HeLa-Mem: Hebbian Learning and Associative Memory for LLM Agents

Jinchang Zhu[1,a*], **Jindong Li**[1*], **Cheng Zhang**[2*], **Jiahong Liu**[3], **Menglin Yang**[1,b†]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Jilin University [3]The Chinese University of Hong Kong
[a]jzhu997@connect.hkust-gz.edu.cn [b]menglinyang@hkust-gz.edu.cn

Abstract

Long-term memory is a critical challenge for Large Language Model agents, as fixed context windows cannot preserve coherence across extended interactions. Existing memory systems encode conversation history as embedding vectors and retrieve information through semantic similarity. This paradigm fails to capture the *associative structure* of human memory, wherein related experiences progressively strengthen interconnections through repeated co-activation. Inspired by cognitive neuroscience, we identify three mechanisms central to biological memory: *association*, *consolidation*, and *spreading activation*, which remain largely absent in current research. To bridge this gap, we propose **HeLa-Mem**, a bio-inspired memory architecture that models memory as a dynamic graph with Hebbian learning dynamics. HeLa-Mem employs a dual-level organization: (1) an *episodic memory graph* that evolves through co-activation patterns, and (2) a *semantic memory store* populated via Hebbian Distillation, wherein a Reflective Agent identifies densely connected memory hubs and distills them into structured, reusable semantic knowledge. This dual-path design leverages both semantic similarity and learned associations, mirroring the episodic-semantic distinction in human cognition. Experiments on LoCoMo demonstrate superior performance across four question categories while using significantly fewer context tokens. Code is available on GitHub.

1 Introduction

Large language models have demonstrated remarkable capabilities in language understanding and generation, enabling increasingly sophisticated interactive agents (Yang et al., 2024; Liu et al., 2026; Li et al., 2026). However, sustaining coherent behavior over long time horizons remains a fundamental challenge (Zhang et al., 2025). Due to their

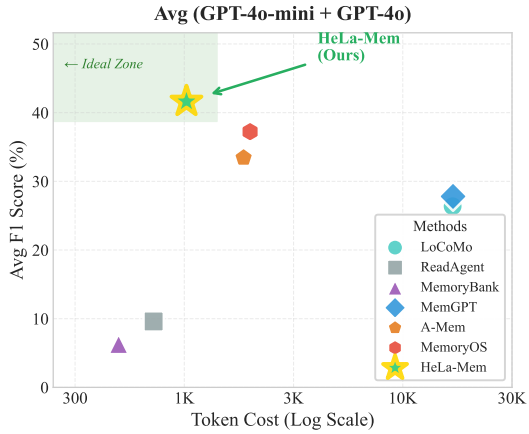

Figure 1: Performance vs. token efficiency on LoCoMo, averaged across GPT-4o-mini and GPT-4o. HeLa-Mem achieves strong performance with fewer tokens, landing in the upper-left ideal region.

reliance on fixed-length context windows, LLMs struggle to maintain consistent representations of past interactions as dialogues extend or span multiple sessions. This limitation often leads to fragmented memory (Wu et al., 2025; Hu et al., 2025; Liang et al., 2025; Liu et al., 2025), resulting in factual inconsistencies, diminished personalization, and unstable agent behavior. Addressing long-term memory coherence is therefore essential for LLM agents operating in settings that require persistent user adaptation, multi-session knowledge retention, or stable persona maintenance.

Current memory mechanisms for LLM agents can be broadly categorized into three methodological paradigms. *Knowledge-organization methods*, such as A-Mem (Xu et al., 2025), structure memory into interconnected semantic networks to enable adaptive management. *Retrieval mechanism-oriented approaches*, exemplified by MemoryBank (Zhong et al., 2024), integrate semantic retrieval with memory forgetting curves for long-term updating. *Architecture-driven methods*, including MemGPT (Packer et al., 2023), employ hierarchi-

*Equal contribution.
†Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13757–13769

cal memory structures with explicit read and write operations to dynamically manage limited context windows. Although effective, these approaches are typically developed in isolation, each prioritizing a single dimension—memory structure, retrieval strategy, or update mechanism—while largely overlooking their mutual interaction and joint contribution to long-term coherence.

More fundamentally, this component-wise optimization often overlooks a critical aspect of long-term coherence: the *dynamic evolution* of memory structure. Human memory is not a static database where items are stored and retrieved in isolation; rather, it is a dynamic system where connections are continuously reorganized by experience. For example, a topic discussed today might trigger a memory from a month ago, not because they share surface-level keywords, but because they are part of the same evolving narrative arc. Current systems, by treating storage and retrieval as separate static processes, fail to capture this evolving connectivity, leading to agents that "remember" facts but lack the "continuity" of a developing relationship.

To better understand how long-term coherence can be maintained, we draw on a fundamental principle of biological memory: *Hebbian learning*. In biological systems, experiences that are repeatedly co-activated gradually develop stronger associations, a phenomenon often summarized as "neurons that fire together wire together" (illustrated in Figure 2). This associative organization allows related memories to be efficiently reactivated through spreading activation and supports the gradual consolidation of episodic experiences into more stable semantic knowledge. Together, association, consolidation, and spreading activation form a tightly coupled memory process that enables biological systems to maintain coherent representations over extended time scales—capabilities that remain largely absent from current artificial memory designs.

Building on this perspective, we propose HeLa-Mem (**He**bbian **L**earning **a**ssociative **Mem**ory), a unified memory architecture for LLM agents. HeLa-Mem represents conversational history as a dynamic graph with Hebbian learning dynamics and operates through coordinated mechanisms of online association, reflective consolidation, and dual-path retrieval, establishing a unified memory management framework that captures both fine-grained details and high-level patterns. HeLa-Mem demonstrates superior performance on LoCoMo while using significantly fewer context tokens (Figure 1). The primary contributions of our work are:

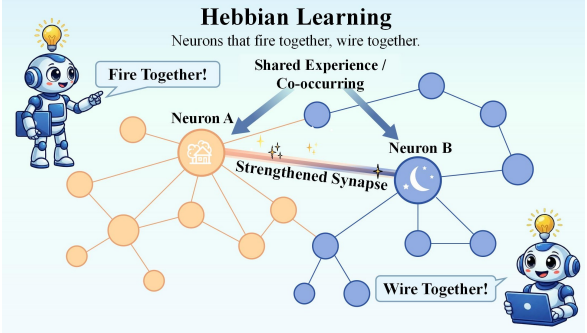

Figure 2: Conceptual illustration of Hebbian learning in associative memory. Two memory nodes (Neuron A and B) representing distinct experiences (e.g., a daytime event and a nighttime event) develop strengthened synaptic connections when co-activated through shared context. This "neurons that fire together, wire together" principle forms the theoretical foundation of HeLa-Mem's dynamic memory graph.

- We propose **HeLa-Mem**, a bio-inspired framework that utilizes an *Online Encoding & Association* mechanism to model conversation history as a dynamic Hebbian graph, where co-activated memories strengthen connections to capture latent context.
- We introduce a *Reflective Consolidation* framework using **Hebbian Distillation**, which identifies hub clusters and transforms them into structured semantic knowledge, preventing graph explosion while retaining key information.
- We implement a *Dual-Path Retrieval* strategy that leverages spreading activation to traverse Hebbian edges, achieving the best average rank (1.25) across all question categories.
- Comprehensive experiments on the LoCoMo benchmark validate HeLa-Mem's effectiveness, achieving superior performance across four categories while using significantly fewer context tokens.

2 Related Work

2.1 Memory for LLM Agents

Existing Large Language Models face fundamental challenges in handling complex scenarios requiring long-term coherence (Hu et al., 2025). Advancements in memory systems addressing this problem can be broadly grouped into three categories.

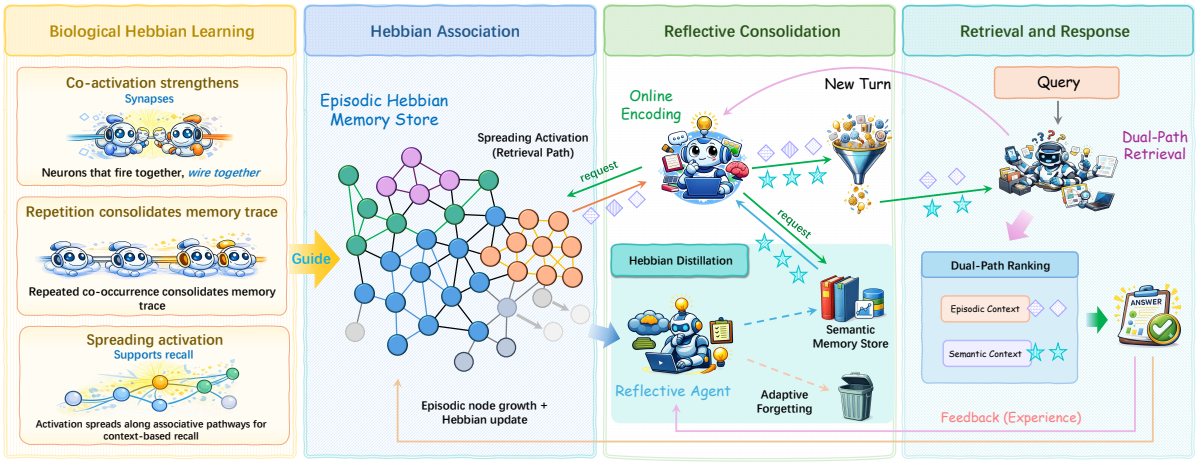

Figure 3: **The architectural overview of HeLa-Mem.** The framework consists of three modules: (1) **Hebbian Association** for dynamic graph construction (Section 3.2); (2) **Reflective Consolidation** for semantic knowledge distillation (Section 3.3); and (3) **Retrieval and Response** using a Dual-Path strategy (Section 3.4).

Knowledge-organization methods focus on capturing and structuring intermediate reasoning states. Think-in-Memory (Liu et al., 2023) stores evolving chains-of-thought, enabling consistency through continual updates. A-Mem (Xu et al., 2025) organizes knowledge into an interconnected note network that spans sessions.

Retrieval mechanism-oriented approaches, pioneered by RAG (Lewis et al., 2020), enrich the model with external memory libraries. MemoryBank (Zhong et al., 2024) logs conversations, events, and user traits in a vector database and refreshes them using a forgetting-curve schedule. Generative Agents (Park et al., 2023) keep memories in natural language and add a reflection loop for relevance filtering. EmotionalRAG (Huang et al., 2024) retrieves memory entries by combining semantic similarity with the agent's emotional state.

Architecture-driven designs alter the core control flow to manage context explicitly. MemGPT (Packer et al., 2023) adopts an OS-like hierarchy with dedicated read/write calls. SCM (Liang et al., 2023) introduces dual buffers and a memory controller that gates selective recall. Mem0 (Chhikara et al., 2025) dynamically extracts and consolidates salient information for scalable long-term memory. MemoryOS (Kang et al., 2025) introduces a three-tier hierarchical storage with short-term, mid-term, and long-term memory units, employing segment-page organization for dynamic updating.

2.2 Hebbian Learning in Neural Networks

Hebbian learning, summarized as "neurons that fire together wire together" (Hebb, 2005), is a foundational principle in neuroscience describing how synaptic connections strengthen through correlated activity (see Figure 2). Formally, given activation states x_i and x_j of two neurons, the connection weight w_{ij} is updated as $\Delta w_{ij} = \eta \cdot x_i \cdot x_j$, where η is the learning rate. This principle has been applied in Hopfield networks (Hopfield, 1982), which demonstrate how recurrent neural networks with symmetric connections can function as associative memories, storing and retrieving patterns through energy minimization. More recently, Ramsauer et al. (2020) show that modern Hopfield networks with continuous states are mathematically equivalent to the attention mechanism in Transformers, revealing deep connections between biological memory principles and contemporary deep learning architectures. In the context of LLM agents, Hebbian dynamics offer a principled approach to capture latent associations between memories that may not be apparent from semantic similarity alone.

3 HeLa-Mem Architecture

Inspired by the synaptic plasticity of the human brain, HeLa-Mem models conversation history as a dynamic associative graph rather than a static log. Our design is guided by three neuroscience intuitions: (1) *Association over Isolation*—memories that co-occur should wire together, forming latent pathways beyond simple semantic similarity; (2) *Active Consolidation*—frequently accessed memory clusters should solidify into stable knowledge, similar to sleep-based consolidation; and (3) *Spreading Retrieval*—recalling one memory should naturally trigger related concepts through

established synaptic routes.

Based on these principles (Figure 3), HeLa-Mem operates through a continuous cognitive lifecycle:

- **Online Encoding & Association**: Conversation turns are encoded into the Episodic Memory Graph, where a Hebbian Learning mechanism dynamically strengthens connections between co-activated memories.
- **Reflective Memory Agent**: Upon reaching associative thresholds, this agent identifies hub nodes and applies *Hebbian Distillation* to consolidate them into stable semantic knowledge, preventing noise accumulation.
- **Dual-Path Retrieval**: During query time, queries activate both specific episodic details and broader semantic knowledge through spreading activation.

3.1 Memory Storage

3.1.1 Episodic Memory Graph

Conversation turns are stored as nodes in a weighted graph. Each node contains the original text, a dense embedding, the timestamp, extracted keywords, and the speaker role. Edges between nodes represent associative connections, with weights indicating the strength of association. Initially, consecutive turns are connected with small weights; these weights evolve through Hebbian learning.

3.1.2 Semantic Memory Store

The semantic level stores distilled knowledge extracted from episodic memories: specifically, we apply *Hebbian Distillation* on hub-centered clusters in the episodic graph to produce *distilled semantic records* with traceable evidence links to their source turns.

- **User Model**: Stable user characteristics (e.g., "enjoys outdoor activities") with confidence scores and supporting evidence.
- **Factual Memory**: Extracted facts with absolute timestamps, such as event dates and relationships.
- **Agent Knowledge**: The agent's established persona, preferences, and behavioral patterns.

This serves as long-term memory that persists beyond conversation windows.

3.2 Online Encoding & Association

Synaptic efficacy in the brain is not fixed; it is plastic, evolving based on activity patterns. We emulate this dynamic through Hebbian learning to capture latent associations that semantic embeddings alone might miss.

Following the neuroscience principle that "neurons that fire together wire together," edge weights strengthen when memories are co-activated during retrieval:

$$w_{ij}^{(t+1)} = \underbrace{(1-\lambda)\cdot w_{ij}^{(t)}}_{\text{synaptic decay}} + \underbrace{\eta \cdot \mathbb{I}(v_i, v_j \in \mathcal{K}_t)}_{\text{active reinforcement}}, \quad (1)$$

where λ is the decay rate, η is the learning rate, and $\mathbb{I}(\cdot)$ is an indicator that the pair (v_i, v_j) is co-activated in the current retrieval set $\mathcal{K}_t$. This dynamic allows frequently correlated memories to strengthen while unused connections fade over time.

3.3 Reflective Memory Agent

While associative learning happens during active retrieval, long-term memory maintenance relies on active consolidation. To prevent memory overload and crystallize important information, we introduce a Reflective Agent that mimics the brain's sleep-based consolidation process through *Hebbian Distillation*.

The Reflective Agent monitors the graph's structural evolution to manage the memory lifecycle (see Figure 4), analogous to sleep-based memory consolidation in the brain.

Hub Detection. Nodes that have accumulated high total edge weight through Hebbian learning are identified as hubs. Specifically, a node v_i is flagged for consolidation if its *associative strength* exceeds a threshold δ_{hub}:

$$D(v_i) = \sum_{j \in \mathcal{N}(i)} w_{ij} > \delta_{hub}. \quad (2)$$

Upon detecting that a node's accumulative weight exceeds the threshold δ_{hub}, the agent triggers Hebbian Distillation. To capture the full context, the agent retrieves the hub node along with its strongly connected neighbors. The LLM synthesizes this cluster of related memories to identify common themes and causal relationships, abstracting them into declarative semantic entries. These distilled records are stored in the Semantic Memory Store, effectively compressing repetitive episodic details into stable, generalizable knowledge.

Adaptive Forgetting. This process is triggered when a node's status falls below critical retention thresholds. A memory is flagged for removal only if it simultaneously satisfies three criteria: (1) its total edge weight is below δ_{prune} (indicating structural irrelevance), (2) its inactive duration exceeds δ_{age} (indicating temporal dormancy), and (3) it has zero recent access. This strict compound criterion ensures that the system selectively removes noise while preserving strong, albeit older, associations.

3.4 Dual-Path Retrieval

Human memory retrieval is rarely a single-step lookup; it is a spreading activation process where one thought triggers another. HeLa-Mem adopts a dual-path retrieval strategy to emulate this interaction between direct recall and associative spreading.

Given a query, retrieval proceeds in two stages.

Base Activation. Each episodic node receives an initial score combining embedding similarity, temporal decay, and keyword overlap:

$$S_{base}(v_i) = (\text{sim}(\mathbf{q}, \mathbf{e}_i) + \alpha \cdot \text{keyword\_match}) \cdot \gamma(v_i), \tag{3}$$

where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity between the query embedding $\mathbf{q}$ and node embedding $\mathbf{e}_i$, $\gamma(v_i) = \exp(-\Delta t/\tau)$ is the temporal decay factor with time constant τ, and α controls the bonus for keyword matches.

Spreading Activation. High-scoring nodes propagate activation through Hebbian edges:

$$S(v_j) = S_{base}(v_j) + \beta \sum_{i \in \mathcal{N}(j)} S_{base}(v_i) \cdot w_{ij}, \tag{4}$$

where $\mathcal{N}(j)$ denotes the neighbors of node v_j in the memory graph, w_{ij} is the Hebbian edge weight, and β controls the spreading activation strength. This enables retrieval of memories that are semantically distant from the query but strongly associated with initially activated content—particularly beneficial for multi-hop reasoning.

Dual-Path Ranking. The final retrieval set is constructed by combining two ranked lists:

$$\mathcal{R}_{final} = \underbrace{\text{Top-}k(S_{base})}_{\text{base path}} \cup \underbrace{\text{Top-}m(S \mid v \notin \text{Top-}k)}_{\text{flip path}}, \tag{5}$$

where the base path selects the top-k nodes by S_{base}, and the flip path promotes up to m additional nodes that rank highest by spreading-augmented score S but were not already selected.

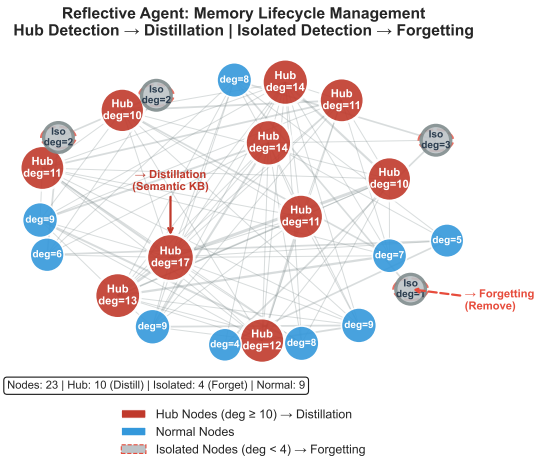

Figure 4: Hebbian memory graph showing the Reflective Agent's dual role. **Hub nodes** (red, high degree) are candidates for Hebbian Distillation, consolidating related memories into structured semantic knowledge. **Isolated nodes** (gray, low degree with dashed circles) are candidates for Adaptive Forgetting, maintaining memory efficiency.

This dual-path approach ensures retrieval of both semantically relevant memories and associatively linked memories that spreading activation surfaces. Semantic memory entries are also retrieved and merged to form the final context.

3.5 Response Generation

The LLM generates responses using the integrated context (episodic memories and semantic knowledge) along with a system prompt that establishes the conversational role. The detailed prompt structure is provided in Appendix A.

4 Experiments

4.1 Experimental Settings

Dataset. We conduct experiments on the LoCoMo benchmark (Maharana et al., 2024), specifically designed for assessing long-term conversational memory capabilities. It consists of ultra-long dialogues averaging 300 turns and about 9K tokens per conversation. Questions span multiple categories to systematically evaluate memory abilities.

Evaluation Metrics. Following prior work on long-term conversational memory (Maharana et al., 2024; Xu et al., 2025), we employ standard F1 and BLEU-1 scores to evaluate performance.

Compared Methods. We compare HeLa-Mem with representative memory methods including LoCoMo (Native), ReadAgent (Lee et al., 2024), MemoryBank (Zhong et al., 2024), MemGPT

(Packer et al., 2023), A-Mem (Xu et al., 2025), Mem0 (Chhikara et al., 2025), LightMem (Fang et al., 2025), and MemoryOS (Kang et al., 2025). Baseline results for GPT-4o-mini, GPT-4o, and Qwen2.5-3b are reported from Xu et al. (2025); results for Qwen2.5-14b are reported from Yan et al. (2025). MemoryOS results marked with † are reproduced by us.

Implementation Details. We evaluate HeLa-Mem across four backbone LLMs: GPT-4o-mini, GPT-4o (Achiam et al., 2023), Qwen2.5-14b, and Qwen2.5-3b (Yang et al., 2024). For HeLa-Mem, the time decay constant is set to $\tau = 60$ days. Episodic retrieval uses $k = 10$, and semantic retrieval uses $k = 5$. The Hebbian learning rate is $\eta = 0.02$, edge decay rate $\lambda = 0.995$, spreading activation strength $\beta = 0.1$, and spreading threshold $\theta = 0.6$.

4.2 Main Results

Table 1 presents the detailed performance breakdown across four different LLM backbones: GPT-4o-mini, GPT-4o, Qwen2.5-14b, and Qwen2.5-3b. Results show that HeLa-Mem consistently outperforms baselines across varying model sizes and capabilities.

Detailed Analysis (GPT-4o-mini). Focusing on GPT-4o-mini as a representative case, HeLa-Mem demonstrates significant advantages. In Multi-hop reasoning, it achieves 40.14%, outperforming MemoryOS (38.39%) and A-Mem (27.02%). This validates the Hebbian graph's ability to bridge disparate information pieces through learned associations.

For Temporal tasks, HeLa-Mem scores 47.29%, surpassing MemoryOS (41.58%). The preservation of absolute timestamps during distillation allows accurate grounding of relative time expressions. In Open Domain questions, it reaches 29.70%, providing useful semantic context even for topics outside the main conversation flow. HeLa-Mem leads in Single-hop tasks (51.89%), demonstrating that the hierarchical retrieval approach remains effective for straightforward factual queries.

Token Efficiency. Notably, HeLa-Mem achieves these results using only ∼1,010 tokens on average. This efficiency stems from the selective nature of Hebbian retrieval, which surfaces only the most strongly associated memories without the computational overhead of processing full context windows.

Robustness Across Backbones. To validate stability, Table 2 summarizes the averaged performance across all backbones. HeLa-Mem achieves the best **Average Rank of 1.25**, significantly surpassing MemoryOS (2.25). This confirms that the theoretical advantages of Hebbian dynamics translate into robust empirical gains regardless of the underlying LLM's scale.

4.3 Ablation Study

To understand the contribution of each component in HeLa-Mem, we conduct ablation experiments by removing key modules. Table 3 presents the results on GPT-4o-mini.

Effect of Reflective Agent. Removing the Reflective Memory Agent causes the largest performance drop (34.74% → 29.87%), with Multi-hop reasoning suffering most severely (36.04% → 30.17%). This confirms that the meta-cognitive component is essential for identifying high-degree hub nodes and triggering Hebbian Distillation, which consolidates related episodic memories into structured semantic knowledge.

Effect of Spreading Activation. Disabling spreading activation leads to a notable performance decline (34.74% → 32.19%), particularly affecting Multi-hop reasoning (36.04% → 33.88%). This validates our dual-path design: without spreading activation, the system degrades to a single semantic path, failing to retrieve memories that are semantically distant from the query but strongly associated through Hebbian connections, which is crucial for multi-hop reasoning that requires bridging disparate pieces of information. This underscores the value of leveraging historically learned pathways rather than relying solely on static semantic similarity.

Effect of Adaptive Forgetting. Interestingly, removing the forgetting mechanism shows minimal impact on the current LoCoMo benchmark. We attribute this to the limited conversation length (∼300 turns), which does not yet saturate the memory capacity. However, forgetting is critical for scalability in reliable agent deployment. Without it, the memory store would grow unboundedly, inevitably increasing retrieval costs and introducing noise from obsolete information. This mechanism ensures that the system's performance remains stable regardless of conversation duration, acting as a garbage collection process for irrelevant associations.

Table 1: Experimental results on LoCoMo dataset of QA tasks across four categories (Multi Hop, Temporal, Open Domain, and Single Hop) using different methods. Results are reported in F1 and BLEU-1 (%) scores. Best performance per model is marked in bold. Missing baselines are marked with "-" (Token Length (↓): lower values are better; †: Reproduced results).

Model	Method	Multi Hop		Temporal		Open Domain		Single Hop		Token Length (↓)
		F1	BLEU	F1	BLEU	F1	BLEU	F1	BLEU	
GPT-4o-mini	LoCoMo	25.02	19.75	18.41	14.77	12.04	11.16	40.36	29.05	16,910
	ReadAgent	9.15	6.48	12.60	8.87	5.31	5.12	9.67	7.66	643
	MemoryBank	5.00	4.77	9.68	6.99	5.56	5.94	6.61	5.16	432
	MemGPT	26.65	17.72	25.52	19.44	9.15	7.44	41.04	34.34	16,977
	A-Mem	27.02	20.09	45.85	36.67	12.14	12.00	44.65	37.06	2,520
	MemoryOS†	38.39	29.52	41.58	35.99	23.75	17.17	45.86	40.70	2,000
	HeLa-Mem	**40.14**	**31.26**	**47.29**	**41.28**	**29.70**	**23.45**	**51.89**	**46.25**	**1,010**
GPT-4o	LoCoMo	28.00	18.47	9.09	5.78	16.47	14.80	**61.56**	**54.19**	16,910
	ReadAgent	14.61	9.95	4.16	3.19	8.84	8.37	12.46	10.29	805
	MemoryBank	6.49	4.69	2.47	2.43	6.43	5.30	8.28	7.10	569
	MemGPT	30.36	22.83	17.29	13.18	12.24	11.87	60.16	53.35	16,987
	A-Mem	32.86	23.76	39.41	31.23	17.10	15.84	48.43	42.97	1,216
	MemoryOS†	**40.23**	**31.89**	43.57	33.55	20.58	15.85	43.85	39.03	2,000
	HeLa-Mem	39.12	29.82	**50.79**	**44.54**	**24.38**	**19.24**	49.69	44.08	**1,036**
Qwen2.5-14b	A-Mem	22.09	15.28	27.19	22.05	13.49	10.74	33.75	30.04	1,300
	MEM0	31.73	24.82	28.96	26.24	15.03	11.28	42.58	35.15	-
	MemoryOS	**38.19**	**29.26**	32.24	27.86	20.27	15.94	46.33	41.62	-
	LightMem	25.45	19.61	32.03	27.70	15.81	11.81	34.92	31.22	-
	HeLa-Mem	36.59	27.02	**36.08**	**29.91**	**24.22**	**20.23**	**49.95**	**45.15**	**944**
Qwen2.5-3b	LoCoMo	4.61	4.29	3.11	2.71	4.55	5.97	7.03	5.69	16,910
	ReadAgent	2.47	1.78	3.01	3.01	5.57	5.22	3.25	2.51	776
	MemoryBank	3.60	3.39	1.72	1.97	6.63	6.58	4.11	3.32	298
	MemGPT	5.07	4.31	2.94	2.95	7.04	7.10	7.26	5.52	16,961
	A-Mem	18.23	11.94	24.32	19.74	**16.48**	**14.31**	23.63	19.23	1,300
	MemoryOS†	19.20	14.84	20.85	16.05	13.57	10.86	25.65	18.78	2,000
	HeLa-Mem	**20.12**	**14.59**	**24.79**	**21.35**	12.24	10.24	**29.51**	**25.91**	**1,072**

Table 2: Averaged results across three backbone LLMs (GPT-4o-mini, GPT-4o, Qwen2.5-3b). Avg Rank is computed across all categories lower is better (†Reproduced results).

Method	Multi-hop		Temporal		Open		Single		Avg Rank
	F1	BL	F1	BL	F1	BL	F1	BL	
LoCoMo	19.21	14.17	10.20	7.75	11.02	10.64	36.32	29.64	5.00
ReadAgent	8.74	6.07	6.59	5.02	6.57	6.24	8.46	6.82	6.38
MemoryBank	5.03	4.28	4.62	3.80	6.21	5.94	6.33	5.19	6.88
MemGPT	20.69	14.95	15.25	11.86	9.48	8.80	36.15	31.07	4.50
A-Mem	26.04	18.60	36.53	29.21	15.24	14.05	38.90	33.09	3.00
MemoryOS†	32.61	25.42	35.33	28.53	19.30	14.63	38.45	32.84	2.25
HeLa-Mem	**33.13**	**25.22**	**40.96**	**35.72**	**22.11**	**17.64**	**43.70**	**38.75**	**1.25**

Table 3: Ablation study on LoCoMo benchmark. Results show F1 / BLEU-1 scores (%).

Variant	Multi-hop		Temporal		Open		Single		Avg F1
	F1	BL	F1	BL	F1	BL	F1	BL	
HeLa-Mem (Full)	**36.04**	**26.56**	**46.23**	**40.48**	29.50	23.55	**45.04**	**39.80**	**34.74**
w/o Forgetting	36.71	27.95	46.50	40.91	**30.58**	**24.45**	45.24	40.01	34.28
w/o Spreading Activation	33.88	25.57	44.36	39.62	27.76	22.28	43.34	38.33	32.19
w/o Reflective Agent	30.17	22.38	42.19	36.92	24.51	19.83	40.46	34.07	29.87

4.4 Reflective Agent: Memory Lifecycle Management

Figure 4 illustrates the structure of the Hebbian memory graph after encoding a multi-session conversation. The graph comprises 23 episodic memory nodes, where edge thickness reflects the association strength accumulated through co-activation during retrieval. Node appearance encodes the lifecycle status assigned by the Reflective Agent:

Hub Nodes (Red, Solid). Ten nodes exhibit degree ≥ 10, indicating dense connectivity within the graph. These nodes tend to occupy central positions, as they serve as anchors connecting multiple conversation threads. The Reflective Agent identifies such high-degree nodes and applies Hebbian Distillation to consolidate their associated episodic clusters into stable semantic entries. For instance, the node with degree 17 in Figure 4 links several temporally dispersed discussions, making it a natural candidate for knowledge extraction.

Isolated Nodes (Gray, Dashed). Four nodes exhibit degree < 4 and show no recent access activity. Their peripheral positions and weak integration suggest limited relevance to the ongoing narrative. The Adaptive Forgetting mechanism flags these nodes for removal, thereby controlling graph growth and reducing retrieval noise over extended conversations. The remaining nine nodes (blue) have moderate connectivity and are retained in the episodic store.

This visualization confirms that Hebbian learning enables automated lifecycle management with-

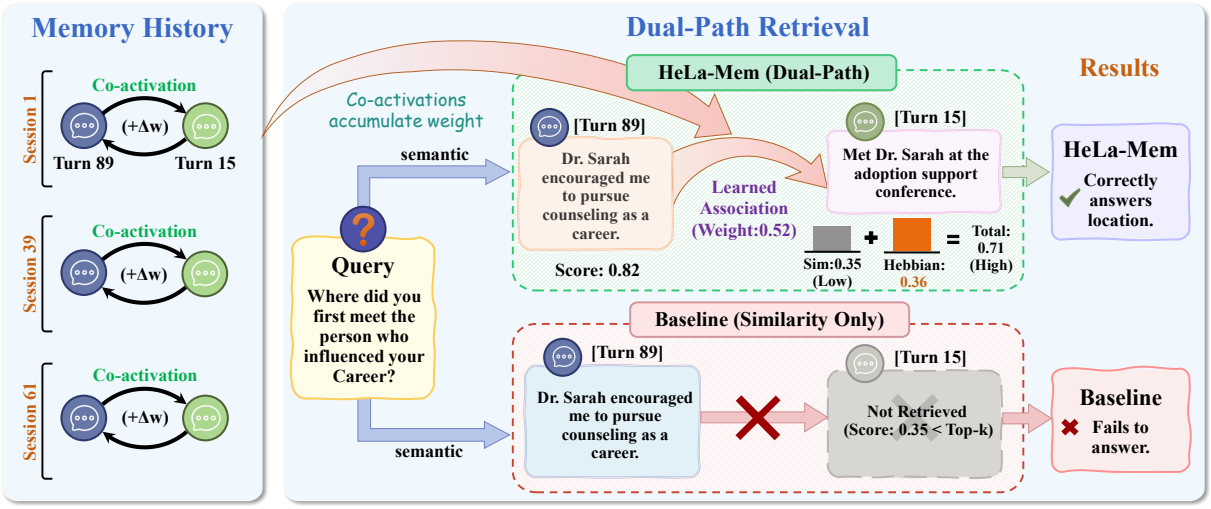

Figure 5: Dual-Path Retrieval for multi-hop reasoning. Given a query requiring both "career influence" and "meeting location," the semantic path retrieves Turn 89 (career context). The Hebbian path then propagates activation through learned associations (edge weight 0.52) to retrieve Turn 15 (location context), which semantic similarity alone would miss. The baseline without Spreading Activation cannot bridge these memories.

out manual annotation (see Appendix B for the edge-weight heatmap).

4.5 Case Study: Trace Analysis of Associative Recall

We analyze the retrieval process for a multi-hop query: "*Where did you first meet the person who influenced your career choice?*" (see Figure 5).

Historical Context. The entities "Dr. Sarah" (Person) and "Adoption Support Conference" (Location) appeared together in Session 1, and were subsequently co-activated in Sessions 39 and 61. Through Hebbian learning, these repeated co-occurrences accumulated a strong associative weight of $w_{89,15} \approx 0.52$ between the career advice memory (Turn 89) and the meeting event (Turn 15). This accumulated weight reflects the frequency and recency of co-activation across the conversation history.

Experimental Trace. The baseline method identifies Turn 89 ("Dr. Sarah encouraged...") as the top candidate due to high semantic similarity (0.82) but fails to retrieve Turn 15 ("Met Dr. Sarah at...") because its low similarity score of 0.35 falls below the retrieval threshold. This results in a "semantic trap" where the model knows the person but not the location.

In contrast, HeLa-Mem utilizes the Hebbian path. Spreading activation from the retrieved Turn 89 propagates through the learned edge (0.52) to Turn 15. The final retrieval score for Turn 15 is dynamically updated:

$$S_{total} = \underbrace{0.35}_{\text{Semantic}} + \underbrace{0.36}_{\text{Hebbian}} \approx \mathbf{0.71} \quad (6)$$

This score boost, derived strictly from historical association, promotes Turn 15 into the active context. By retrieving both the cue (Person) and the target (Location), HeLa-Mem correctly synthesizes the answer: "*At the adoption support conference.*" This demonstrates how Hebbian associations complement semantic retrieval for complex reasoning.

5 Conclusion

We introduce **HeLa-Mem**, a bio-inspired memory architecture that models conversation history as a dynamic graph driven by Hebbian learning principles. Unlike static context windows, HeLa-Mem mimics the brain's plasticity, where "neurons that fire together, wire together," enabling the spontaneous emergence of associative pathways for retrieval. Building on this foundation, our Reflective Agent distills transient episodes into structured semantic knowledge, while Adaptive Forgetting ensures long-term scalability. Experiments on the LoCoMo benchmark demonstrate that this synergy between associative retention and semantic consolidation yields superior performance across diverse question types and robustness across diverse LLM backbones. These findings suggest that incorporating neuro-symbolic dynamics offers a promising direction for evolving static LLMs into lifelong learning agents.

6 Limitations

While HeLa-Mem effectively models long-term memory consolidation, it faces a "cold start" challenge: Hebbian weights require sufficient interaction history to accumulate, meaning the benefits of associative retrieval are less pronounced in early conversation stages. Future work may explore initializing Hebbian edges using semantic similarity as a prior, allowing the graph structure to bootstrap before sufficient co-occurrences accumulate. Additionally, the quality of both Semantic Memory and Hub Detection relies on the capabilities of the underlying LLM; hallucinations or reasoning errors during the distillation process could propagate into the long-term storage, potentially affecting future retrieval accuracy.

7 Ethical Considerations

We use the publicly available LoCoMo benchmark and do not collect any private user data. The proposed memory architecture is intended to enhance the consistency of LLM agents. However, we acknowledge that long-term memory systems could potentially reinforce biases present in the underlying LLM if not carefully monitored. The distilled semantic memories should be treated with the same caution as standard LLM generations regarding accuracy and bias.

Acknowledgments

This work was partially supported by the Guangdong Provincial Natural Science Foundation General Program (Grant No. 2026A1515012118).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2025. LightMem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*.

Donald Olding Hebb. 2005. *The organization of behavior: A neuropsychological theory*. Psychology press.

John J Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2025. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Le Huang, Hengzhi Lan, Zijun Sun, Chuan Shi, and Ting Bai. 2024. Emotional rag: Enhancing role-playing agents through emotional retrieval. In *2024 IEEE International Conference on Knowledge Graph (ICKG)*, pages 120–127. IEEE.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory os of ai agent. *arXiv preprint arXiv:2506.06326*.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. *arXiv preprint arXiv:2402.09727*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Jindong Li, Yali Fu, Jiahong Liu, Linxiao Cao, Wei Ji, Menglin Yang, Irwin King, and Ming-Hsuan Yang. 2026. Discrete tokenization for multimodal llms: A comprehensive survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–24.

Jiafeng Liang, Hao Li, Chang Li, Jiaqi Zhou, Shixin Jiang, Zekun Wang, Changkai Ji, Zhihao Zhu, Runxuan Liu, Tao Ren, Jinlan Fu, See-Kiong Ng, Xia Liang, Ming Liu, and Bing Qin. 2025. Ai meets brain: Memory systems from cognitive neuroscience to autonomous agents. *arXiv preprint arXiv:2512.23343*.

Xinnian Liang, Bing Wang, Huijia Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Scm: Enhancing large language model with self-controlled memory framework. *arXiv preprint arXiv:2304.13343*.

Jiahong Liu, Zexuan Qiu, Zhongyang Li, Quanyu Dai, Wenhao Yu, Jieming Zhu, Minda Hu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2025. A survey of

personalized large language models: Progress and future directions. *arXiv preprint arXiv:2502.11528*.

Jiahong Liu, Wenhao Yu, Quanyu Dai, Zhongyang Li, Jieming Zhu, Menglin Yang, Tat-Seng Chua, and Irwin King. 2026. Perfit: Exploring personalization shifts in representation space of LLMs. In *The Fourteenth International Conference on Learning Representations*.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023. Think-in-memory: Recalling and post-thinking enable LLMs with long-term memory. *arXiv preprint arXiv:2311.08719*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Charles Packer, Vivian Fang, Shishir_G Patil, Kevin Lin, Sarah Wooders, and Joseph_E Gonzalez. 2023. MemGPT: Towards LLMs as operating systems.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Thomas Adler, Lukas Gruber, Markus Holzleitner, Milena Pavlović, Geir Kjetil Sandve, Victor Greiff, David Kreil, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. 2020. Hopfield networks is all you need. *arXiv preprint arXiv:2008.02217*.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. 2025. From human memory to AI memory: A survey on memory mechanisms in the era of LLMs. *arXiv preprint arXiv:2504.15965*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-MEM: Agentic memory for LLM agents. *arXiv preprint arXiv:2502.12110*.

BY Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. 2025. General agentic memory via deep research. *arXiv preprint arXiv:2511.18423*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.*, 43(6).

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

A LLM Prompts

This appendix provides the core LLM prompts used in the HeLa-Mem system.

A.1 Hebbian Distillation

The Reflective Agent uses the following prompt to extract structured knowledge from memory clusters identified as hubs.

Prompt 1: Semantic Memory Extraction Prompt

```
System: You are a knowledge extraction engine analyzing conversation memories.
Extract ONLY factual information with direct evidence.
Output concise, structured entries.

User: Analyze the following memory cluster and extract:

1. USER CHARACTERISTICS:
   - Observable traits (with evidence)
   - Content preferences (with evidence)
   - Interaction patterns

2. FACTUAL INFORMATION:
   - Events with dates and locations
   - Stated preferences
   - Mentioned relationships

Format: Concise bullet points with supporting evidence.

Memory Cluster: {conversation}
```

A.2 Response Generation

The system uses the following prompt to generate responses using retrieved episodic and semantic memories.

Prompt 2: Response Generation Prompt

```
System: You are an AI assistant with access to conversation history.
Answer questions concisely using the provided context.
For dates, use format "15 July 2023".

User:
<EPISODIC MEMORIES>
{episodic_context}

<SEMANTIC KNOWLEDGE>
{semantic_knowledge}

<USER CHARACTERISTICS>
{user_model}

Question: {query}

Provide an extremely concise answer using concrete entities.
Output only the answer content, without labels.
```

B Hebbian Weight Visualization

Figure 6 shows the Hebbian edge weight matrix for the first 20 memory nodes. Stronger weights (darker colors) indicate associations formed through co-activation. The matrix exhibits both local associations near the diagonal and cross-topic connections between distant nodes, demonstrating that Hebbian learning captures semantic relationships beyond temporal adjacency. Nodes with red borders have high total connectivity, making them candidates for Hebbian Distillation.

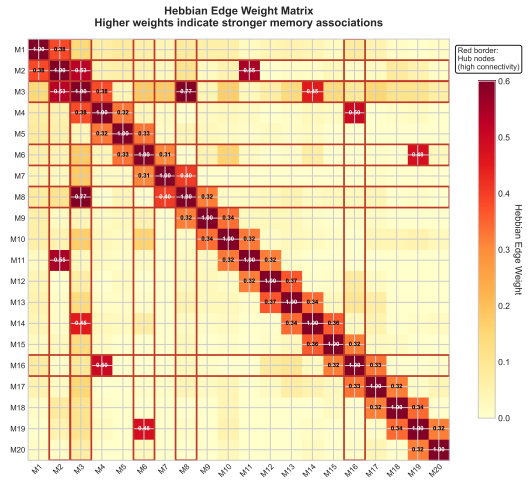

Figure 6: Hebbian edge weight matrix for the first 20 memory nodes. Stronger weights (darker colors) indicate more frequent co-activation. Nodes with red borders have high total connectivity across multiple memories.

C Dataset Statistics

We utilize the LoCoMo benchmark, focusing on the long-context conversation split. Table 4 provides the detailed statistics of the 10 conversations used in our experiments, while Table 5 details the distribution across different question categories.

D Additional Benchmark: LongMemEval-S

We additionally evaluate HeLa-Mem on LongMemEval-S, a 500-item long-term conversational memory benchmark. We use GPT-4o-mini as both the backbone model and the LLM judge. For retrieval, HeLa-Mem uses top-15 episodic memories and top-5 semantic memories (20 total). The best setting on this benchmark uses learning rate $\eta = 0.02$, decay rate $\lambda = 0.995$, spreading activation strength

Table 4: LoCoMo dataset overview.

Metric	Value
Number of Conversations	10
Avg. Turns per Conversation	~ 300
Avg. Tokens per Conversation	$\sim 9,000$
Total Question-Answer Pairs	1,986

Table 5: Distribution of question types in the evaluation set.

Category	Count	Percentage
Single-hop	841	42.3%
Multi-hop	282	14.2%
Temporal	321	16.2%
Open-domain	96	4.8%
Adversarial	446	22.5%
Total	**1,986**	**100.0%**

$\beta = 0.1$, spreading threshold $\theta = 0.4$, keyword weight 0.7, and max flipped items $m = 3$. In Table 7, **Single** denotes the merged single-hop group combining Single-User, Single-Asst, and Single-Pref. Baseline numbers are reported from Fang et al. (2025) under the same total retrieval budget.

Table 6: Overall accuracy on LongMemEval-S.

Method	ACC (%)
LangMem	37.20
MemoryOS	44.80
Mem0	53.61
FullText	56.80
NaiveRAG	61.00
A-MEM	62.60
HeLa-Mem	**65.40**

Table 7: Category-wise accuracy on LongMemEval-S.

Method	Temporal	Multi-Session	Knowledge-Update	Single
LangMem	15.79	20.30	66.67	55.13
MemoryOS	32.33	31.06	48.72	64.74
Mem0	40.15	46.21	70.12	62.82
FullText	31.58	45.45	76.92	78.21
NaiveRAG	39.85	48.48	67.95	**85.90**
A-MEM	47.36	48.87	64.11	84.62
HeLa-Mem	**50.38**	**57.14**	**78.21**	78.85

The category-wise values are not directly averaged to obtain the overall ACC because the category sizes are unequal.

HeLa-Mem achieves the best overall accuracy and the best performance on the three reasoning-

intensive categories: Temporal, Multi-Session, and Knowledge-Update.

E LLM Usage Statement

We use publicly available large language model tools as writing assistants to check grammar and polish a small number of sentences. All technical content, claims, and contributions are conceived, written, and verified by the authors. For schematic figures, several icons or visual elements are refined with the assistance of LLM-based design tools, while the figure layout, semantics, and interpretation are fully determined by the authors. Since this paper involves LLM-related research, all model usage that affects experiments, analysis, or results is explicitly documented in Section Experiments. No other parts of the manuscript are generated or substantively rewritten by an LLM.