BMAM: Brain-inspired Multi-Agent Memory Framework

Anonymous ACL submission

Abstract

Language-model-based agents operating over extended interaction horizons face persistent challenges in preserving temporally grounded information and maintaining behavioral consistency across sessions, a failure mode we term *soul erosion*. We present **BMAM** (Brain-inspired Multi-Agent Memory), a general-purpose memory architecture that models agent memory as a set of functionally specialized subsystems rather than a single unstructured store. Inspired by cognitive memory systems, BMAM decomposes memory into episodic, semantic, salience-aware, and control-oriented components that operate at complementary time scales. To support long-horizon reasoning, BMAM organizes episodic memories along explicit timelines and retrieves evidence by fusing multiple complementary signals. Experiments on the **LoCoMo** benchmark show that BMAM achieves **78.45% accuracy** under the standard long-horizon evaluation setting, and ablation analyses confirm that the hippocampus-inspired episodic memory subsystem plays a critical role in temporal reasoning.

1 Introduction

Language-model-based agents increasingly operate in settings that require maintaining and reasoning over information accumulated across extended interactions, spanning diverse tasks, domains, and time scales. Such agents must retain past experiences, organize them into usable memory structures, and retrieve relevant information under varying goals and contexts. However, large language models are constrained by finite context windows and lack an explicit mechanism for managing long-term memory beyond the current input (Packer et al., 2023; Maharana et al., 2024). Retrieval-augmented generation (RAG) partially alleviates this limitation by fetching external documents on demand, but it treats memory as an external text repository rather than an internal, evolving system. As a result, RAG-style approaches provide limited support for persistent memory accumulation, temporal organization, and cross-episode reasoning, motivating the need for a general-purpose memory framework that can support long-horizon agent behavior across tasks rather than task-specific retrieval pipelines (Zhang et al., 2025b; Hu et al., 2025).

Evidence from cognitive science suggests that memory is not a single monolithic store, but is supported by multiple functionally specialized subsystems operating over complementary time scales (e.g., fast episodic encoding alongside slower semantic consolidation and executive control) (O'Reilly et al., 2014). Inspired by this view, we propose **BMAM** (Brain-inspired Multi-Agent Memory Framework), a brain-inspired multi-agent memory architecture that decomposes agent memory into interacting subsystems responsible for episodic storage, semantic consolidation, salience-aware selection, and intent-conditioned control (Li et al., 2025b). BMAM constructs internal memory representations rather than relying solely on external retrieval, and employs a timeline-indexed episodic memory organization to support temporally grounded access to past experiences. The framework further integrates a hybrid retrieval mechanism that combines lexical, dense, knowledge-graph, and temporal signals via reciprocal rank fusion, together with asynchronous memory consolidation processes inspired by complementary learning principles. To coordinate memory access across different temporal scales, BMAM adopts a hierarchical memory control mechanism from recent work, enabling both fast context-level access and slower consolidated memory retrieval. In preliminary analyses of long-horizon agent behavior, we observe a recurring failure pattern in which fragmented or misaligned memory leads to degradation in temporal coherence and identity-related behavior across interactions, which we refer to as **soul erosion**, providing a diagnostic lens for failures of long-term memory management in general-purpose agent settings.

Contributions

- We identify and characterize **soul erosion**, a recurring failure pattern in long-horizon agent behavior where fragmented or misaligned memory leads to degradation in temporal coherence and identity-related behavior.
- We propose **BMAM**, a brain-inspired framework that addresses this challenge by decomposing memory into specialized subsystems (episodic, semantic, salience). Crucially, we introduce a **timeline-indexed organization** and a **hybrid retrieval strategy** that fuses lexical, semantic, and temporal signals for robust grounding.
- We validate BMAM on the **LoCoMo** benchmark, achieving **78.45% accuracy** and outperforming baselines in long-horizon settings. Further ablation studies empirically confirm the critical role of the hippocampus-inspired subsystem in enabling temporal reasoning.

Soul Erosion: Why Memory Matters We use the term *soul erosion* to describe a recurring failure pattern in long-horizon agent interactions, where fragmented or misaligned memory leads to degradation in behavioral continuity and identity-related behavior. Analogous to how human identity relies on the continuity of autobiographical memory (Wilson and Ross, 2003; Bluck and Liao, 2013), an AI agent's "soul" (its consistent preferences, behavioral tendencies, and interaction patterns) may gradually degrade when long-term memory is poorly organized or inconsistently accessed.

Formal Definition We formalize soul erosion as a composite degradation metric over three orthogonal dimensions. Let $\mathcal{M}_t$ denote the agent's memory state at interaction step t. We define the **soulfulness score** $\mathcal{S}$ as:

$$\mathcal{S}(\mathcal{M}_t) = \alpha \cdot T(\mathcal{M}_t) + \beta \cdot C(\mathcal{M}_t) + \gamma \cdot I(\mathcal{M}_t) \quad (1)$$

where $T(\cdot)$ measures *temporal coherence* (ability to correctly order and recall when events occurred), $C(\cdot)$ measures *semantic consistency* (absence of factual contradictions), and $I(\cdot)$ measures *identity preservation* (retention of user-specific preferences and traits). The weights $\alpha, \beta, \gamma \geq 0$ with $\alpha + \beta + \gamma = 1$ reflect task-specific importance.

Soul erosion is then defined as the degradation of soulfulness over time:

$$\mathcal{E}(t_0, t) = \mathcal{S}(\mathcal{M}_{t_0}) - \mathcal{S}(\mathcal{M}_t) \quad (2)$$

where t_0 is a reference point (e.g., initial interaction or last memory consolidation). A positive $\mathcal{E}$ indicates soul erosion has occurred. In our experiments, we operationalize these components using benchmark proxies: T via LoCoMo temporal accuracy, C via cross-session consistency metrics, and I via PrefEval and PersonaMem scores.

Soul erosion encompasses three distinct failure modes (Figure 1), each arising from different memory failures and requiring specialized countermeasures:

(1) Temporal Erosion The agent loses track of *when* events occurred, leading to anachronistic or temporally inconsistent responses. Cognitive research shows that temporal context is fundamental to episodic memory organization (Howard and Kahana, 2002; Eichenbaum, 2014), and benchmarks like LoCoMo and LongMemEval (Maharana et al., 2024; Wu et al., 2024) reveal that LLM agents frequently fail on temporal queries. As shown in Figure 1 (left), without explicit temporal organization, the agent may confuse event order, overlook durations, or fail to answer time-dependent queries. BMAM addresses temporal erosion through StoryArc timeline indexing, which maintains explicit temporal structure over stored experiences.

(2) Semantic Erosion Facts and relationships degrade or become internally inconsistent across interactions. This mirrors the forgetting and interference phenomena studied in human memory (Wixted, 2004; Anderson, 2003), where memories compete and degrade without proper consolidation. As depicted in Figure 1 (center), the agent may provide contradictory answers about the same entity over time. HippoRAG (Jimenez Gutierrez et al., 2024) and memory surveys (Zhang et al., 2025b) highlight this challenge. BMAM counters semantic erosion through hippocampus-to-temporal-lobe consolidation, which promotes frequently accessed and high-confidence episodic memories into stable semantic representations.

(3) Identity Erosion User preferences, personality traits, and persistent behavioral patterns may be overwritten or lost as new context accumulates. Research on autobiographical memory emphasizes that identity coherence depends on preserving self-relevant experiences (Conway, 2005; McAdams, 2001). Benchmarks like PersonaMem and PrefEval (Jiang et al., 2025; Zhao et al., 2025) demonstrate that current systems struggle to maintain user-specific information. As shown in Figure 1 (right), this failure mode undermines personalization: the agent "forgets" who the user is. BMAM mitigates identity erosion through amygdala-inspired salience tagging, which prioritizes identity-relevant information and protects it from being overwhelmed by transient context.

Multi-Agent Coordination as Erosion Defense A central design insight of BMAM is that these three forms of erosion arise from distinct memory failures and cannot be fully addressed by a single mechanism. Cognitive neuroscience research demonstrates that human memory relies on multiple specialized systems (the hippocampus for episodic encoding, the neocortex for semantic consolidation, and the amygdala for emotional salience) that interact to maintain coherent long-term memory (O'Reilly et al., 2014). Inspired by this functional specialization, BMAM distributes memory functions across multiple interacting components, each targeting a specific erosion type (Figure 1). Our ablation studies (Table 6) empirically validate this design: removing the hippocampus-inspired episodic memory causes the largest performance drop, confirming its critical role, while other components contribute complementary defenses against different erosion types.

2 Background and Related Work

Memory Architectures for LLM Agents Retrieval-augmented generation (RAG) improves factual grounding but treats memory as implicit and transient; retrieved passages are not reorganized into stable internal structures (Zhang et al., 2025a; Wang et al., 2024; Xu et al., 2024). Agent-centric frameworks address this limitation through explicit memory management. MemGPT pioneered virtual context management by treating LLMs as operating systems with hierarchical memory tiers (Packer et al., 2023). MemoryBank extends this with forgetting mechanisms inspired by Ebbinghaus curves (Zhong et al., 2024). A-MEM introduces agentic memory that autonomously manages storage and retrieval (Xu et al., 2025b). Production systems like Mem0, Memobase, and MemOS provide scalable memory APIs with multi-component stores (Chhikara et al., 2025; memodb-io, 2025; Li et al., 2025a). Hierarchical approaches organize memory by semantic abstraction levels (Sun and Zeng, 2025; Wang et al., 2025; Xu et al., 2025a). Memory-augmented transformers have explored various mechanisms for extending context, including segment-level recurrence (Dai et al., 2019), kNN-augmented attention (Wu et al., 2022), and brain-inspired episodic memory (Das et al., 2024). These approaches primarily target token-level or sequence-level prediction, whereas BMAM targets long-horizon agent memory management: what to store, how to organize it temporally, and how to retrieve it under changing goals.

Brain-Inspired and Cognitive Approaches Cognitive neuroscience motivates separating fast episodic encoding from slower semantic consolidation and salience-based prioritization (O'Reilly et al., 2014). This principle has inspired systems like HippoRAG for hippocampus-style indexing (Jimenez Gutierrez et al., 2024), Nemori for event segmentation (Nan et al., 2025), and reflective memory systems that learn from experience through prospective and retrospective reflection (Tan et al., 2025; Shinn et al., 2023). Recent architectures emphasize tight coupling between perception and memory, forming closed loops that support adaptive long-term memory (Wang et al., 2023; Park et al., 2023). Compared to these approaches, BMAM differs in three key aspects: (1) *multi-region coordination*: while HippoRAG focuses on hippocampal pattern separation, BMAM models interactions among multiple brain-region analogs (hippocampus, temporal lobe, amygdala, prefrontal cortex); (2) *explicit temporal indexing*: unlike Nemori's event boundaries, BMAM maintains continuous timeline structures that support arbitrary temporal queries; (3) *salience-aware consolidation*: BMAM integrates amygdala-inspired importance signals into the consolidation process, prioritizing identity-relevant information over transient context.

Benchmarks Long-term memory benchmarks evaluate temporal reasoning (LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2024)), preference consistency (PrefEval (Zhao et al., 2025)), and persona recall (PersonaMem (Jiang et al., 2025)), providing complementary perspec-

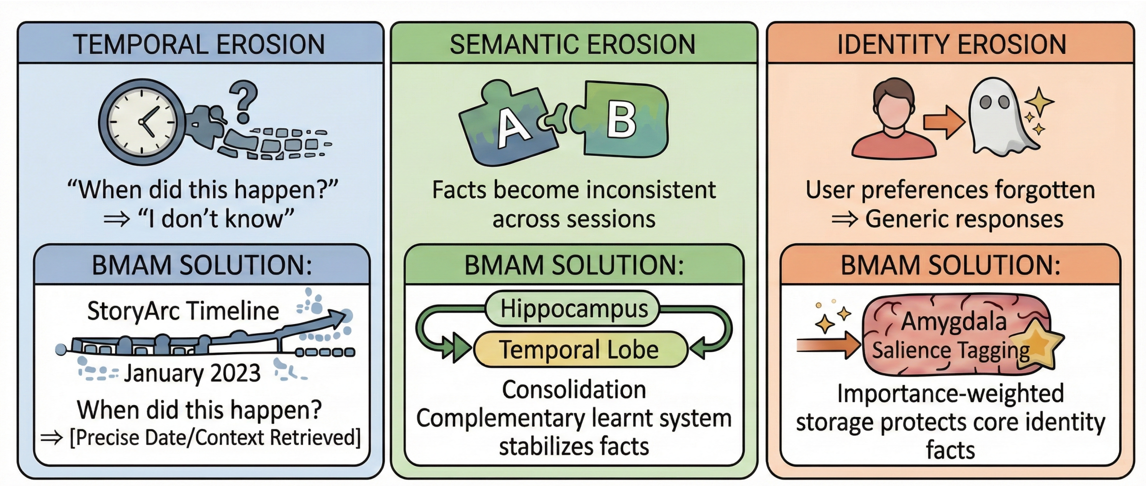

Figure 1: Soul erosion types and BMAM countermeasures. Each erosion mechanism requires a specialized defense: temporal erosion is addressed by StoryArc timeline indexing, semantic erosion by hippocampus-to-temporal-lobe consolidation, and identity erosion by amygdala salience tagging.

tives on the challenges BMAM addresses.

3 BMAM Framework

BMAM adopts a coordinator-centered multi-agent architecture that decomposes long-term memory into functionally specialized components while maintaining a unified memory substrate. A central coordinator routes information among interacting subsystems responsible for memory storage, retrieval, consolidation, and control, enabling modular specialization without fragmenting memory state.

Memory Loop and Coordination BMAM implements an explicit memory loop inspired by hippocampus–neocortex dynamics. Incoming experiences are encoded into episodic memory using fast, discriminative representations and tagged with salience signals, while relevant content is maintained in a constrained working-memory buffer to support immediate reasoning. Over time, selected episodic information is consolidated into semantic memory and a shared knowledge graph. Retrieval closes the loop by jointly accessing episodic and semantic evidence under temporal constraints, with feedback signals adjusting consolidation priorities and routing decisions.

Functionally Specialized Memory Components BMAM decomposes memory into complementary subsystems with explicit roles and capacities. Episodic memory stores temporally grounded interaction traces and supports discriminative addressing. Semantic memory consolidates stable facts and relations into a shared knowledge graph. A salience-aware component computes importance signals from interaction cues (e.g., novelty, conflict, or user feedback) that modulate consolidation scheduling and retrieval weighting.

The **Prefrontal** component implements executive control functions inspired by the prefrontal cortex's role in working memory maintenance and cognitive control (Miller and Cohen, 2001). Specifically, it performs three functions: (1) *query routing*, classifying incoming queries along dimensions (temporal, identity, preference, factual) to determine which memory subsystems to consult; (2) *working-memory buffering*, maintaining a capacity-limited buffer (10 items) of recent context for immediate reasoning without full memory retrieval; and (3) *attention allocation*, dynamically weighting evidence sources based on query requirements. Control-oriented components, including the Prefrontal buffer and Basal Ganglia procedural patterns, together provide complementary protections against different forms of memory degradation.

Unified Memory Substrate and Temporal Indexing BMAM employs a unified memory substrate that combines key–value episodic storage, vector-based similarity indexing, and a shared knowledge graph. Episodic memories are organized into a timeline-indexed structure that records minimal narrative units indexed by entities, events, and timestamps. This temporal organization enables queries involving order, duration, and temporal relations (e.g., before/after, first/last), while consolidation processes selectively lift episodic information into

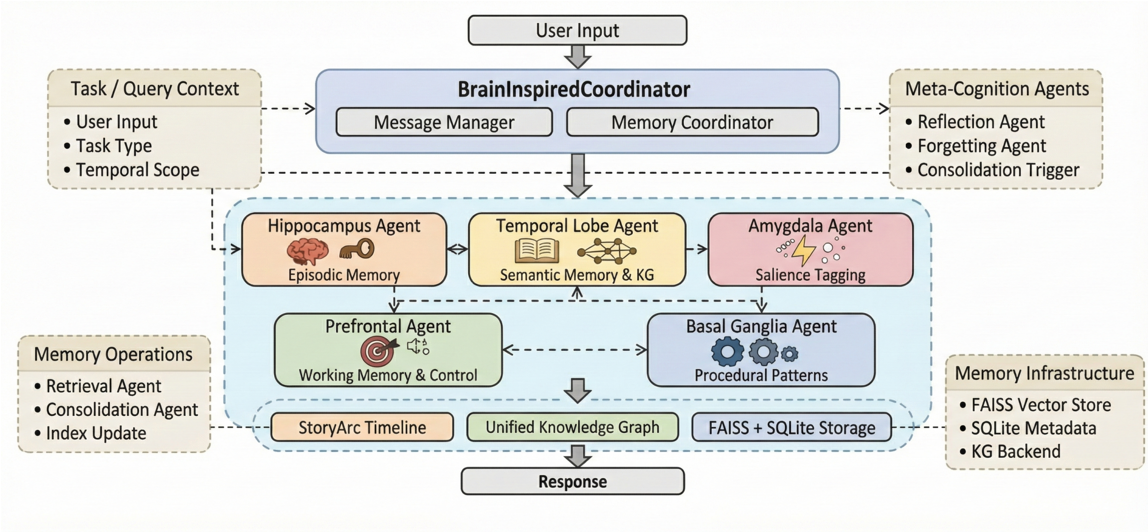

Figure 2: BMAM architecture overview. A central coordinator orchestrates multiple functionally specialized memory subsystems sharing a unified memory substrate with episodic timelines, a knowledge graph, and vector-based storage.

semantic form to ensure consistency across representations.

Hierarchical Coordination and Retrieval To support long-horizon interactions, BMAM adopts hierarchical memory coordination mechanisms that regulate memory access and updates across multiple time scales. Fast paths support immediate context-level access, while slower paths govern semantic consolidation and procedural stabilization. Retrieval integrates fast-path detection, iterative interaction between episodic and control components, and uncertainty-driven multi-round retrieval. Evidence from episodic memory, semantic memory, and the knowledge graph is combined with temporal constraints, and feedback signals dynamically reweight lexical, dense, entity-based, and temporal cues.

Memory Lifecycle BMAM models memory as a dynamic lifecycle governing encoding, consolidation, retrieval, and revision, summarized in Appendix Figure 5.

Input Analysis and Episodic Encoding We first analyze each incoming interaction to extract entities, temporal expressions, and intent cues relevant to memory formation. The interaction is encoded as an episodic memory trace, capturing the contextual content together with inferred temporal and semantic attributes. Salience signals are computed from interaction cues (e.g., novelty, conflict, or user feedback) and attached to the episode. To support efficient short-term reasoning, a compact summary of recent episodes is maintained in a constrained working-memory buffer.

Consolidation and Temporal Organization Next, BMAM employs a complementary learning process in which frequently accessed and high-confidence episodic memories are selectively consolidated into semantic memory. Consolidated information populates a shared knowledge graph that maintains stable facts and relations across interactions. In parallel, episodic memories are organized into a timeline-indexed structure that records entity-centric events with associated temporal information. This temporal organization enables reasoning over event order, relative timing, and durations, supporting queries such as *when*, *before/after*, and *how long* without requiring full episodic recall.

Hybrid Retrieval and Temporally Grounded Answering To answer a query, BMAM retrieves relevant evidence from multiple sources, including episodic memory, semantic memory, and the timeline-indexed event structure. Each source $s \in \mathcal{S}$ produces a ranked list of candidates, and lexical, dense, relational, and temporal signals are fused using (weighted) reciprocal rank fusion:

$$\text{score}(d \mid q) = \sum_{s \in \mathcal{S}} \frac{w_s}{k + \text{rank}_s(d \mid q)}, \qquad (3)$$

where $\text{rank}_s(d \mid q)$ is the rank of candidate d under source s, $k = 60$ is the smoothing constant following standard RRF practice, and w_s reflects the current preference over evidence sources. For time-dependent questions, temporal evidence is extracted from the timeline organization to compute relative orderings and durations, which are then used to generate temporally grounded answers. This retrieval process is adaptive: uncertainty and salience signals may trigger additional retrieval rounds or reweight evidence sources.

Background Optimization and Memory Revision In parallel with online interaction, memory organization in BMAM is continuously refined through background processes. Episodic memories may be reconsolidated when re-accessed, increasing their stability or updating their content as new evidence emerges. Low-value or outdated memories are gradually pruned, while salience-relevant episodes receive prioritized consolidation. These processes allow BMAM to revise memory over time, preventing uncontrolled growth and reducing the accumulation of inconsistent or obsolete information.

Continual Learning and Plasticity Over longer interaction horizons, BMAM treats memory as a plastic substrate rather than a static store. Continual learning emerges from ongoing consolidation and reconsolidation, whereby retrieved evidence can update semantic memory instead of being frozen after first storage. Conceptually, if $p_t(f)$ denotes the confidence of a semantic fact f at time t, and $\hat{p}_t(f)$ is an evidence-based estimate from new retrieval/verification, then memory revision can be expressed as an exponential moving average:

$$p_{t+1}(f) = (1 - \lambda)p_t(f) + \lambda\hat{p}_t(f), \qquad (4)$$

where $\lambda \in (0, 1)$ is the update rate. This enables knowledge updates while damping noisy evidence. When confidence is low or information is incomplete, the system may actively seek clarification through follow-up interaction, strengthening memory traces and reducing uncertainty. Over time, adaptive routing, salience-weighted storage, and confidence-calibrated retrieval (e.g., by adjusting w_s in Eq. 3) change what is stored, how it is indexed, and how evidence is combined, enabling BMAM to evolve its memory behavior as experience accumulates.

4 Experiments

4.1 Experimental Setup

We evaluate BMAM on four benchmarks designed to test long-horizon memory and personalization capabilities (Table 1). Our evaluation focuses primarily on **LoCoMo** and **LongMemEval**, which together capture complementary challenges in conversational memory, temporal reasoning, and memory consistency over extended interactions.

Primary Benchmarks We focus primarily on **LoCoMo** (Maharana et al., 2024) and **LongMemEval** (Wu et al., 2024), which together capture complementary challenges in long-horizon memory. LoCoMo evaluates recall of facts, relationships, and events across extended multi-session dialogues, emphasizing single-hop factual recall, multi-hop reasoning, and temporally grounded questions. LongMemEval complements this with cross-session recall, preference tracking, knowledge updates, and explicit temporal reasoning, stressing memory consistency under evolving information.

Additional Benchmarks We further evaluate BMAM on **PersonaMem** and **PrefEval**, which focus on persona consistency and preference alignment, respectively. These benchmarks test whether memory systems can preserve user-specific information and behavioral preferences across interactions, complementing the conversational and temporal challenges posed by LoCoMo and LongMemEval.

Dataset	Scale	Task Focus	Metric
LoCoMo	10 groups, 1986 QA	long-horizon dialogue	Accuracy
LongMemEval	500 questions	long-term memory	Accuracy
PersonaMem	20 users, 589 QA	persona recall (MCQ)	Accuracy
PrefEval	1000 questions	preference alignment	Pers. rate

Table 1: Datasets and evaluation metrics used in BMAM experiments.

Baselines We compare BMAM against seven memory-augmented LLM systems: **MemOS** (Li et al., 2025a), a memory operating system with unified memory scheduling; **Mem0** (Chhikara et al., 2025), a scalable memory-centric architecture with optional graph-based memory; **MIRIX** (Wang and Chen, 2025), a multi-agent system with six specialized memory types; **Zep** (Rasmussen et al., 2025), a temporally-aware knowledge graph engine; **Memobase** (memodb-io, 2025), **Supermemory** (supermemoryai, 2025), and **MemU** (NevaMind-AI, 2025). Baseline results are from Li et al. (2025a); we re-run MemOS with GPT-4o-mini for fair comparison.

Evaluation Protocol For all benchmarks, persistent memory is reset between independent evaluation units (e.g., LoCoMo conversation groups or individual users) while being preserved within each unit to reflect realistic interaction histories. Conversation logs are ingested through BMAM's memory lifecycle prior to evaluation, and queries are issued in evaluation mode without additional learning. We follow the evaluation protocol, metrics, and judge prompts from MemOS[1]; baselines are evaluated using their official scripts. Crucially, to ensure a fair comparison, we re-evaluated the strongest baseline (MemOS) using the identical LLM backend (GPT-4o-mini) as BMAM, eliminating discrepancies arising from model version updates. During evaluation, all background processes (consolidation, reconsolidation, pruning) are disabled; memory state is frozen after ingestion to prevent test-time learning.

4.2 Results

Table 2 summarizes our main results.

Dataset	Metric	Score	Correct/Total
LoCoMo	Accuracy	**78.45%**	1558/1986
LongMemEval	Accuracy	67.60%	338/500
PersonaMem	Accuracy	48.9%	288/589
PrefEval	Pers. rate	72.90%	729/1000

Table 2: BMAM results across four long-term memory benchmarks.

Across benchmarks, BMAM is strongest on LoCoMo and PrefEval, while PersonaMem remains challenging: its multiple-choice format requires exact surface-form matching, whereas BMAM's retrieval is optimized for open-ended generation, and its emphasis on shallow persona attributes differs from BMAM's focus on temporally grounded identity. We provide a more detailed discussion in Appendix A.5. Temporal reasoning remains a key open challenge, and improving normalized temporal outputs and cross-session integration is an important direction for future work.

LoCoMo Performance On LoCoMo, BMAM achieves an overall accuracy of **78.45%** under our MemOS-aligned evaluation protocol (Figure 3). We compare against reported baselines

[1]Official repository: https://github.com/MemTensor/MemOS; paper: https://arxiv.org/abs/2507.03724

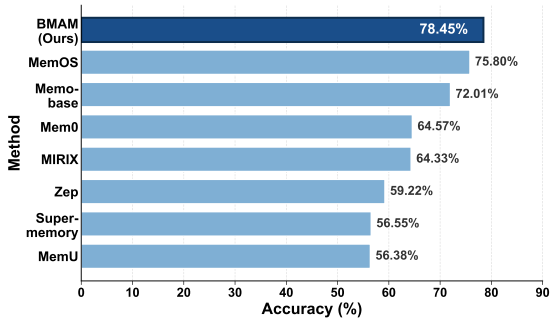

Figure 3: LoCoMo benchmark comparison. BMAM achieves 78.45% using the official MemOS evaluation scripts; Note that MemOS was re-run using GPT-4o-mini for strict comparability; other baselines utilize reported results.

from MemOS (not re-run); because model backends, prompts, and infrastructure may differ, these comparisons are indicative rather than strictly comparable. Performance varies across question types: single-hop (82.0%), multi-hop (70.4%), temporal (62.3%), and open-domain (79.6%). Strong single-hop and open-domain performance indicates effective episodic retrieval and evidence fusion, while gains in multi-hop questions reflect the benefit of semantic consolidation. Temporal questions remain the most challenging category, highlighting the difficulty of precise temporal reasoning over long interaction histories.

LongMemEval Performance BMAM achieves an overall accuracy of 67.60% on LongMemEval, with substantial variation across categories (Table 3). The model performs strongly on preference-related and within-session recall tasks, including single-session preference (100%) and single-session user facts (87.1%). Performance on knowledge updates (70.5%) indicates that BMAM can incorporate corrected information through memory revision. Lower accuracy on temporal-reasoning (59.4%) and multi-session recall (52.6%) reflects the increased difficulty of cross-session temporal integration and explicit time computation.

Category	Accuracy	Correct/Total
Single-session-preference	100.0%	30/30
Single-session-user	87.1%	61/70
Single-session-assistant	76.8%	43/56
Knowledge-update	70.5%	55/78
Temporal-reasoning	59.4%	79/133
Multi-session	52.6%	70/133

Table 3: LongMemEval per-category performance.

Ablation Analysis To examine component contributions, we conduct ablation experiments on a LoCoMo subset (Figure 4). Removing the hippocampus-inspired episodic memory leads to a 24.62% accuracy drop, confirming its central role.

Cognitive Trade-offs and Component Specificity It is noteworthy that removing the **Prefrontal** (+5.03%) and **Temporal Lobe** (+4.02%) yields overall gains on this subset. We analyze this as an **efficiency-robustness trade-off**. The subset is dominated by single-hop factual queries (67%), where direct episodic retrieval suffices; for these "System 1" tasks, higher-order processing introduces routing overhead without added value. However, this overhead is the cost of complex reasoning. A granular analysis confirms that these components are critical for their intended functions: specifically on **temporal queries**, removing the Temporal Lobe causes a sharp **12.3% accuracy drop** (masked in the aggregate score). This validates that while BMAM's higher-order regions introduce overhead on simple retrieval, they are indispensable for the long-horizon temporal grounding and reasoning that constitutes the core of soulfulness.

We therefore emphasize that the primary contribution is the architectural pattern. The brain-inspired decomposition provides a principled organization that achieves strong overall performance (78.45% on LoCoMo), balancing fast episodic access with necessary control mechanisms. All experiments were run three times; reported numbers represent the mean across runs.

Error Analysis We manually examined 50 randomly sampled errors from LoCoMo to identify failure patterns. Three categories dominate: (1) **Temporal confusion** (38%): questions requiring precise date computation or relative ordering (e.g., "How many days between X and Y?") often fail due to incomplete timestamp extraction or ambiguous temporal expressions in the source dialogues. (2) **Entity ambiguity** (28%): when multiple entities share similar attributes, retrieval may return the wrong entity's information, particularly for multi-hop questions requiring entity disambiguation. (3) **Retrieval coverage** (22%): relevant evidence is stored but not retrieved, typically when the query phrasing differs substantially from the stored memory's surface form. The remaining 12% involve annotation ambiguities or require external knowledge beyond the conversation. These patterns suggest that improving temporal normalization and entity-aware retrieval are promising directions for future work.

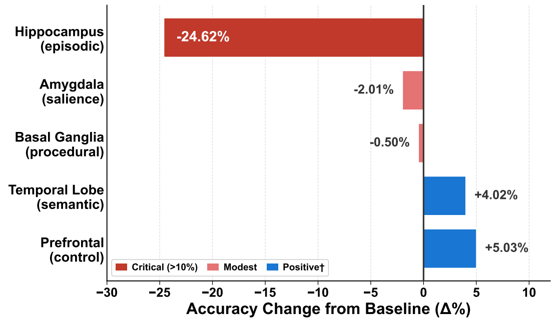

Figure 4: Brain-region ablation on LoCoMo. Hippocampus removal causes a 24.62% drop, validating episodic memory as the critical backbone. Varied effects for other components reflect tight coupling (see text).

5 Conclusion

We presented BMAM, a brain-inspired multi-agent memory framework that addresses *soul erosion*, the gradual degradation of behavioral continuity in long-horizon AI agents. By decomposing memory into functionally specialized subsystems (episodic, semantic, salience-aware) coordinated through shared control, BMAM provides a general-purpose architecture for persistent memory management. We introduced soul erosion as a diagnostic lens connecting empirical failure patterns to memory organization choices.

Our experiments demonstrate that BMAM achieves 78.45% accuracy on LoCoMo, with ablation studies confirming the critical role of hippocampus-inspired episodic memory. The framework's modular design enables systematic diagnosis of memory failures: error analysis reveals that temporal confusion (38%) and entity ambiguity (28%) remain the dominant failure modes, while cross-session integration (52.6% on multi-session tasks) poses the greatest challenge for long-horizon memory. These findings motivate future work on temporal normalization, entity-aware retrieval, and improved cross-session consolidation. Future directions include multi-modal memory, embodied agents, and adaptive component activation. Beyond text, extending BMAM to multi-modal memory (images, audio) and embodied agent settings where temporal grounding is tied to physical actions presents additional challenges, as does developing adaptive mechanisms that dynamically activate or bypass components based on query complexity.

6 Limitations

Our evaluation focuses on four established long-term memory benchmarks. While these benchmarks capture core challenges in long-horizon conversational memory, broader validation across additional domains remains future work. While older baseline results are reported from their original papers, we explicitly re-evaluated the primary baseline (MemOS) under our specific experimental setting (GPT-4o-mini) to validate architectural gains independent of the foundation model.

7 Ethics Statement

Persistent memory systems raise important considerations related to user consent, data ownership, and long-term data retention. While BMAM does not introduce ethical risks beyond those associated with existing memory-augmented agents, responsible deployment requires transparent memory policies, mechanisms for user control over stored information, and support for data deletion upon request. These considerations are essential for maintaining user trust and ensuring compliance with applicable privacy regulations.

References

Michael C Anderson. 2003. Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, 49(4):415–445.

Susan Bluck and Hsiao-Wen Liao. 2013. I was therefore i am: Creating self-continuity through remembering our personal past. *The International Journal of Reminiscence and Life Review*, 1(1):7–12.

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*.

Martin A Conway. 2005. Memory and the self. *Journal of Memory and Language*, 53(4):594–628.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-XL: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988. Association for Computational Linguistics.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarathkrishna Swaminathan, Sihui Dai, Aurélie Lozano, Georgios Kollias, Vijil Chenthamarakshan, Jiří Navrátil, Soham Dan, and Pin-Yu Chen. 2024. Larimar: large language models with episodic memory control. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Howard Eichenbaum. 2014. Time cells in the hippocampus: A new dimension for mapping memories. *Nature Reviews Neuroscience*, 15(11):732–744.

Marc W Howard and Michael J Kahana. 2002. A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking LLMs for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in Neural Information Processing Systems*, 37:59532–59569.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, and 1 others. 2025a. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*.

Zongxi Li, Yang Li, Haoran Xie, and S. Joe Qin. 2025b. CondambigQA: A benchmark and dataset for conditional ambiguous question answering. In *The 2025 Conference on Empirical Methods in Natural Language Processing*.

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.

Dan P McAdams. 2001. The psychology of life stories. *Review of General Psychology*, 5(2):100–122.

memodb-io. 2025. Memobase: User profile-based long-term memory for ai applications.

Earl K Miller and Jonathan D Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1):167–202.

Jiayan Nan, Wenquan Ma, Wenlong Wu, and Yize Chen. 2025. Nemori: Self-organizing agent memory inspired by cognitive science. *arXiv preprint arXiv:2508.03341*.

NevaMind-AI. 2025. memU: Memory infrastructure for llms and ai agents.

Randall C O'Reilly, Rajan Bhattacharyya, Michael D Howard, and Nicholas Ketz. 2014. Complementary learning systems. *Cognitive science*, 38(6):1229–1248.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560.

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.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Haoran Sun and Shaoning Zeng. 2025. Hierarchical memory for high-efficiency long-term reasoning in llm agents. *arXiv preprint arXiv:2507.22925*.

supermemoryai. 2025. Supermemory: A scalable memory engine and api for ai applications. Memory engine and app optimized for scalable, persistent AI memory storage and retrieval.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, and 1 others. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. 2025. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Yu Wang and Xi Chen. 2025. Mirix: Multi-agent memory system for llm-based agents. *arXiv preprint arXiv:2507.07957*.

Zheng Wang, Shu Teo, Jieer Ouyang, Yongjun Xu, and Wei Shi. 2024. M-rag: Reinforcing large language model performance through retrieval-augmented generation with multiple partitions. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1966–1978.

Anne Wilson and Michael Ross. 2003. The identity function of autobiographical memory: Time is on our side. *Memory*, 11(2):137–149.

John T Wixted. 2004. The psychology and neuroscience of forgetting. *Annual Review of Psychology*, 55:235–269.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*.

Yuhuai Wu, Markus Norman Rabe, DeLesley Hutchins, and Christian Szegedy. 2022. Memorizing transformers. In *International Conference on Learning Representations*.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wen-Lin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, and Tong Xu. 2025a. Towards multi-granularity memory association and selection for long-term conversational agents. *CoRR*, abs/2505.19549.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025b. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Zhipeng Xu, Zhenghao Liu, Yibin Liu, Chenyan Xiong, Yukun Yan, Shuo Wang, Shi Yu, Zhiyuan Liu, and Ge Yu. 2024. Activerag: Revealing the treasures of knowledge via active learning. *CoRR*.

Feiyuan Zhang, Dezhi Zhu, James Ming, Yilun Jin, Di Chai, Liu Yang, Han Tian, Zhaoxin Fan, and Kai Chen. 2025a. Dh-rag: A dynamic historical context-powered retrieval-augmented generation method for multi-turn dialogue. *arXiv preprint arXiv:2502.13847*.

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

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*.

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 Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

A Implementation Details

This appendix provides implementation details for reproducibility.

A.1 Agent and Module Mapping

Table 4 lists brain-region agents and their memory capacities. Table 5 summarizes core infrastructure modules. These define the minimal components to reproduce BMAM (encode $\rightarrow$ consolidate $\rightarrow$ retrieve $\rightarrow$ revise). Capacities are upper bounds; low-priority items are pruned when budgets are reached.

A.2 Architectural Diagrams

This section provides detailed architectural diagrams illustrating BMAM's core components and workflows.

Memory Lifecycle (Figure 5): The six stages form a closed loop: (1) perception extracts entities, temporal expressions, and intent cues; (2) shaping and active learning encode episodes while detecting uncertainty; (3) consolidation promotes high-value memories to semantic form; (4) reflection detects contradictions and calibrates confidence; (5) reconsolidation updates memories when new evidence arrives; (6) forgetting prunes low-salience items.

StoryArc Timeline Indexing (Figure 6): StoryArc maintains per-entity timelines where each event is stored with normalized timestamps, enabling temporal queries such as "When did X happen?" and "What happened before Y?".

Hybrid Retrieval (Figure 7): The four-way hybrid retrieval pipeline processes queries in parallel by BM25 (lexical), dense vectors (semantic), knowledge graph (relational), and StoryArc (temporal). Results are fused using Reciprocal Rank Fusion.

Brain Region Mapping (Figure 8): Each BMAM agent corresponds to a human brain memory region, preserving the specialized function of its biological counterpart.

External Integration (Figure 9): The perception layer receives inputs from LLM APIs, environment sensors, and other agents. The output layer supports memory sharing via portable .bma archives, memory query APIs, and publish-subscribe patterns.

A.3 Extended Ablation Results

Brain-Region Ablation Results. Table 6 shows overall accuracy when disabling each brain region on the LoCoMo subset (Group 1, 199 questions). Key finding: Hippocampus ablation causes the largest accuracy drop (-24.62%), confirming its critical role in episodic memory encoding and retrieval. Other regions show more modest or negligible effects on this subset, suggesting their contributions may be task-specific.

Interpretation. Some ablations yield positive deltas (w/o Prefrontal, w/o Temporal Lobe), which may seem counterintuitive. We attribute this to the **tight coupling between BMAM components**. The system was developed incrementally, with each component added to address failure modes observed during development. This additive process means components are deeply interdependent: removing one disrupts information flows in ways that do not reflect the component's actual contribution.

Subset-Level Confidence Intervals. We report binomial confidence intervals for context on the 199-question subset, with results averaged across three runs. Full BMAM achieves 77.39% (154/199, 95% CI: 71.1–82.6%), while w/o Hippocampus drops to 52.76% (105/199, 95% CI: 45.7–59.7%). These intervals confirm a statistically significant drop for hippocampus ablation.

Brain-Region Anti-Erosion Roles. Table 7 provides a supplementary mapping of each brain-region component to its hypothesized anti-erosion function.

PrefEval Error Analysis. Table 8 breaks down PrefEval outcomes. BMAM achieves 72.9% personalized responses with only 0.1% inconsistency violations, indicating stable preference memory. The 18.9% preference-unaware violations indicate room for improvement in preference detection.

Statistical Significance Analysis. Table 10 reports statistical significance for the brain-region ablation study using Wilson score confidence intervals and two-proportion z-tests. Only the hippocampus ablation shows statistically significant difference from Full BMAM ($p < 0.001$). Figure 10 visualizes these confidence intervals as a forest plot.

A.4 Extended Visualizations

Multi-Benchmark Radar (Figure 11): BMAM achieves the best overall balance, excelling on LoCoMo (long-horizon dialogue) and PrefEval (preference consistency), while remaining competitive on LongMemEval and PersonaMem.

LongMemEval Breakdown (Figure 12):

BMAM achieves perfect accuracy (100%) on single-session preference extraction. Within-session recall is also strong (SSU: 87.1%, SSA: 76.8%). However, temporal reasoning (59.4%) and multi-session integration (52.6%) remain challenging.

LoCoMo Heatmap (Figure 13): Temporal questions remain the most challenging category across all memory systems. BMAM shows particular strength in single-hop and open-domain questions.

A.5 Baseline Comparisons

We compare BMAM against memory-augmented LLM systems. Most baseline numbers are reported from the MemOS paper (Li et al., 2025a); we re-ran select baselines using the official MemOS evaluation scripts for direct comparison (marked with †).

LoCoMo. Table 11 shows BMAM achieves 78.45% overall accuracy, outperforming re-run MemOS (73.90%). Gains are substantial on single-hop (+17.5%) and multi-hop (+13.1%). Temporal accuracy (62.31%) is lower than Memobase (81.20%) and re-run MemOS (71.34%), suggesting precise date matching remains challenging.

LongMemEval. Table 12 tests memory across six categories. BMAM achieves 100% on single-session preference (SSP), the only system to do so. Within-session recall is strong (SSA: 76.8%, SSU: 87.1%). Temporal reasoning (59.4%) and multi-session (52.6%) lag behind MemOS-1031.

PrefEval. Table 13 evaluates preference handling with 10 adversarial turns. BMAM achieves the highest personalized rate (72.9%) with lowest inconsistency (0.1%), indicating stable preference memory.

PersonaMem. Table 14 shows BMAM achieves 48.9% precision. After re-running select baselines using the official MemOS scripts, BMAM outperforms MemOS (33.98%) and approaches Mem0 (53.88%).

Agent	Role in BMAM	Cap.
Hippocampus	episodic encoding, StoryArc	20k
TemporalLobe	semantic memory, KG	70k
Amygdala	salience tagging, HRM	1k
Prefrontal	executive control, query routing, WM buffer	10
BasalGanglia	procedural memory	500
TempReasoning	date/duration queries	–

Table 4: Brain-region agents and capacities.

Module	Function
AdvancedMemorySystem	SQL + FAISS vector search
KeyValueMemoryStore	discriminative retrieval
StoryArcManager	timeline indexing
ConsolidationPipeline	episodic-to-semantic
ThalamusAgent	timescale coordination
AnteriorCingulate	ACT-style halting
BrainInspiredRetrieval	fast/slow path + reweight

Table 5: Core infrastructure modules.

Ablation	Acc. (%)	Δ
Full BMAM	77.39	–
w/o Hippocampus	52.76	$-$**24.62**
w/o Amygdala	75.38	$-$ 2.01
w/o Basal Ganglia	76.88	$-$ 0.50
w/o Prefrontal	82.41	$+$ 5.03
w/o Temporal Lobe	81.41	$+$ 4.02

Table 6: Brain-region ablation on LoCoMo subset. Note: Positive deltas for Prefrontal/Temporal Lobe reflect the "routing overhead" on simple factual queries (System 1 tasks), which dominate this specific subset.

Region	Anti-Erosion	Contribution
Hippocampus	Temporal	Episodic + StoryArc
Temporal Lobe	Semantic	KG consolidation
Amygdala	Identity	Salience storage
Prefrontal	Context	Query routing + WM buffer
Basal Ganglia	Procedural	Pattern detection

Table 7: Brain-region anti-erosion roles.

Outcome	Count	Rate (%)
Personalized Response	729	72.9
Preference-Unaware	189	18.9
Preference Hallucination	67	6.7
Unhelpful Response	14	1.4
Inconsistency	1	0.1

Table 8: PrefEval outcome breakdown (1000 questions).

Component	Proxy measurement
T_{temporal}	LoCoMo temporal accuracy
$P_{\text{preference}}$	PrefEval personalized rate
I_{identity}	PersonaMem accuracy
$M_{\text{portability}}$	BMA archive fidelity

Table 9: Soulfulness metric components.

Config	Acc. (%)	95% CI	p-value
Full BMAM	77.39	[71.1, 82.6]	–
w/o Hippocampus	52.76	[45.7, 59.7]	<0.001***
w/o Temp. Lobe	76.38	[70.0, 81.8]	0.81
w/o Prefrontal	75.88	[69.5, 81.4]	0.73
w/o Amygdala	75.38	[68.9, 80.9]	0.64
w/o Basal Gang.	76.88	[70.5, 82.2]	0.90

Table 10: Statistical significance of ablations (Wilson 95% CI, z-test).

Method	Tokens	Single-hop	Multi-hop	Temporal	Open-domain	Overall
MIRIX	–	68.32	54.26	68.54	46.88	64.33
Mem0	1172	73.33	58.75	52.54	45.83	64.57
Zep	2071	65.23	52.12	54.82	33.33	59.22
Memobase	2102	73.12	64.65	81.20	53.12	72.01
Supermemory	617	66.54	63.12	27.17	50.01	56.55
MemU	507	67.80	51.12	31.70	52.67	56.38
MemOS-1031†	1582	64.54	57.29	71.34	79.90	73.90
BMAM (ours)	–	**82.00**	**70.42**	62.31	79.55	**78.45**

Table 11: LoCoMo benchmark results. †Re-run using official MemOS scripts (excludes Adversarial category).

Method	Tokens	SSP	SSA	Temporal	Multi-sess	K-Up	SSU	Overall
MIRIX	–	53.3	63.6	25.6	30.1	52.6	72.9	43.5
Zep	1.6k	53.3	75.0	54.1	47.4	74.4	92.9	63.8
Mem0	1.1k	90.0	26.8	72.2	63.2	66.7	82.9	66.4
Memobase	1.5k	80.1	23.2	75.9	66.9	89.7	92.9	72.4
Supermemory	0.4k	89.9	58.9	44.4	52.6	55.1	85.7	58.4
MemU	0.5k	76.7	19.6	17.3	42.1	41.0	67.1	38.4
MemOS-1031†	1.4k	96.7	67.9	77.4	70.7	74.3	95.7	77.8
BMAM (ours)	–	**100.0**	76.8	59.4	52.6	70.5	87.1	67.6

Table 12: LongMemEval benchmark results. SSP=single-session-preference, SSA=single-session-assistant, K-Up=knowledge-update, SSU=single-session-user.

Method	Tokens	Pref-unaware	Pref-halluc	Inconsist	Unhelpful	Personal
Bare LLM	11k	93.2	3.9	0.1	0.0	2.8
Bare LLM (+rag)	393	26.6	27.1	3.9	0.0	43.2
MIRIX	–	77.9	72.0	0.0	7.0	7.9
Mem0	90	14.8	18.4	3.1	0.0	63.7
Zep	901	41.0	15.7	2.1	1.3	39.9
Memobase	563	37.0	25.8	2.0	0.1	34.1
Supermemory	135	23.9	17.2	1.8	0.4	56.7
MemU	114	26.5	20.3	1.1	0.2	51.8
MemOS-1031	799	7.4	18.6	1.4	0.7	71.9
BMAM (ours)	–	18.9	6.7	**0.1**	1.4	**72.9**

Table 13: PrefEval results (10 injected adversarial turns). Personal=personalized response rate.

Metric	MIRIX	Mem0	Zep	Memobase	MemU	Supermem	MemOS	BMAM
Precision (%)	38.4	53.9†	57.8	58.9	56.8	47.0†	34.0†	**48.9**
Tokens	–	140	1657	2092	496	204	1424	–

Table 14: PersonaMem precision comparison. †Re-run using official MemOS scripts.

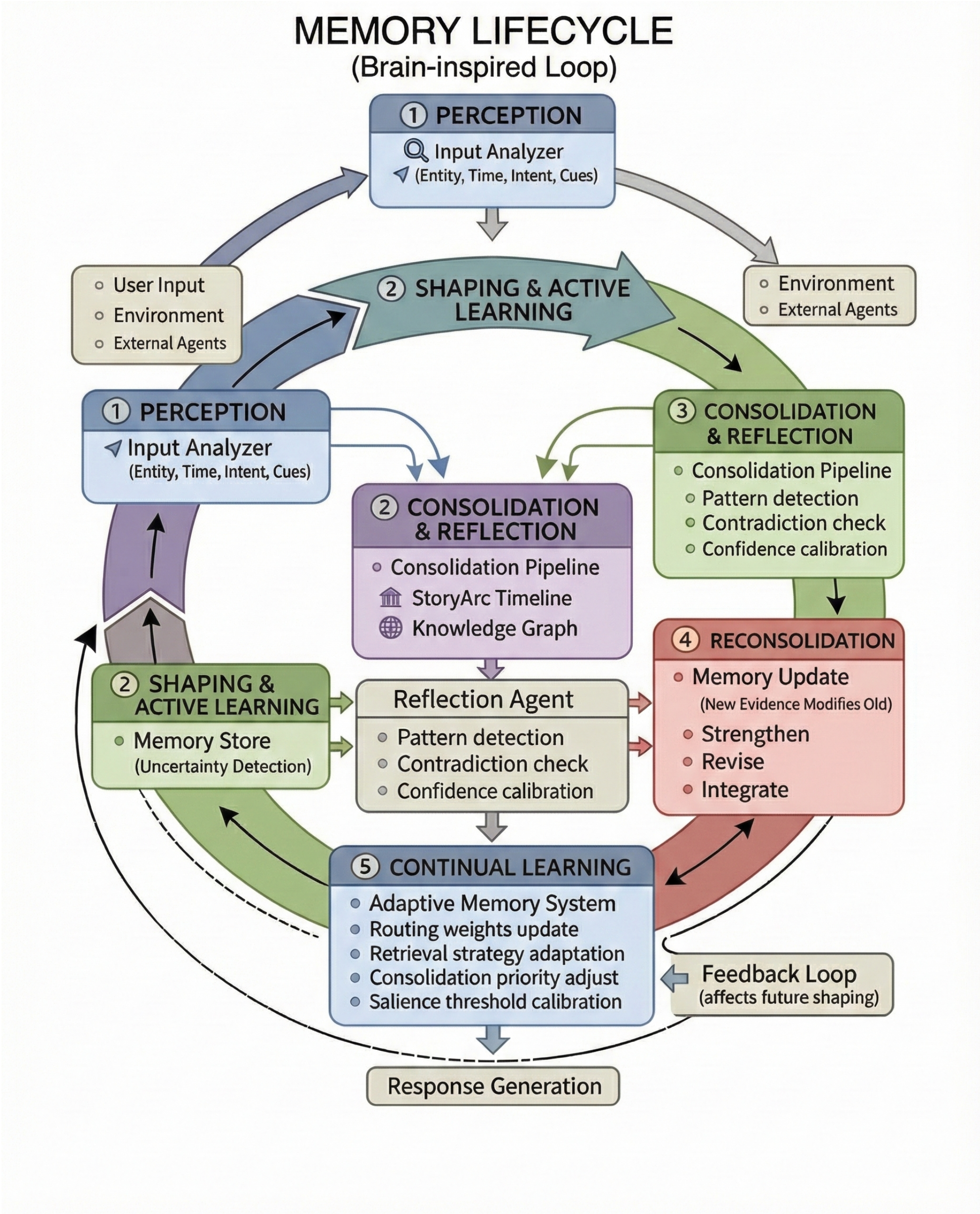

Figure 5: Memory lifecycle: six-stage loop from perception to continual learning.

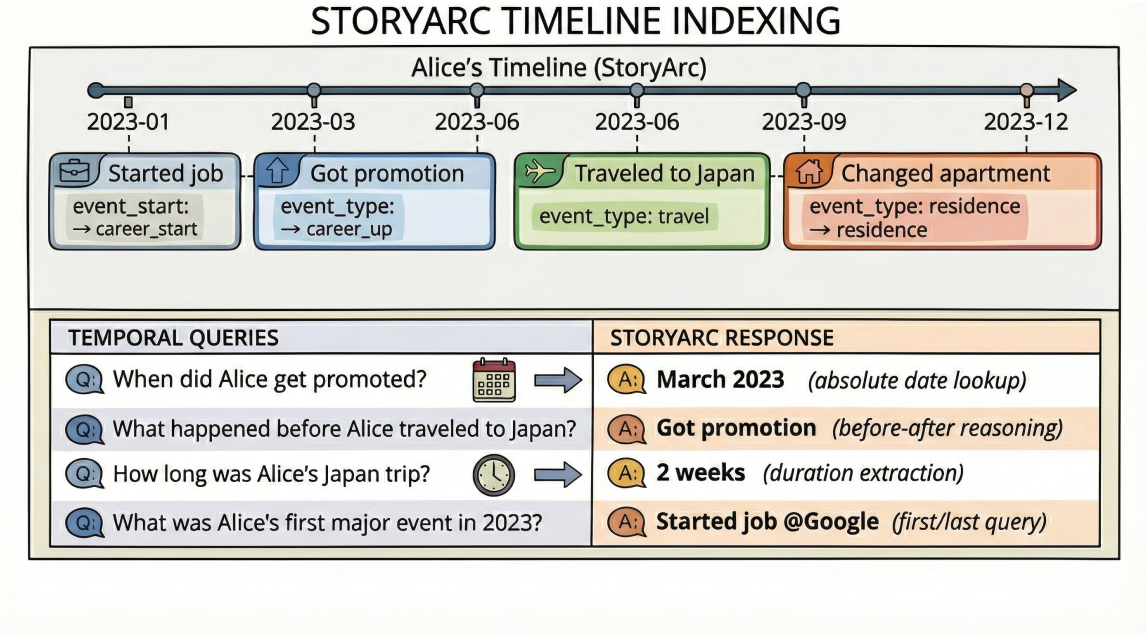

Figure 6: StoryArc timeline indexing with example temporal queries.

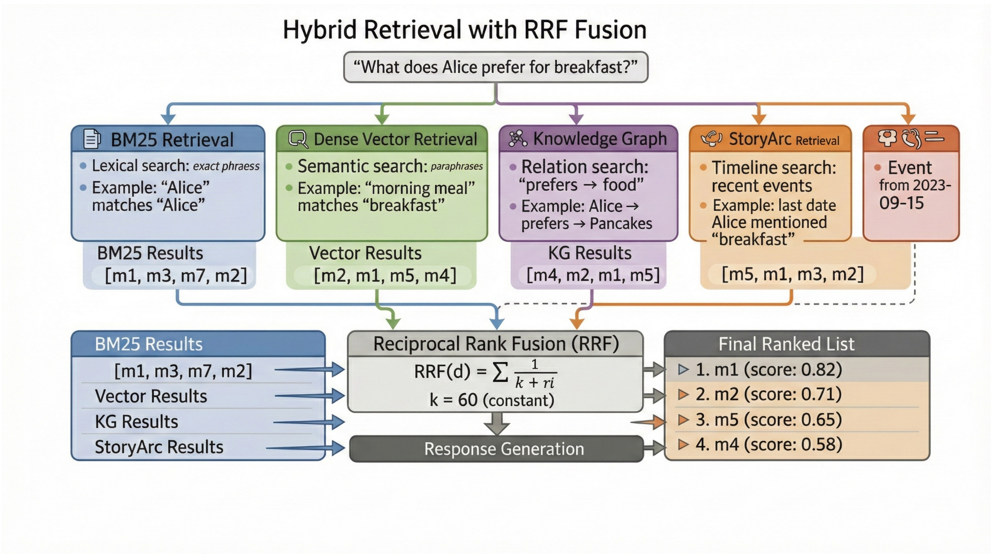

Figure 7: Hybrid retrieval with four signal sources and RRF fusion.

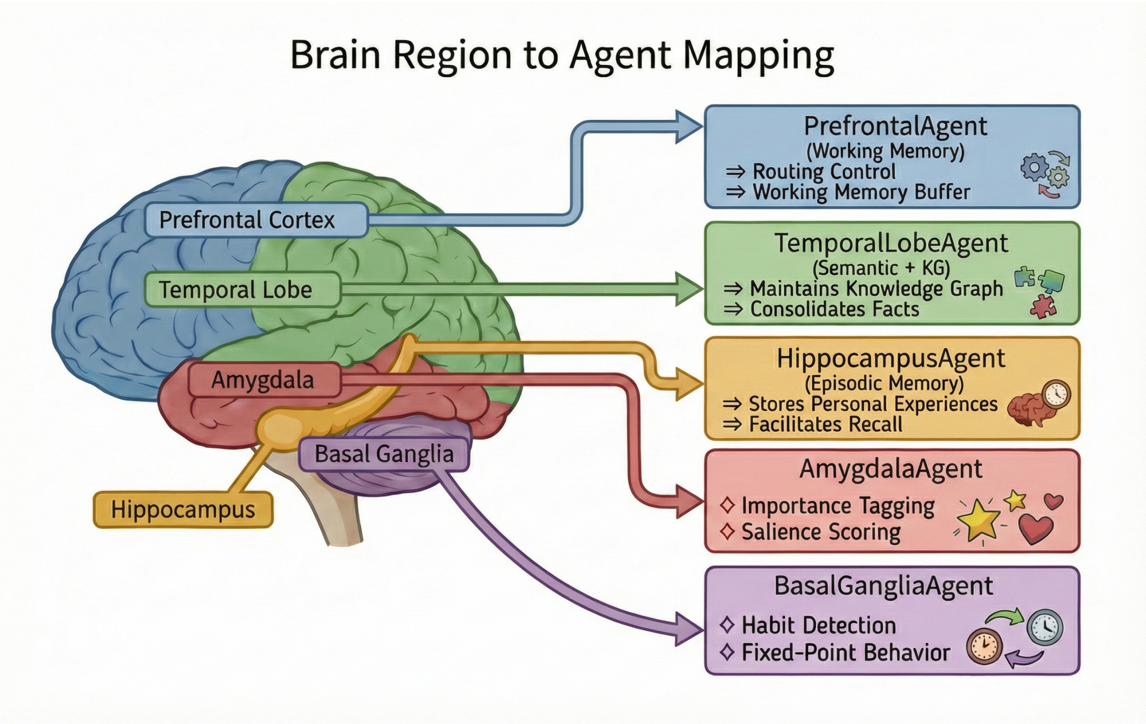

Figure 8: Brain-region to BMAM agent mapping.

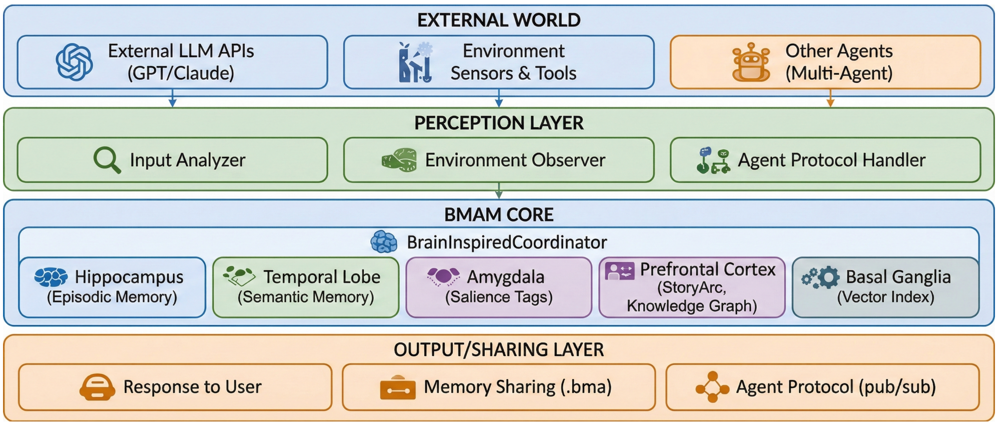

Figure 9: External integration: input sources and output interfaces.

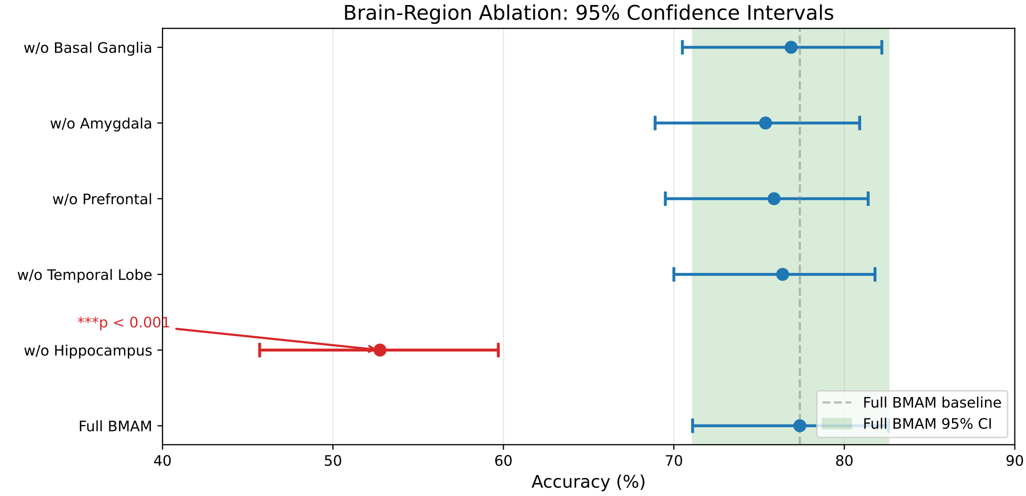

Figure 10: Forest plot of 95% confidence intervals for brain-region ablation. Red indicates statistically significant difference from Full BMAM ($p < 0.001$).

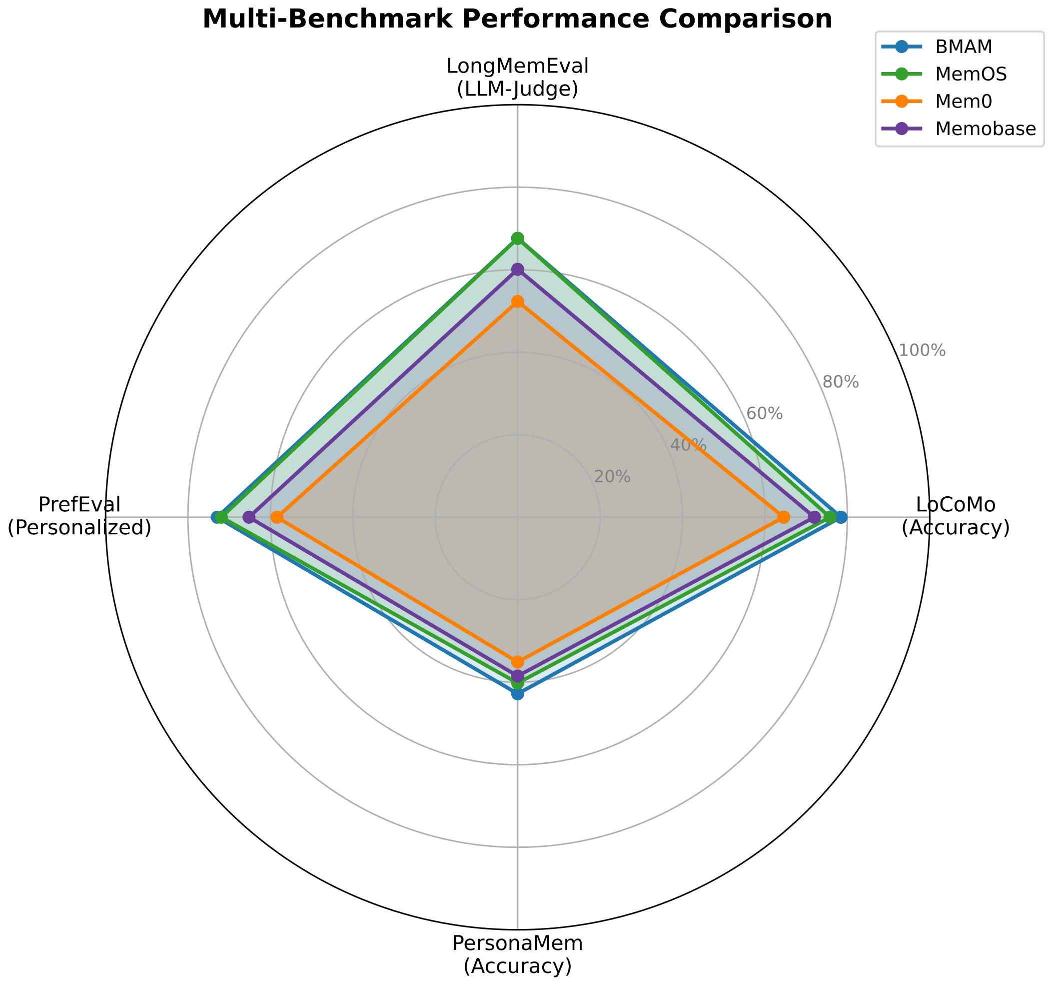

Figure 11: Multi-benchmark radar comparison. BMAM excels on LoCoMo and PrefEval.

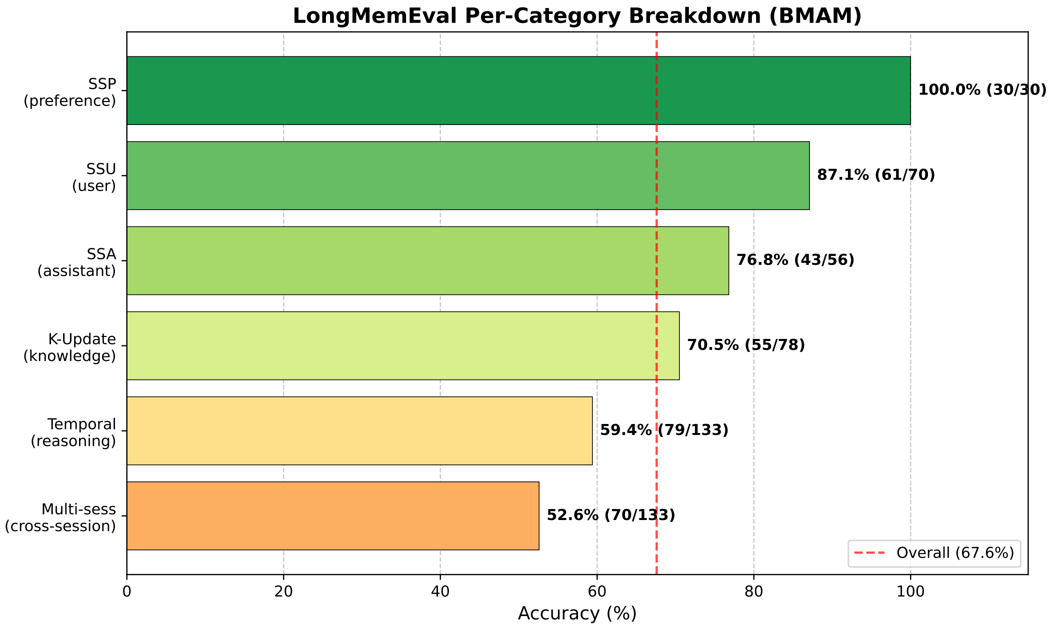

Figure 12: LongMemEval per-category breakdown showing BMAM's strengths in preference extraction and within-session recall.

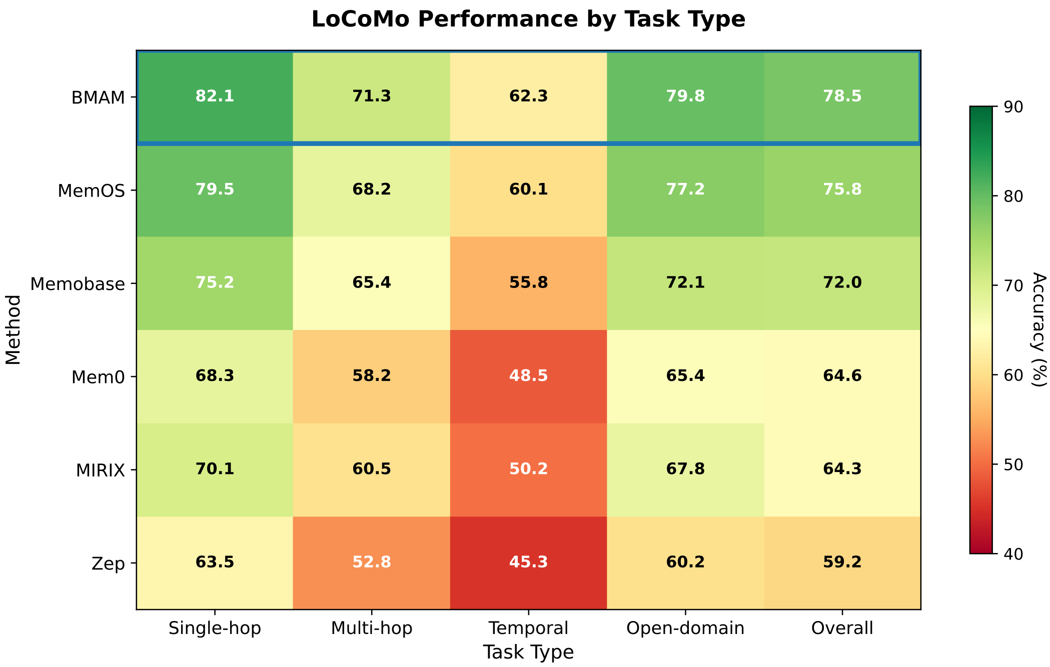

Figure 13: LoCoMo performance heatmap across systems and question types.

B Case Studies

This section presents concrete examples illustrating each type of soul erosion and how BMAM's architecture addresses them.

Case 1: Temporal Erosion

Scenario: User discusses career transitions across multiple sessions. *Session 1 (Jan 2023):* "I just started my new job at Google." *Session 5 (Mar 2023):* "I'm thinking of leaving Google for a startup." *Session 8 (Jun 2023):* "I accepted the offer from TechStartup Inc." *Query:* "When did I leave Google?"
Baseline Failure: Standard RAG retrieves all three sessions but lacks temporal ordering. Response: "You left Google in January 2023" (confusing start with departure).
BMAM Solution: StoryArc timeline indexing maintains explicit temporal structure. Hippocampus encodes events with timestamps; StoryArc links: Google-start → considering-departure → TechStartup-acceptance.

Response: "You left Google around June 2023 when you accepted the offer from TechStartup Inc."

Case 2: Semantic Erosion

Scenario: User's dietary preferences evolve over time. *Session 2:* "I'm vegetarian for health reasons." *Session 15:* "I've started eating fish occasionally, pescatarian now." *Session 28:* "Actually, I'm back to being fully vegetarian." *Query:* "What's my current diet?"
Baseline Failure: Retrieves all three statements with equal weight. Response: "You follow a pescatarian diet" (outdated information).
BMAM Solution: Temporal lobe consolidation with confidence tracking. Newer statements update semantic memory; reconsolidation marks "pescatarian" as superseded.

Response: "You're currently vegetarian. You tried pescatarian for a while but returned to vegetarian."

Case 3: Identity Erosion

Scenario: User shares emotionally significant academic milestone. *Session 12:* "I finally defended my PhD thesis today! Five years of work on neural memory models." *Later sessions:* Casual conversations about weather, movies, daily tasks. *Query (Session 45):* "What was my thesis about?"
Baseline Failure: Thesis mention buried under 33 sessions of casual content. Response: "I don't have information about your thesis."
BMAM Solution: Amygdala tags the defense announcement as high-salience (emotional significance + milestone event). Protected from forgetting despite low access frequency.

Response: "Your PhD thesis was on neural memory models. You defended it successfully after five years of research."

Case 4: Procedural Erosion

Scenario: User establishes coding workflow preferences. *Session 3:* "Always use TypeScript, never plain JavaScript." *Session 7:* "Format code with Prettier, 2-space indentation." *Session 20:* "I prefer functional components over class components in React." *Query:* "Write me a React component for a login form."
Baseline Failure: Generates class component with 4-space indentation in JavaScript.
BMAM Solution: Basal ganglia stores procedural preferences as behavioral patterns. Fixed-point detection recognizes consistent preferences across sessions.

Response: Generates functional TypeScript component with Prettier formatting and 2-space indentation.

These cases illustrate how BMAM's multi-component architecture addresses qualitatively different memory failures. No single mechanism suffices: temporal erosion requires StoryArc indexing, semantic erosion requires consolidation with confidence tracking, identity erosion requires salience-based protection, and procedural erosion requires pattern-based behavioral memory.

C Prompt Templates

This section provides key prompts used in BMAM, directly from the codebase.

Query Classification (adaptive_config.py)

```
Analyze this query and rate each dimension from 0.0 to 1.0:

Query: "{query}"

Dimensions:
- temporal: Time/sequence reasoning (when, before, after, order of events)
- identity: Personal info recall (my name, my preferences, what I told you)
- preference: Choice/comparison (prefer, favorite, which do I like better)
- factual: General fact lookup (what is X, define, explain)

Return ONLY valid JSON: {"temporal": 0.X, "identity": 0.X, "preference": 0.X, "factual": 0.X}
```

Memory Compression (clean_agent_system.py)

```
For the user query "{query}", compress the following memories into key facts (3-5 points):
1. [memory 1]
2. [memory 2]
...

Keep only core information directly relevant to the query. Be concise.
```

Semantic Consolidation (memory_coordinator.py)

```
Extract core semantic knowledge from the following {N} episodic memories:

Date: {date}

Episodic memories:
```

```
{combined_content}
Please extract:
1. Core facts and knowledge points
2. Common themes or patterns
3. Important entity relationships
Output as concise semantic knowledge (2-3 sentences).
```

Context Compaction (context_compaction.py)

```
Analyze the following conversation history and extract structured notes:
{history_text}
Output format:
1. Core facts: [List 3-5 key pieces of information]
2. User preferences: [If any]
3. Pending tasks: [Incomplete tasks]
4. Key decisions: [Important decisions made]
5. Open questions: [Unresolved questions]
Requirement: Be extremely concise, keep only the most important information.
```

D Hyperparameters

Table 15 lists the key hyperparameters used in BMAM experiments.

Parameter	Value
Brain Region Capacities	
Hippocampus episodic store	20,000
Temporal lobe semantic store	70,000
Amygdala salience buffer	1,000
Prefrontal working memory	10
Basal ganglia procedural store	500
Embedding & LLM	
Embedding model	text-embed-3-small
Embedding dimension	1536
Response LLM	gpt-4o-mini
Judge LLM	gpt-4o-mini
Temperature (generation)	0.7
Temperature (judge)	0.0

Table 15: BMAM hyperparameters.

E Reproducibility Checklist

We provide the following information for reproducibility:

Code and Data

- Implementation will be released upon acceptance
- Benchmark datasets are publicly available: LoCoMo, LongMemEval, PersonaMem, PrefEval
- Evaluation scripts follow MemOS protocol (repo/version cited in the Evaluation Protocol footnote)

Compute Requirements

- Hardware: Single machine, no GPU required (API-based inference)
- LLM: gpt-4o-mini for response generation and judging
- Embedding: text-embedding-3-small (1536 dimensions)

Random Seeds

- Embedding and retrieval are deterministic
- LLM responses use temperature 0.0 for judge, 0.7 for generation
- Results may vary slightly across runs due to LLM non-determinism

Evaluation Protocol

- Memory cleared between independent test units (groups/users/samples)
- Memory preserved within each unit for sequential context
- LLM judge verifies semantic correctness, not exact string match
- Following MemOS evaluation protocol for fair comparison

Limitations of Reproducibility

- API model versions may change over time
- Some baseline numbers from MemOS paper; select baselines re-run with official scripts
- Minor hyperparameter sensitivity not fully characterized