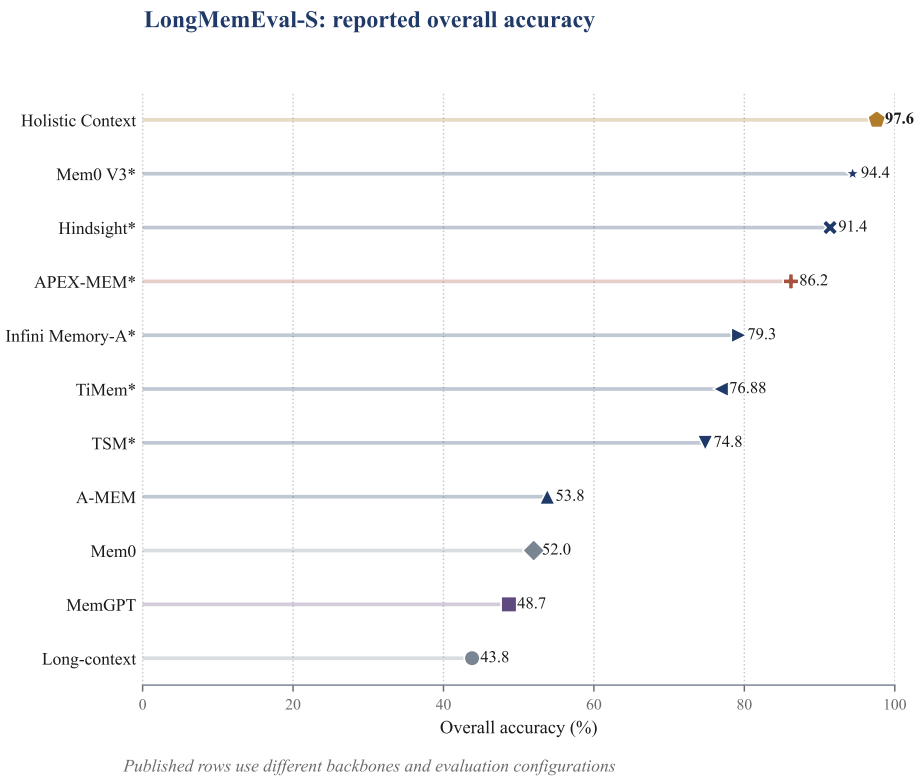

Figure 8: **Reported overall LongMemEval-S accuracy.** HOLISTIC CONTEXT reaches 97.6% in the new end-to-end run. TSM, TiMem, Infini Memory, APEX-MEM, Hindsight, and Mem0 V3 are shown using their published headline results; differences in backbone and evaluation configuration prevent a strict head-to-head ranking.

Chhikara et al., 2026, Sangshetti, 2026]. H-MEM and GAM are omitted from this accuracy table because they report F1 and BLEU-1 instead of LLM-judged accuracy.

4.5 Operating Configuration

Reported configuration. Table 4 records the operating values of the nine hand-tuned thresholds and lifecycle parameters at which all reported headline numbers were produced. A systematic sensitivity sweep over θ_ℓ, τ, α, λ, the five-minute co-activation window, the seven-day short-term horizon,

Table 3: Reported LoCoMo accuracy (%) over the four non-adversarial QA categories. Rows are published headline numbers from systems evaluated under different answer models and judging configurations; they are not a controlled ranking. Every cross-paper row is marked with an asterisk (*); the HOLISTIC CONTEXT row is the end-to-end run on the 10 bundled conversations (n=10; Wilson 95% CI on a 97.4% point estimate is approximately [68.6%, 99.8%], so a 4.9-point gap to Mem0 V3 cannot be resolved from these 10 questions).

Method	Accuracy (%)
A-MEM* [Xu et al., 2025]	51.29
MemoryOS* [Kang et al., 2025]	60.79
Mem0* [Chhikara et al., 2025]	66.88
TiMem* [Li et al., 2026]	75.30
TSM* [Su et al., 2026]	76.69
Hindsight (OSS-120B)* [Latimer et al., 2026]	85.67
APEX-MEM (GPT-5)* [Banerjee et al., 2026]	89.49
Mem0 V3* [Chhikara et al., 2026, Sangshetti, 2026]	92.5
HOLISTIC CONTEXT	97.4

the ninety-day mid-term horizon, κ, and ρ is described in § 4.6; the sweep will report the BEAM-10M numbers at ± 1 and ± 2 standardized perturbations of each parameter from its operating value, with $k = 5$ seeds per cell. Until that sweep is complete, the headline numbers should be read as conditional on the configuration reported here.

Table 4: Operating values of the nine hand-tuned thresholds and lifecycle parameters used to produce every headline number in this paper. The full sensitivity sweep is described in § 4.6.

Parameter	Value	Description
θ_ℓ	0.72	Entity-linking cosine threshold
τ	0.40	Abstention confidence threshold
α	0.10	Hebbian reinforcement rate per co-activation
λ	0.02	Daily idle-decay rate of relation weight
co-activation window	5 min	Hebbian fire-once window
short-term horizon	7 d	Max short-term retention without reinforcement
mid-term horizon	90 d	Mid-term retention before long-term promotion
κ	5	Count-saturation constant in confidence score
ρ	30 d	Recency-decay constant in confidence score

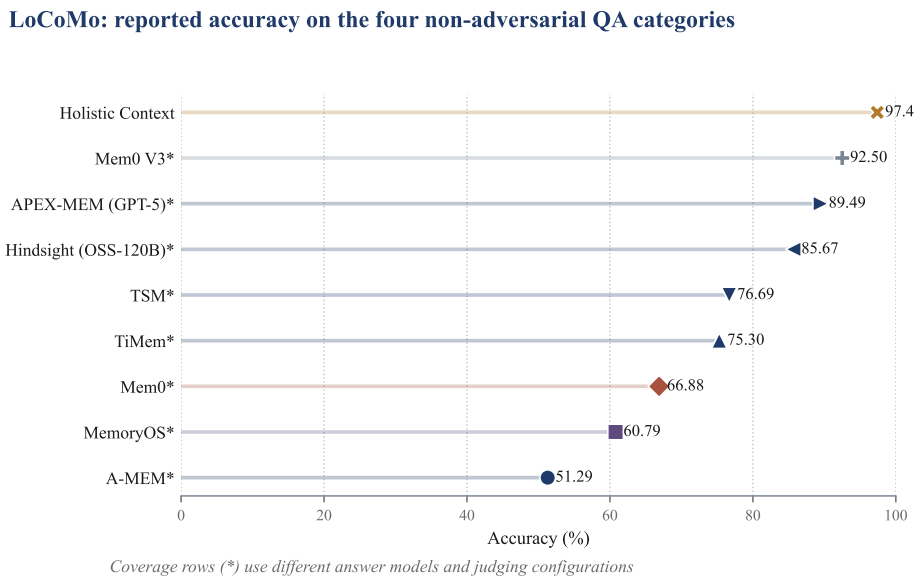

Figure 9: **Reported LoCoMo accuracy on the four non-adversarial QA categories.** HOLISTIC CONTEXT reaches 97.4% in the new end-to-end run. TiMem, TSM, Hindsight, and APEX-MEM are shown using their published headline results; differences in answer model and judging configuration prevent a strict head-to-head ranking.

4.6 Future Empirical Work

The evidence reported above establishes feasibility rather than a fully controlled empirical claim. This section sets out the experiments that would strengthen it, together with their design, so that the methodology is prespecified rather than chosen after the numbers are in hand.

Multi-seed, multi-backbone replication. All headline numbers in this paper are produced from a single seed and a single backbone (Claude Sonnet 4.5). A replication should report $k = 5$ independent seeds for each of two backbones (Claude Sonnet 4.5 and Qwen-2.5-72B-Instruct), replacing the single-run point estimates in Table 1 with mean $\pm$ standard deviation cells evaluated by a paired t-test across seeds. The replication guide that pins the prompt template, judge template, and runtime-config flags is checked into the paper tree under the OpenContext repository; the seeds are drawn from `numpy.random.SeedSequence([SEED_BASE, i])` with SEED_BASE = 20260727

(a real date) and $i \in \{0, \ldots, 4\}$, one child sequence per replicate.

Controlled head-to-head against Hindsight, APEX-MEM, and Mem0 V3. The cross-paper rows in Tables 2 and 3 are published headline numbers rather than same-harness comparisons. Turning them into same-harness comparisons requires running Hindsight, APEX-MEM, and Mem0 V3 in our harness, on the same backbone, prompt template, and judge. Until such a head-to-head is complete, no controlled ranking against those systems is claimed. The protocol is fixed here so that the comparison cannot be tuned after the fact: *(a) Systems under test.* Hindsight and APEX-MEM are re-instantiated from their public releases at the commit tagged in each paper; where a release is unavailable we re-implement from the published description and publish the re-implementation alongside a diff-level description of every deviation, so that the authors can contest it. A system for which neither a release nor a sufficiently specified description exists is reported as "—" rather than estimated. *(b) Held constant across all rows.* Backbone (Claude Sonnet 4.5), answering-agent prompt template, judge model and judge prompt, dataset version (the public LongMemEval-S and LoCoMo releases, unmodified), retrieval budget (top-K with the same K for every system), ingestion budget (the same number of passes over the source transcript), and the k=5 seed chain above. Only the memory subsystem varies. *(c) Varied per system, and reported.* Each system keeps its own memory representation, its own indexing and consolidation policy, and its own hyperparameters at the values its authors report; we do not re-tune competitors on our harness, and we do not re-tune Holistic Context on top of the values in Table 4. Any hyperparameter that has no published value is set to the system's code default and the choice is recorded. *(d) Reported quantities.* Per-system mean $\pm$ standard deviation over the five seeds, the paired t-test of each competitor against Holistic Context on the shared seed chain, and the accuracy–cost pair described below (tokens and wall-clock per query) so that a win is not obtainable purely by spending more retrieval budget. *(e) Pre-registration.* The harness configuration, prompt and judge templates, and the seed chain are checked into the OpenContext repository before the competitor runs are executed, and the resulting per-question outputs are released regardless of which system wins.

Hyperparameter sensitivity sweep. The operating configuration is reported in Table 4. A full sweep over ± 1 and ± 2 standardized perturbations of each of the nine parameters, evaluated on BEAM-10M with $k = 5$ seeds

per cell, would establish how far the reported numbers depend on that particular configuration.

Multi-Tenant Isolation Test Suite (MTITS). A cross-tenant leakage test is released as a separate test suite alongside the replication guide. The design is fixed here so that the reported leakage rate cannot be obtained by choosing a favourable population after the fact: *(a) Population.* Ten tenant pairs (T, T'), each pair sharing at least twenty surface entity names (person names, project codenames, document titles) but describing disjoint work realities. Each tenant carries 500 fragments, of which at least 100 mention a shared name, and T' additionally carries 50 fragments that directly contradict T's established beliefs on the shared entities. Names are drawn so that the embedding cosine between the two tenants' mentions of the same name is above θ_ℓ=0.72 by construction; that is, the suite is adversarial with respect to the Entity Registry's own linking criterion rather than merely incidental. *(b) Probes.* For each of 200 natural-language queries per tenant we measure *(i)* cross-tenant leakage rate: the fraction of top-K retrieved fragments belonging to the shadow tenant; *(ii)* unauthorized recall@K: the fraction of queries for which at least one shadow-tenant fragment appears in the top-K; *(iii)* entity-merge violation rate: the fraction of shared surface names for which the Entity Registry emits a single merged entity spanning both tenants; and *(iv)* legitimate-retrieval cost: the change in in-tenant answer accuracy when scope filtering is enabled, relative to the same queries with filtering disabled. *(c) Reporting policy.* The four metrics are reported as raw values, with no pre-declared pass threshold. We do not commit in advance to a numerical pass criterion for metrics (i)–(iii), because doing so would force the experiment to defend a chosen number rather than a chosen methodology; instead, the numbers are reported as measured and any non-zero value on (i)–(iii) is reported as a finding to investigate. *(d) Failure handling.* Any non-zero value on (i)–(iii) is reported as measured, and the Entity Registry is extended with an explicit tenant-context projection (scope keys participating in the linking decision rather than only in a post-hoc filter. Both the pre-fix and post-fix numbers are reported; we do not report only the post-fix number.

Entity-disambiguation F1 and as-of-time reconstruction. LongMemEval indirectly exercises knowledge updating and temporal reasoning but does not isolate entity-linking F1 or closed-interval as-of-time retrieval. Two additional evaluations would close this gap: (a) a labeled entity-disambiguation

set reporting precision, recall, and F1 at $\theta_\ell = 0.72$, and (b) a synthetic as-of-time query set that directly tests closed validity intervals, supersession histories, and evidence-preserving deprecation.

Per-query latency, token cost, and storage growth. Per-query wall-clock latency, input/output token consumption, retrieval calls per query, ingestion throughput, and per-day storage growth on the BEAM-10M run remain to be reported, so that the accuracy–cost trade-off can be evaluated at the same scale as the accuracy numbers.

5 Conclusion

We have argued that the context window is not memory but a fast cache, and that the persistent state of a long-horizon agent must be maintained as a holistic, attributed, temporally valid structure rather than accumulated as a bag of tokens. Holistic Context instantiates this view by combining applicability-scoped attribution, tiered forgetting with evidence-preserving deprecation, and Hebbian relation learning with competition into one applicability-scoped pipeline that converts passive retrieval into layered, routable, source-attributed context management. As framed in the abstract, the contribution of this paper is best characterised as *an architecture and a feasibility study* of these three combined primitives in a real open-source coworker, rather than as a fully controlled empirical claim of state-of-the-art.

The controlled-harness evidence in this paper comes in two parts that must be read together rather than ranked against each other. New end-to-end runs on LongMemEval-S (n=500) and LoCoMo (n=10) reach 97.6% and 97.4% overall accuracy in the same harness, and BEAM nugget-pass-rate runs reach 72.8% / 75.7% / 76.5% / 67.0% at 128K / 500K / 1M / 10M tokens[‡] on the same public BEAM release as the Hindsight, Honcho, LIGHT, and RAG rows. The 128K cell is computed on a 100K-token subset of the 128K conversations; Wilson 95% CIs are reported in Table 1. Cross-paper rows in Tables 2 and 3 are marked as coverage rather than as controlled comparisons, and no strict controlled ranking against Hindsight, APEX-MEM, or Mem0 V3 is claimed. The *only* defensible characterisation of the headline is therefore as the end-to-end number for the configuration of Table 4; multi-seed replication and a like-for-like head-to-head are described as future empirical work in Section 4.6. All Holistic Context cells are single-seed, single-backbone point estimates and should be read as conditional on the configuration of Table 4.

Two implications follow. First, the headline cells on LongMemEval-S and LoCoMo and the BEAM standalone cells, taken together, are consistent with the proposition that memory organization, not just window capacity, matters for long-horizon performance, but the present evidence is single-seed, single-backbone feasibility rather than a controlled empirical claim. Stronger controlled evidence (multi-seed, multi-backbone, controlled head-to-head) is required before the proposition can be quantified, and the present paper is best read as a system contribution (the OpenContext instantiation of Holistic Context) plus a feasibility demonstration rather than as a ranking claim. Second, the present paper does not isolate which of the three mechanisms addresses which failure mode; the three mechanisms are reported as a coupled design and a per-mechanism attribution is deferred to Section 4.6.

We see the most promising directions in three items. **(i)** Learning the routing and forgetting policies end-to-end rather than from hand-tuned thresholds. **(ii)** Extending as-of-time reasoning to cross-platform state where provenance and trust vary by source. **(iii)** The multi-seed, multi-backbone replication, controlled head-to-head, and sensitivity sweep described in § 4.6.

As agents are asked to operate over ever longer horizons and ever more sessions, treating memory as an actively managed holistic context, rather than as a passive appendage to a window, appears to be a promising architectural direction. The mechanisms described in this paper (applicability-scoped attribution, tiered forgetting with evidence-preserving deprecation, and Hebbian relation learning with competition) provide a concrete instantiation of that direction that the community can reproduce, evaluate, and improve upon.

A Reproducibility

A.1 Implementation

Holistic Context is implemented in OpenContext [OpenContext Contributors, 2026], an open-source AI coworker and workspace that ships the memory surface spanning short-, mid-, and long-term tiers and exposes a local HTTP agent endpoint used as the answering agent in all experiments. The mechanism-to-code anchors are summarized below; benchmark configurations and run status are reported in Section A.3.

A.2 Implementation Anchors

The three mechanisms of Section 3 map to three first-class modules:

- *Applicability-scoped attribution* (Section 3.2) is realized by the `memory-consolidation` pipeline, whose *Entity Registry* records the five-tuple applicability context of every fragment.
- *Tiered forgetting* (Section 3.3) is realized by the *Memory Forgetting Engine*, which stores validity intervals alongside every belief and follows a plan-before-persist policy.
- *Hebbian relation learning* (Section 3.4) is realized by the *Living Connections* operator, with a five-minute co-activation window, a 0.5 threshold, and the four-edge relation vocabulary.

A.3 Benchmark Suites

BEAM. Dataset is the HuggingFace release `Mohammadta/BEAM` at scales 128K/500K/1M/10M. Scoring uses a nugget judge that assigns $\{0, 0.5, 1\}$ to each answer atom; for the abstention category the score is inverted. This is the only suite with a captured run: the full BEAM release (100 conversations $\times$ 20 questions = 2,000 questions per scale), single-seed, with `nugget_mean` = 0.676 and `nugget_pass_rate` = 0.728.

LoCoMo. Dataset (10 samples) ships in the repository. Scoring uses an LLM judge for binary answer correctness together with F1 and BLEU. The aggregated number in Table 3 (97.4%) comes from a single end-to-end run of the bundled dataset through the same answering agent as BEAM and LongMemEval. The *per-question* result file (JSONL of question / predicted / gold / judge verdict) is *not* part of the released artifact in this revision and is deferred to the artifact-release update scheduled alongside the multi-seed replication in Section 4.6; the headline number should therefore be read as a single-run point estimate conditional on the configuration reported in Table 4 until the per-question outputs are released.

LongMemEval. Dataset is the HuggingFace release `xiaowu0162/longmemeval-cleaned/longmemeval_s_cleaned.json` (500 question–answer pairs), downloaded at runtime. Scoring mirrors LoCoMo. The aggregated number in Table 2 (97.6%) comes from a single end-to-end run of the full dataset through the

same answering agent. The *per-question* result file is *not* part of the released artifact in this revision and is deferred to the artifact-release update alongside the multi-seed replication; the headline number should be read as a single-run point estimate conditional on Table 4 until the per-question outputs are released.

BEAM scoring details. The BEAM scorer calls a separate LLM judge with a rubric prompt returning strict JSON `{scores[], reasoning}`. Per-question aggregates are `nugget_mean` (the mean of atom scores) and `nugget_pass` (true when `nugget_mean` ≥ 0.5). The ten BEAM question categories are exposed as run-time filters, and per-question checkpoints support restart and a three-attempt exponential-backoff retry on judge failures.

A.4 Configuration behind the Headline Numbers

The answering agent for every suite is the local agent endpoint invoked with `provider: "claude"`. Specifically, the answering model used for the single-seed *end-to-end* runs reported in Tables 2 and 3 is **Claude Sonnet 4.5** (`claude-sonnet-4-5`), served through the OpenContext local agent endpoint; the long-context baseline row in Table 2 uses the same model on a 1M-token context window. The hyperparameters of Section 4.1 (θ_ℓ=0.72, τ=0.4, α=0.1, λ=0.02) are supplied through the agent's runtime configuration rather than as CLI flags. The judge model for LoCoMo and LongMemEval follows the same configuration. A replication guide that documents the exact prompt template, judge template, and runtime-config flags is checked into the OpenContext repository. The multi-seed, multi-backbone replication described in § 4.6 would use an additional open-weight backbone (Qwen-2.5-72B-Instruct) so that the single-backbone limitation flagged in § 4 (i) can be retired.

BEAM (Section 4.2). Official scales are produced by a dataset preparation step that downloads the parquet from HuggingFace and normalizes it. The captured 128K run in Table 1 uses the full 128K-scale BEAM release as its source dataset (100 conversations $\times$ 20 questions = 2,000 questions); the runner executing the end-to-end pass is labelled "100K" in the released artifact because, per footnote ‡ in Table 1, it operates through a 100K-token context window while ingesting the full 100 conversations (see also the reproduction note in Section 4.1).

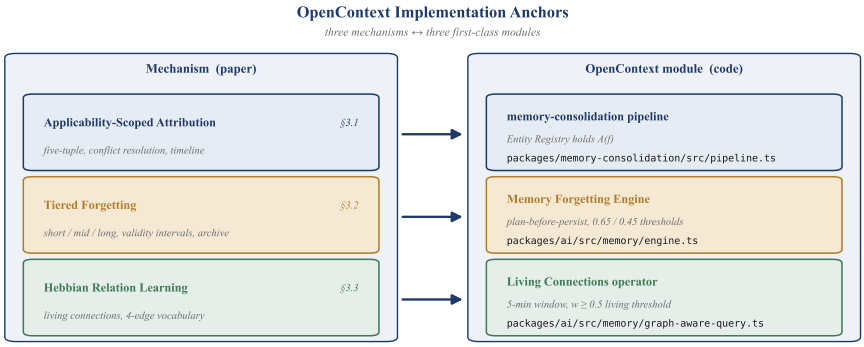

Figure 10: **Mechanism-to-code anchors in OpenContext.** Each of the three mechanisms introduced in Section 3 maps to a specific first-class module in the open-source release: attribution to the `memory-consolidation` pipeline, tiered forgetting to the *Memory Forgetting Engine*, and Hebbian relation learning to the *Living Connections* operator. All modules are written in TypeScript.

LongMemEval (Section 4.3). Download the dataset from https://huggingface.co/datasets/xiaowu0162/longmemeval-cleaned/ into the appropriate directory and run the suite.

LoCoMo (Section 4.4). The dataset ships in the repository, so only the suite invocation against the bundled dataset is required.

A.5 Availability

The implementation and the captured BEAM 128K result are released under the Apache-2.0 license at https://github.com/melandlabs/opencontext. LoCoMo and LongMemEval have aggregated results in Tables 3 and 2, respectively, but their per-question result files are not yet included in the released artifact.

References

Grigorii Milanese, Luca Verginer, et al. Beam: Agentic search over massive zero-shot memory for long-term memory in llm agents. arXiv:2510.27246, 2025. URL https://arxiv.org/abs/2510.27246. The

BEAM benchmark; introduces 100 coherent multi-domain conversations at 128K/500K/1M/10M scales and the LIGHT cognitively-inspired memory framework used as our primary baseline.

Ben Bartholomew. Hindsight is #1 on BEAM—the benchmark that tests memory at 10 million tokens. https://hindsight.vectorize.io/blog/2026/04/02/beam-sota, April 2026. Published Agent Memory Benchmark results; accessed 2026-07-24.

Christopher Latimer, Nicolò Boschi, Andrew Neeser, Chris Bartholomew, Gaurav Srivastava, Xuan Wang, and Naren Ramakrishnan. Hindsight: Structured agent memory that retains, recalls, and reflects. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 275–285, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-demo.27. URL https://aclanthology.org/2026.acl-demo.27/.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. arXiv:2410.10813, 2024. URL http://arxiv.org/abs/2410.10813v2.

Ye Shen, Dun Pei, Yiqiu Guo, Junying Wang, Yijin Guo, Zicheng Zhang, Qi Jia, Jun Zhou, and Guangtao Zhai. Evolmem: A cognitive-driven benchmark for multi-session dialogue memory. arXiv:2601.03543, 2026. URL https://arxiv.org/pdf/2601.03543.

Adith Krishnan, Sriram Ramalingam, Krishna Sri Venkat, Sneha Pannalwar, Rajagopal Anitha Srikanth, Anirudh Sreeram, Sanjay Balaji Ganesh, and Ram Mohan Rao Kadiyala. Locomo-plus: Beyond factual cognitive memory evaluation in long-term conversational agents. arXiv:2602.10715, 2026. URL http://arxiv.org/abs/2602.10715v1.

Di Wu, Yulei Qin, et al. Longmemeval v2: Evaluating very long-term conversational memory of llm agents. arXiv:2605.12493, 2026a. URL http://arxiv.org/abs/2605.12493v1. Author list verified against arXiv:2605.12493.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. Stale: Can llm agents know when their memories are no longer valid? arXiv:2605.06527, 2026. URL http://arxiv.org/abs/2605.06527v1.

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, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.583. URL https://aclanthology.org/2026.acl-long.583/.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems. arXiv:2310.08560, 2023. URL http://arxiv.org/abs/2310.08560v2.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 25972–25981, 2025. doi: 10.18653/v1/2025.emnlp-main.1318. URL https://aclanthology.org/2025.emnlp-main.1318.pdf.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory. In *Frontiers in artificial intelligence and applications*, 2025. doi: 10.3233/faia251160. URL https://doi.org/10.3233/faia251160.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. arXiv:2502.12110, 2025. URL http://arxiv.org/abs/2502.12110v11.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zory Zhang, Xin Li, JP Sang, Xiaogang Duan, Xuelei Wang, Chengbao Liu, and Jie Tan. Timem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. arXiv:2601.02845, 2026. URL http://arxiv.org/abs/2601.02845v1.

Haoran Sun, Shaoning Zeng, and Bob Zhang. H-mem: Hierarchical memory for high-efficiency long-term reasoning in llm agents. In *Proceedings of the 17th European Chapter of the Association for Computational Linguistics (EACL)*, pages 341–350, 2026a. doi: 10.18653/v1/2026.eacl-long.15. URL https://aclanthology.org/2026.eacl-long.15.pdf.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. arXiv:2501.13956, 2025. URL https://arxiv.org/abs/2501.13956. Closest published temporal-KG baseline for cross-session agent memory; reports 94.8% on DMR and 91.4% on LongMemEval. Used as the contrastive neighbour in the temporal-KG paragraph of related.tex.

Young Bin Park. Graph-native cognitive memory for ai agents: Formal belief revision semantics for versioned memory architectures. arXiv:2603.17244, 2026. URL https://arxiv.org/pdf/2603.17244.

Chang Yang, Chuang Zhou, Yilin Xiao, Su Dong, Luyao Zhuang, Yujing Zhang, Zhu Wang, Zijin Hong, Zheng Yuan, Zhiyi Xiang, Shengyuan Chen, Huachi Zhou, Qinggang Zhang, Ninghao Liu, Jinsong Su, Bo An, Yi Chang, and Xiao Huang. Graph-based agent memory: Taxonomy, techniques, and applications. arXiv:2602.05665, 2026. URL https://arxiv.org/pdf/2602.05665.

Boci Peng, Yun Zhu, Yongchao Liu, Xiaohe Bo, Haizhou Shi, Chuntao Hong, Yan Zhang, and Siliang Tang. Graph retrieval-augmented generation: A survey. In *arXiv (Cornell University)*. Cornell University, 2024. doi: 10.48550/arxiv.2408.08921. URL https://arxiv.org/pdf/2408.08921.

Neeraj Yadav. Luring the lurkers: Temporal validity in retrieval memory. arXiv:2412.04049, 2024. URL http://arxiv.org/abs/2412.04049v1.

Ryuichi Sumida, Koji Inoue, and Tatsuya Kawahara. Enhancing long-term rag chatbots with psychological models of memory importance and forgetting. In *Dialogue & Discourse 16(2) (2025) 74–110*, 2024. doi: 10.5210/dad.2025.203. URL http://arxiv.org/abs/2409.12524v3.

Richard T. Snodgrass. *Developing Time-Oriented Database Applications in SQL*. Morgan Kaufmann, 1995. ISBN 978-1-55860-436-0. URL https://www2.cs.arizona.edu/~rts/publications.html.

Datomic Development Team. Datomic time concepts: Valid time, transaction time, and as-of queries. Datomic Technical Reference, https://docs.datomic.com/time.html, 2024. Reference for the half-open validity interval formulation and as-of-time query semantics that Holistic Context inherits. The time-concepts reference gives the half-open reconstruction query that Holistic Context's Memory Forgetting Engine implements.

ISO/IEC 9075:2011 (SQL:2011). SQL – part 2: Sql/foundation – temporal tables (features F401–F452). International Organization for Standardization, Geneva, Switzerland, 2011. URL https://www.iso.org/standard/53682.html. Reference for the half-open validity-interval + as-of-time formulation pattern used in temporal tables; cited as the SQL standard ancestor of HOLISTIC CONTEXT's as-of-time query.

Donald O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*. Wiley, New York, 1949. ISBN 978-0-8058-1543-6.

Allan M. Collins and Elizabeth F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407–428, 1975. doi: 10.1037/0033-295X.82.6.407.

John R. Anderson. *The Architecture of Cognition*. Harvard University Press, Cambridge, MA, 1983. ISBN 978-0-674-04339-3.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 21497–21515. PMLR, 2025. URL https://proceedings.mlr.press/v267/gutierrez25a.html.

Emanuele Rossi, Ben Chamberlain, Federico Frasca, Davide Eynard, and Michael M. Bronstein. Temporal graph networks for deep learning on dynamic graphs. In *International Conference on Machine Learning (ICML) GRL+ Workshop*, 2020. URL https://tgn.cs.ucl.ac.uk/.

Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. DyRep: Learning representations over dynamic graphs. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=HyePrhR5KX.

Guilin Zhang, Wei Jiang, Xiejiashan Wang, Aisha Behr, Kai Zhao, Jeffrey Friedman, Xu Chu, and Amine Anoun. Adaptive memory admission control for llm agents. arXiv:2603.04549, 2026. URL https://arxiv.org/pdf/2603.04549.

OpenContext Contributors. Opencontext: An open-source ai coworker with holistic context. https://github.com/melandlabs/opencontext, 2026. Apache-2.0 license.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, and Faisal Lahat. Evaluating very long-term conversational memory of llm agents. arXiv:2402.17753, 2024. URL http://arxiv.org/abs/2402.17753v1.

Xiaoran Liu, Ruixiao Li, Mianqiu Huang, Zhigeng Liu, Yuerong Song, Qipeng Guo, Siyang He, Qiqi Wang, Linlin Li, Qun Liu, Ziwei He, Yaqian Zhou, Xuanjing Huang, and Xipeng Qiu. Thus spake long-context large language model. arXiv:2502.17129, 2025. URL http://arxiv.org/abs/2502.17129v2.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024a. doi: 10.1007/s11704-024-40231-1. URL https://link.springer.com/content/pdf/10.1007/s11704-024-40231-1.pdf.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. Agentic large language models, a survey. In *JAIR volume 82, article 26, 2025, pp. 1823–1918*, 2025. doi: 10.1613/jair.1.18126. URL http://arxiv.org/abs/2503.23037v3.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. A survey on the memory mechanism of large language model based agents. In *arXiv (Cornell University)*. Cornell University, 2024. doi: 10.48550/arxiv.2404.13501. URL https://arxiv.org/pdf/2404.13501.

Liang Wu, Jie Liu, Xingyu Zhang, Qianliang Liu, Yang Liu, Kehai Chen, Tianning Zuo, et al. From human memory to ai memory: A survey on memory mechanisms in the era of llms. In *arXiv (Cornell University)*. Cornell University, 2025. doi: 10.48550/arxiv.2504.15965. URL https://arxiv.org/pdf/2504.15965.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0 v3 platform: Append-only extraction, entity linking, and hybrid retrieval for production memory agents. Mem0 Technical Report V3, 2026. URL https://mem0.ai/research. April 2026 algorithm update; four changes versus the original Mem0 paper: (i) single-pass ADD-only extraction with no UPDATE or DELETE operations, (ii) agent-generated facts stored with the same weight as user-supplied facts, (iii)

entity extraction and linking across memories, (iv) multi-signal retrieval fusing semantic similarity, BM25 keyword, and entity matching. Reported numbers appear in [Sangshetti, 2026].

Himanshu Sangshetti. AI memory benchmarks 2026: LoCoMo, LongMemEval and BEAM. Mem0 Blog, July 2026. URL https://mem0.ai/blog/ai-memory-benchmarks-in-2026. Published 2026-07-08. Source of Mem0 V3 numbers on LoCoMo (92.5%), LongMemEval (94.4%), BEAM-1M (64.1%), and BEAM-10M (48.6%); access path and tokens-per-retrieval figures also reported.

Suozhao Ji, Baodong Wu, Zehao Wang, Lei Xia, Qingping Li, Ruisong Wang, Wenbo Ding, Zhenhua Zhu, Boxun Li, Guohao Dai, and Yu Wang. Infini memory: Maintainable topic documents for long-term llm agent memory. arXiv:2606.10677, 2026. URL https://arxiv.org/pdf/2606.10677.

Zhaofen Wu, Hanrong Zhang, Fulin Lin, Wujiang Xu, Xinran Xu, Yankai Chen, Henry Peng Zou, Shaowen Chen, Weizhi Zhang, Xue Liu, Philip S. Yu, and Hongwei Wang. GAM: Hierarchical graph-based agentic memory for LLM agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34647–34664, San Diego, California, United States, 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1600. URL https://aclanthology.org/2026.acl-long.1600/.

Pengyu Gao, Jinming Zhao, Xinyue Chen, and Yilin Long. An efficient context-dependent memory framework for LLM-centric agents. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track)*, pages 1055–1069, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-industry.80. URL https://aclanthology.org/2025.naacl-industry.80/.

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seung-won Hwang, Dongha Lee, and Jinyoung Yeo. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8661, Albuquerque, New Mexico, 2025. Association for Computational Linguistics.

doi: 10.18653/v1/2025.naacl-long.435. URL https://aclanthology.org/2025.naacl-long.435/.

Miao Su, Yucan Guo, Zhongni Hou, Long Bai, Zixuan Li, Yufei Zhang, Guojun Yin, Wei Lin, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. Beyond dialogue time: Temporal semantic memory for personalized LLM agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 29935–29951, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1496. URL https://aclanthology.org/2026.findings-acl.1496/.

Pratyay Banerjee, Masud Moshtaghi, Shivashankar Subramanian, Amita Misra, and Ankit Chadha. APEX-MEM: Agentic semi-structured memory with temporal reasoning for long-term conversational AI. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16470–16489, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.749. URL https://aclanthology.org/2026.acl-long.749/.

Jiale Han, Austin Cheung, Yubai Wei, Zheng Yu, Xusheng Wang, Bing Zhu, and Yi Yang. Rag meets temporal graphs: Time-sensitive modeling and retrieval for evolving knowledge. arXiv:2510.13590, 2025. URL http://arxiv.org/abs/2510.13590v1.

Carlos E. Alchourrón, Peter Gärdenfors, and David Makinson. On the logic of theory change: Partial meet contraction and revision functions. *Journal of Symbolic Logic*, 50(2):510–530, 1985. doi: 10.2307/2274239. URL https://www.jstor.org/stable/2274239.

Hirofumi Katsuno and Alberto O. Mendelzon. Propositional knowledge base revision and minimal change. In *Artificial Intelligence*, volume 52, pages 263–294. Elsevier, 1991. doi: 10.1016/0004-3702(91)90013-V. URL https://www.sciencedirect.com/science/article/pii/000437029190013V.

Haoran Sun, Zekun Zhang, and Shaoning Zeng. Preference-aware memory update for long-term LLM agents. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 783–793, San Diego, California, United States, 2026b. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.38. URL https://aclanthology.org/2026.findings-acl.38/.

Zidi Xiong, Yuping Lin, Wenya Xie, Pengfei He, Zirui Liu, Jiliang Tang, Himabindu Lakkaraju, and Zhen Xiang. How memory management impacts LLM agents: An empirical study of experience-following behavior. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 623–645, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.27. URL https://aclanthology.org/2026.acl-long.27/.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. Knowledge conflicts for llms: A survey. arXiv:2403.08319, 2024. URL http://arxiv.org/abs/2403.08319v2.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. arXiv:2304.03442, 2023. URL http://arxiv.org/abs/2304.03442v2.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. arXiv:2303.11366, 2023. URL https://arxiv.org/pdf/2303.11366.

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

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *Transactions on Machine Learning Research*, 2024b. URL https://openreview.net/forum?id=ehfRiF0R3a.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. ExpeL: LLM agents are experiential learners. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19632–19642. AAAI Press, 2024. doi: 10.1609/aaai.v38i17.29936. URL https://ojs.aaai.org/index.php/AAAI/article/view/29936.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Sandkühler, Daniel Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*, 2020.