Structured Episodic Event Memory

Zhengxuan Lu[1,3], **Dongfang Li**[2*], **Yukun Shi**[2],
Beilun Wang[1*], **Longyue Wang**[4], **Baotian Hu**[2,3]
[1]Southeast University, Nanjing, China
[2]Harbin Institute of Technology (Shenzhen), Shenzhen, China
[3]Shenzhen Loop Area Institute, Shenzhen, China
[4]Alibaba Group, Hangzhou, China
zhengxuanlu@slai.edu.cn, lidongfang@hit.edu.cn

Abstract

Current approaches to memory in Large Language Models (LLMs) predominantly rely on static Retrieval-Augmented Generation (RAG), which often results in scattered retrieval and fails to capture the structural dependencies required for complex reasoning. For autonomous agents, these passive and flat architectures lack the cognitive organization necessary to model the dynamic and associative nature of long-term interaction. To address this, we propose **S**tructured **E**pisodic **E**vent **M**emory (**SEEM**), a hierarchical framework that synergizes a graph memory layer for relational facts with a dynamic episodic memory layer for narrative progression. Grounded in cognitive frame theory, SEEM transforms interaction streams into structured Episodic Event Frames (EEFs) anchored by precise provenance pointers. Furthermore, we introduce an agentic associative fusion and Reverse Provenance Expansion (RPE) mechanism to reconstruct coherent narrative contexts from fragmented evidence. Experimental results on the LoCoMo and LongMemEval benchmarks demonstrate that SEEM significantly outperforms baselines, enabling agents to maintain superior narrative coherence and logical consistency.

1 Introduction

Large Language Models (LLMs) have evolved into sophisticated agents capable of complex reasoning and long-term interaction (Achiam et al., 2023; Xi et al., 2025). However, LLM-based agents remain limited by their finite context windows and the lack of a stable long-term memory system (Packer et al., 2023). This constraint causes reasoning capabilities to degrade over extended sessions, as the agent cannot effectively recall critical information once it exceeds the immediate context. Developing a robust long-term memory is therefore a central challenge in building autonomous agents.

To address this, Retrieval-Augmented Generation (RAG) has emerged as a standard paradigm to supplement LLMs with external knowledge (Lewis et al., 2020). Traditional RAG systems rely on vector similarity to retrieve local text passages (Karpukhin et al., 2020). While efficient, they often struggle with multi-hop reasoning tasks that require understanding the structural dependencies between disparate facts. Recent advancements, such as GraphRAG (Edge et al., 2024) and Mem0 (Chhikara et al., 2025), attempt to solve this by organizing information into graph databases. Nevertheless, these approaches face significant structural limitations. Most existing systems rigidly bind semantic content to fixed graph structures or predefined schemas. This rigidity mitigates the memory from dynamically reorganizing as new knowledge arrives. Consequently, these systems frequently suffer from scattered retrieval (Gutiérrez et al., 2025), where the retrieved context is fragmented into isolated pieces, failing to provide the coherent narrative required for complex reasoning.

To bridge this gap, we propose **S**tructured **E**pisodic **E**vent **M**emory (**SEEM**), a hierarchical framework that transforms continuous interaction streams into a cohesive dual-layer architecture. This system is composed of an Episodic Memory Layer (EML), which captures dynamic narrative progression by extracting and fusing structured Episodic Event Frames (EEFs) inspired by cognitive frame theories (Minsky, 1975; Fillmore, 1976), and a complementary Graph Memory Layer (GML) that organizes static factual details into a relational graph. Both layers are anchored to their original source passages via precise provenance pointers, which ensures that abstract memory units remain traceable to raw passages. During inference, these layers are synergized through a hybrid retrieval process utilizing a Reverse Provenance Expansion (RPE) mechanism, allowing the agent to reconstruct a coherent and logically consistent context

* Corresponding authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6125–6141

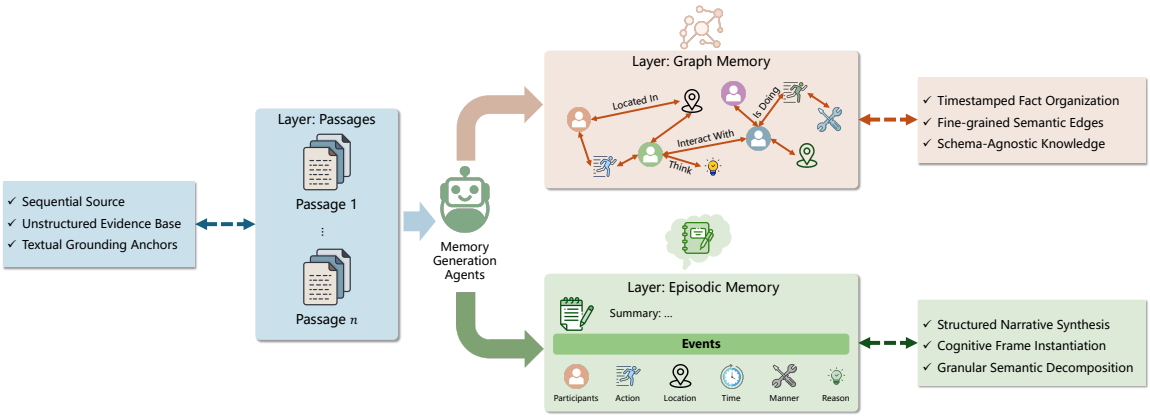

Figure 1: **Overview of the SEEM hierarchical memory architecture.** The system transforms unstructured interaction passages into a dual-layer representation, integrating a semantic Graph Memory Layer for static facts with a structured Episodic Memory Layer for event-centric details. This hierarchical design enables the agent to effectively synergize stable factual knowledge with dynamic narrative contexts for coherent long-term reasoning.

for complex reasoning. Extensive experiments are conducted on the LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025a) benchmarks. Our results demonstrate that SEEM consistently outperforms competitive memory-augmented and dense retrieval baselines. Notably, it surpasses HippoRAG 2 (Gutiérrez et al., 2025) by an absolute margin of 4.4% on LongMemEval. Moreover, supplemental tests under incremental construction settings confirm its stability and robustness for real-world sequential deployment.

Our contributions are summarized as follows:

- We introduce SEEM, a hierarchical framework that synergizes GML for relational facts with EML to capture dynamic narrative progression.
- We propose the EEFs and RPE mechanism, which transform interaction passages into multi-attribute cognitive units linked by provenance pointers to mitigate the scattered retrieval problem.
- We provide extensive empirical validation demonstrating that SEEM outperforms competitive memory-augmented and dense retrieval baselines in maintaining logical consistency and narrative coherence.

2 Related Work

Vector-based RAG. RAG addresses the parametric constraints of LLMs by accessing external corpora via vector similarity (Lewis et al., 2020). However, standard RAG systems predominantly rely on flat vector spaces, which operate in a de-contextualized manner (Gao et al., 2023). This often fails to capture the structural dependencies required for complex multi-hop reasoning, resulting in scattered retrieval where the retrieved context lacks the coherence necessary for consistent long-term interactions (Tang and Yang, 2024; Gutiérrez et al., 2025).

Structured Semantic Memory. To bridge semantic gaps, structure-augmented approaches organize memory into knowledge graphs or hierarchical summaries. GraphRAG (Edge et al., 2024) and RAPTOR (Sarthi et al., 2024) utilize summaries to link related text segments, while HippoRAG 2 (Gutiérrez et al., 2025) leverages graph algorithms to facilitate associative retrieval. Despite these gains, such methods often suffer from lack of structural differentiation, where high-level thematic abstracts and fine-grained facts are entangled (Edge et al., 2024). Furthermore, heavy reliance on LLM-generated summarization can introduce noise, causing performance on basic factual tasks to deteriorate compared to standard RAG (Cuconasu et al., 2024; Wu et al., 2025b).

Episodic Memory. A fundamental distinction exists between general semantic memory and episodic memory grounded in specific spatiotemporal contexts (Tulving et al., 1972). While recent systems such as Mem0 (Chhikara et al., 2025) and Graphiti (Rasmussen et al., 2025) track interaction histories, they may struggle to preserve coherent event contexts due to selective summarization or rigid entity-centric relations. Specifically, these methods frequently fail to integrate essential situational dimensions, including time, causality, and participants, into a unified representation. Con-

sequently, there remains a need for a hierarchical memory to handle the spatiotemporal dynamics of continuous interactions. In contrast, our proposed framework is designed to address this specific gap.

3 Methodology

The SEEM framework transforms a continuous stream of interaction passages into a hierarchical memory architecture composed of two complementary layers. The Episodic Memory Layer (EML) focuses on capturing the narrative progression by extracting and fusing structured Episodic Event Frames (EEFs) while the Graph Memory Layer (GML) organizes static factual relations into a structured relational graph. Both layers are grounded in original passages through a system of provenance pointers, which maintain the link between abstract memory units and their raw passage. During inference, these layers are integrated through a hybrid retrieval process utilizing the Reverse Provenance Expansion (RPE) mechanism to reconstruct a coherent and logically consistent context.

3.1 Problem Formulation

The task of memory-augmented generation in long-term interactions is defined as follows. Given a chronological sequence of interaction passages $\mathcal{P} = \{p_1, p_2, \ldots, p_T\}$, where each passage p_t represents a discrete unit of historical context, and a current user query $q \in \mathcal{Q}$, the objective is to generate a response a that is factually consistent with $\mathcal{P}$ and contextually relevant to q. The set $\mathcal{Q}$ denotes the space of possible user queries.

We formulate this problem as the optimization of a conditional probability $P(a \mid q, \mathcal{P})$. Due to the significant length and semantic density of $\mathcal{P}$, the task requires the construction of an intermediate memory representation $\mathcal{M}$ to bridge the gap between historical evidence and current reasoning. The process is decomposed into two core stages:

Memory Consolidation. We define a transformation function $\Phi : \mathcal{P} \rightarrow \mathcal{M}$ that transforms the raw interaction sequence into a structured representation. This stage is designed to preserve essential thematic and relational information while mitigating the noise inherent in raw text.

Conditioned Generation. A retrieval augmented generation function $G(q, \mathcal{M}) \rightarrow a$ is employed to identify a relevant subset $\mathcal{M}_{sub} \subseteq \mathcal{M}$ based on the query q, leading to the final response generation:

$$a = \arg\max_{a'} P(a' \mid q, \mathcal{M}_{sub}; \theta) \tag{1}$$

where θ denotes the parameters of the underlying generative model. The core challenge lies in designing a representation space $\mathcal{M}$ that can effectively encode the narrative continuity and factual dependencies within $\mathcal{P}$. A desirable property of this transition from $\mathcal{P}$ to $\mathcal{M}$ is provenance preservation, which enables the final generation process to remain grounded in the original source evidence.

3.2 Episodic Memory Generation and Fusion

We introduce a structured episodic memory layer to maintain a coherent understanding of long-term interactions. Instead of storing raw interaction turns, we transform a sequence of passages $\mathcal{P}$ into discrete, event-centric units. As illustrated in Figure 1, this process consists of two phases: (1) extracting structured episodic event frames from each passage and (2) performing associative consolidation to merge related frames.

3.2.1 Episodic Event Frame Extraction

We treat each passage p_t as a source signal that is instantiated into a cognitive frame. Following the principles of frame semantics (Fillmore, 1976), an episodic event frame $\mathbf{e}_t$ captures the structured semantics of p_t. The selection of six core slots, namely Participants, Action, Time, Location, Causality, and Manner, is grounded in Case Grammar to provide a comprehensive event-centric representation while avoiding excessive complexity and noise associated with more fine-grained schemas. This design provides a balanced structural foundation for multi-hop reasoning over long-term interactions. An LLM-based agent $\mathcal{F}_{\text{ext}}$ is employed to parse p_t into semantic role representations together with a high-level summary. Each frame is linked to its source passage via a **provenance pointer** ρ_t^{eml}, which preserves grounding in the original text. The formal definition is given as follows:

$$\begin{aligned}\mathbf{e}_t &= \mathcal{F}_{\text{ext}}(p_t; \theta) \\ &= \Big\langle \rho_t^{eml}, v_{\text{sum}}, \Big\{ \big\langle v_{\text{par}}, v_{\text{act}}, v_{\text{tmp}}, \\ &\qquad v_{\text{spa}}, v_{\text{cau}}, v_{\text{man}} \big\rangle^{(k)} \Big\}_{k=1}^{N_t} \Big\rangle\end{aligned} \tag{2}$$

where v_{sum} denotes the event summary, and the remaining components correspond to semantic roles: Participants (v_{par}), Action (v_{act}), Time (v_{tmp}), Location (v_{spa}), Causality (v_{cau}), and Manner (v_{man}).

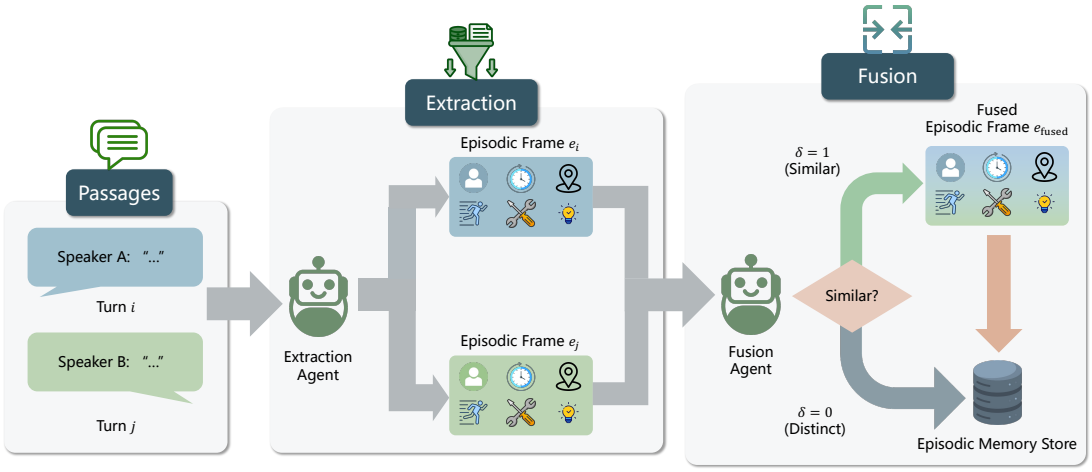

Figure 2: Overview of the associative consolidation and fusion. The $\mathcal{F}_{\text{ext}}$ first transforms raw interaction passages into episodic representations $\mathbf{e}$. The $\mathcal{F}_{\text{judge}}$ then evaluates the similarity between frames, and the fusion agent $\mathcal{F}_{\text{fuse}}$ performs associative consolidation by merging semantically related events. This mechanism maintains a coherent and synthesized episodic memory store.

The variable N_t denotes the number of extracted event instances from passage p_t. Let $\mathcal{E}$ denote the space of all episodic frames constructed in the EML, such that each $\mathbf{e}_t \in \mathcal{E}$ follows the structure defined in Equation (2). This hierarchical representation enables the agent to navigate memory through both high-level abstractions and fine-grained textual anchors.

3.2.2 Associative Consolidation and Fusion

We employ an associative fusion mechanism to mitigate memory fragmentation, consolidating related observations into coherent event representations. When a new candidate frame $\mathbf{e}_t$ is generated, the system retrieves the most relevant historical frame $\mathbf{e}_{\text{prev}}$ and uses an LLM-based judge $\mathcal{F}_{\text{judge}}$ to determine whether they correspond to the same event:

$$\delta_t = \mathcal{F}_{\text{judge}}(\mathbf{e}_t, \mathbf{e}_{\text{prev}}) \tag{3}$$

If $\delta_t = 1$, the fusion agent $\mathcal{F}_{\text{fuse}}$ produces a fused frame $\mathbf{e}_{\text{fused}}$ by integrating the attributes of both inputs and updating the summary v_{sum}. The provenance pointers are aggregated accordingly, with $\rho^{eml}(\mathbf{e}_{\text{fused}})$ linking to the union of all associated source passages. This design allows the fused frame to serve as a unified entry point to evidence distributed across multiple interaction turns.

3.3 Graph Memory Construction

While the EML captures the narrative flow, the GML organizes static facts into a consistent relational structure.

3.3.1 Fact Extraction and Grounding

For each passage p_t, the system extracts a set of relational quadruples $\mathcal{K}_t$ to form a schema-agnostic knowledge graph:

$$\mathcal{K}_t = \{(s, r, o, \tau) \mid s, o \in \mathcal{V}, r \in \mathcal{R}, \tau \in \mathcal{T}\} \tag{4}$$

where $\mathcal{V}$, $\mathcal{R}$, and $\mathcal{T}$ denote the sets of entities, relations, and temporal attributes, respectively. The variables s and o denote subject and object entities, r denotes the relation, and τ represents the temporal validity of the fact. Each node in the graph is linked to its source passage p_t via provenance pointers ρ_t^{gml}. Nodes that exceed a vector similarity threshold are merged to maintain graph integrity, thereby bridging lexical variations across different passages.

3.4 Hybrid Retrieval and Context Integration

During inference, we integrate the structured facts from the GML with the narrative details from the EML through a multi-stage retrieval process.

3.4.1 Relational Propagation and Passage Retrieval

The system initiates retrieval by extracting structured quadruples from the query q to ensure structural alignment with the GML. A shared semantic encoder transforms each query-derived quadruple into a dense vector representation. The retrieval engine then computes the semantic similarity between these query vectors and the pre-indexed embeddings of the facts store within the GML using cosine similarity. By ranking these scores across the relational space, the system identifies the most

relevant facts to form the initial seed set $\mathcal{K}_{top}$. We then execute a propagation algorithm (Haveliwala, 2002) using $\mathcal{K}_{top}$ as the seed set to compute a distribution over graph nodes. This relational traversal identifies the set of most relevant initial passages $\mathcal{P}_{ret}$ through their provenance pointers.

3.4.2 Reverse Provenance Expansion

Initial retrieval often suffers from context fragmentation, as critical details of an event may be distributed across multiple turns that lack direct lexical overlap with the query. We address this issue by leveraging the episodic representations $\mathbf{e}$ in the EML as a semantic bridge. While the GML retrieves isolated factual fragments, the RPE mechanism uses these fragments as entry points to activate the fused narrative structures encoded in $\mathbf{e}$, thereby reconstructing the complete episodic context required for coherent reasoning.

We first retrieve the event frames associated with the initially retrieved passages:

$$\mathcal{E}_{ret} = \{\Psi(p) \mid p \in \mathcal{P}_{ret}\} \tag{5}$$

where $\Psi : \mathcal{P} \to \mathcal{E}$ maps each passage p to its corresponding frame $\mathbf{e} = \Psi(p)$ constructed during memory consolidation. Each $\mathbf{e} \in \mathcal{E}_{ret}$ corresponds to an instantiated frame of the form defined in Equation (2).

We then perform reverse provenance expansion. For each retrieved frame $\mathbf{e}$, we access its aggregated provenance links, denoted by $\rho^{eml}(\mathbf{e})$, which trace $\mathbf{e}$ back to the source passages accumulated during the fusion process described in Section 3.2. The evidence set is then expanded as:

$$\mathcal{P}_{final} = \mathcal{P}_{ret} \cup \bigcup_{\mathbf{e} \in \mathcal{E}_{ret}} \rho^{eml}(\mathbf{e}) \tag{6}$$

This design ensures that once any fragment of an event is activated, all corresponding textual supports are incorporated into the final context, enabling reconstruction of a coherent and complete narrative for downstream reasoning.

3.4.3 Context Synthesis

The final reasoning context $\mathbf{C}$ is synthesized by serializing the expanded passages $\mathcal{P}_{final}$, the structured event frames $\mathcal{E}_{ret}$, and the relational facts $\mathcal{K}_{top}$. This composite context enables the LLM to resolve temporal ambiguities and maintain logical consistency by cross-referencing high-level facts with nuanced episodic evidence.

Finally, the agent generates the response a conditioned on the query q and the synthesized context $\mathbf{C}$. We model this process as autoregressive sequence generation, where the LLM acts as a decoder G that selects the most probable candidate response:

$$\begin{aligned} a &= G(q, \mathbf{C}) \\ &= \arg\max_{a'} P(a' \mid q, \mathbf{C}; \theta) \\ &= \arg\max_{a'} \prod_{i=1}^{|a'|} P(y_i \mid y_{<i}, q, \mathbf{C}; \theta) \end{aligned} \tag{7}$$

where $a' = (y_1, y_2, \ldots, y_{|a'|})$ denotes a candidate response sequence and y_i is its i-th token. By prepending the structured memory evidence directly to the input space, the model can perform integrated reasoning across both episodic and relational knowledge, so that the final output remains grounded in the original evidence while being guided by the high-level semantic structure captured during memory construction.

4 Experimental Setup

4.1 Datasets

To rigorously evaluate the long-term memory and reasoning capabilities of our framework, we conduct experiments on two representative benchmarks: (1) **LongMemEval** (Wu et al., 2025a) serves as a comprehensive testbed for memory-augmented chat assistants, designed to simulate dynamic, evolving user-agent interactions. The dataset comprises 500 manually curated questions that assess five core memory competencies. These include *information extraction* (spanning single-session user, assistant, and preference details), *multi-session reasoning* for synthesizing fragmented information, *temporal reasoning* regarding event timelines, and *knowledge updates* to track changing user states. This benchmark is particularly challenging due to its requirement for maintaining factual consistency across extensible chat histories. (2) **LoCoMo** (Maharana et al., 2024) focuses on the comprehension of extremely long-term, open-domain conversations. Derived from long-form multi-session dialogues that span up to 32 sessions with an average of 16k tokens, this benchmark provides a rigorous assessment of long-range dependency modeling. We utilize its question answering component, which consists of 1,986 samples categorized into five distinct reasoning types: *single-hop* and *multi-hop* reasoning for

Method	LoCoMo			LongMemEval
	BLEU-1	F1	J	Acc.
Dense Retrieval				
KaLM-Embedding-V2.5 (Zhao et al., 2025)	44.4	47.9	64.6	55.6
NV-Embed-v2 (Lee et al., 2025)	53.0	57.9	74.7	58.4
Memory-based Frameworks				
Mem0 (Chhikara et al., 2025)	34.2	43.3	54.1	56.7
A-MEM (Xu et al., 2025)	45.7	44.6	61.9	55.2
HippoRAG 2 (Gutiérrez et al., 2025)	53.8	58.3	76.2	60.6
SEEM (Ours)	**56.1**	**61.1**	**78.0**	**65.0**

Table 1: Performance comparison on LoCoMo and LongMemEval. The best results are highlighted in **bold**.

context retrieval, *temporal* understanding, *open-domain* knowledge integration, and *adversarial reasoning* to test robustness against hallucinations on unanswerable queries.

4.2 Metrics

We evaluate SEEM using a combination of lexical and semantic metrics to capture both surface-level similarity and high-level factual consistency. For LoCoMo, we employ token-level F1 (Maharana et al., 2024) and BLEU-1 (Papineni et al., 2002) for lexical comparison. To further assess semantic correctness and factual accuracy, we utilize LLM-as-a-Judge (J). Specifically, the judge evaluates model responses using the multi-dimensional evaluation prompts introduced in Mem0 (Chhikara et al., 2025), with DeepSeek-V3.2 (Liu et al., 2025) serving as the underlying scoring engine. For LongMemEval, we strictly adhere to the evaluation protocol described in Wu et al. (2025a), which utilizes the LLM to perform binary assessments of answer correctness and reports the resulting accuracy. These metrics collectively provide a rigorous basis for measuring performance across diverse long-term interaction scenarios.

4.3 Baselines

We compare SEEM against the following approaches: **KaLM-Embedding-V2.5** (Zhao et al., 2025) employs a compact decoder-only architecture modified with bidirectional attention and mean-pooling, leveraging high-quality data scaling and advanced training techniques to achieve competitive performance as a versatile and efficient embedding model. **NV-Embed-v2** (Lee et al., 2025) optimizes a decoder-only LLM architecture by incorporating a latent attention layer and bidirectional attention mechanisms to yield high-performance generalist text embeddings for dense retrieval. **HippoRAG 2** (Gutiérrez et al., 2025) adopts a neurobiologically grounded framework that synergizes Personalized PageRank with retrieval-augmented generation, facilitating complex multi-hop reasoning through the integration of dense vector retrieval and sparse knowledge graph structures. **A-MEM** (Xu et al., 2025) implements an agentic memory system inspired by the Zettelkasten method, enabling the dynamic construction and autonomous evolution of interconnected memory notes to refine knowledge representations over time. **Mem0** (Chhikara et al., 2025) provides a scalable memory architecture that dynamically extracts and consolidates conversational history into salient facts, supporting explicit operations to maintain long-term consistency in agentic interactions.

4.4 Implementation Details

We standardize the backbone models across all methods to ensure a fair comparison. We primarily employ Qwen3-Next-80B-A3B-Instruct (Yang et al., 2025) for both information extraction and downstream question answering tasks.To further validate the model-agnostic robustness and efficiency of the SEEM framework, we additionally incorporate Pangu-Embedded-7B (Chen et al., 2025), a compact model with 7 billion parameters, as a secondary backbone. This allows us to observe whether the episodic memory benefits scale down effectively to smaller-parameter models. Furthermore, cross-model validation using GPT-OSS-120B (Agarwal et al., 2025) is provided in Appendix A.1. Regarding retrieval configurations, we align the hyperparameters based on the granularity of the retrieved units. For standard RAG baselines that rely on dense retrieval, we set the retrieval count k to 5, fetching the top-5 original interaction

Method	Multi-hop (Count: 282)	Temporal (Count: 321)	Open-domain (Count: 96)	Single-hop (Count: 841)	Adversarial (Count: 446)
A-MEM	29.4	39.7	15.0	37.6	78.3
HippoRAG 2	31.9	53.4	**34.7**	54.2	94.2
SEEM (Ours)	**32.3**	**54.6**	26.6	**58.2**	**96.9**

Table 2: Detailed F1 performance breakdown across five question categories on the LoCoMo benchmark. Sample counts for each category are indicated in parentheses. Best results are highlighted in **bold**.

Method	LoCoMo			LongMemEval		
	F1	EM	J	F1	EM	Acc.
Context (Relevant Only)	38.60	11.40	69.28	22.60	3.20	68.60
NV-Embed-v2 (Lee et al., 2025)	46.10	29.10	51.56	16.53	2.00	57.20
Memory-based Frameworks						
HippoRAG 2 (Gutiérrez et al., 2025)	44.60	27.90	50.96	34.82	25.10	46.59
SEEM (Ours)	**48.60**	**30.80**	**56.11**	**45.50**	**29.60**	**60.80**

Table 3: Performance comparison on LoCoMo and LongMemEval using the `Pangu-Embedded-7B` backbone. The best results are highlighted in **bold**.

messages. Similarly, for the memory-augmented baselines Mem0 and A-MEM, we retrieve the top-10 processed memory chunks. For HippoRAG 2, which operates on a session-based retrieval logic, we utilize the top-5 retrieved chunks to construct the context for final response generation. In our proposed SEEM framework, we configure the system to retrieve the top-5 relevant text chunks alongside their associated episodic memories to construct the reasoning context. To balance narrative continuity with information density, we employ a selective RPE strategy. Specifically, the total size of the final expanded evidence set $\mathcal{P}_{final}$ is restricted to at most twice the initial retrieval budget.

5 Results

5.1 Main Results

Table 1 summarizes the performance of SEEM and several baseline methods on the LoCoMo and LongMemEval benchmarks, while Table 2 provides a detailed breakdown across different question categories. Overall, experimental results indicate that SEEM yields the highest scores across most evaluation metrics, reflecting its capacity for managing long-term agentic memory.

Comparison with Dense Retrieval. As shown in the first group of Table 1, while advanced dense retrieval models such as NV-Embed-v2 exhibit competitive performance in fetching local information, they remain limited by the absence of a structured memory state. SEEM exceeds the performance of NV-Embed-v2 by 3.2% in F1 score and 3.3% in LLM-as-a-Judge (J) score on LoCoMo. This performance gap suggests that pure vector-based retrieval, although efficient, may not fully capture the intricate relational and temporal dependencies of long-term interactions. By integrating structured EEFs and relational quadruples, SEEM provides a context that is more logically grounded compared to simple embedding-based matching.

Comparison with Memory-based Frameworks. SEEM consistently outperforms the evaluated memory-based systems across both benchmarks. On LoCoMo, SEEM achieves an F1 score of 61.1 and a J score of 78.0, exceeding HippoRAG 2, by 2.8% and 1.5% respectively. It is observed that older memory frameworks yield lower performance scores, likely due to their reliance on flatter storage structures when processing extremely long interaction streams. In contrast, our hierarchical architecture facilitates an organized representation of complex event sequences. This trend is evident on LongMemEval, where SEEM achieves 65.0% accuracy, representing a 4.4% absolute improvement over HippoRAG 2.

Performance by Question Category. The categorical breakdown in Table 2 provides further insights into the framework's strengths. SEEM exhibits superior performance in four out of five categories, with notable gains in *single-hop* and *temporal* reasoning. The advantage in temporal queries suggests that the event-centric indexing within the

episodic layer effectively maintains chronological narrative flow. Furthermore, SEEM achieves a high score in the *adversarial* category, indicating that its provenance-based grounding helps distinguish factual evidence from distractors. Conversely, SEEM shows lower performance in the *open-domain* category compared to HippoRAG 2. This suggests that for queries lacking specific narrative anchors, a purely graph-based retrieval approach without episodic expansion may be more efficient.

Semantic vs. Lexical Performance. A key observation is that the performance gains of SEEM are particularly evident in the LLM-as-a-Judge (J) and LongMemEval accuracy (Acc.) metrics. These metrics prioritize semantic alignment and factual correctness over surface-level word overlap (measured by BLEU-1). The scores in these categories indicate that SEEM does not merely retrieve relevant text but also reconstructs the underlying narrative logic. This synthesis is primarily driven by the RPE mechanism, which ensures that retrieved fragments are expanded into complete event contexts to support accurate reasoning.

Efficacy on Parameter-Constrained Backbones. The results in Table 3 reveal a compelling trend: SEEM effectively narrows the performance gap between small-scale models and their larger counterparts. Even when powered by Pangu-Embedded-7B, SEEM consistently outperforms all baseline frameworks, including those utilizing more complex retrieval-augmented strategies. Specifically, the significant lead in the reasoning assessment (J) on LoCoMo suggests that SEEM's episodic memory provides higher information density and narrative coherence, which is crucial for smaller LLMs that often struggle with "distraction" in long-context windows. Furthermore, the robust accuracy on LongMemEval confirms that SEEM does not rely on massive parameter counts to maintain long-term context; instead, it provides a specialized memory scaffolding that enables a 7B model to achieve reasoning depth previously reserved for much larger architectures. This highlights SEEM as a model-agnostic, parameter-efficient solution for long-range agentic memory.

5.2 Hyperparameter Sensitivity Analysis

We analyze the impact of the initial retrieval size $|\mathcal{P}_{ret}|$ on the reasoning performance of SEEM. This parameter controls the number of seed passages retrieved from the GML before any expansion occurs.

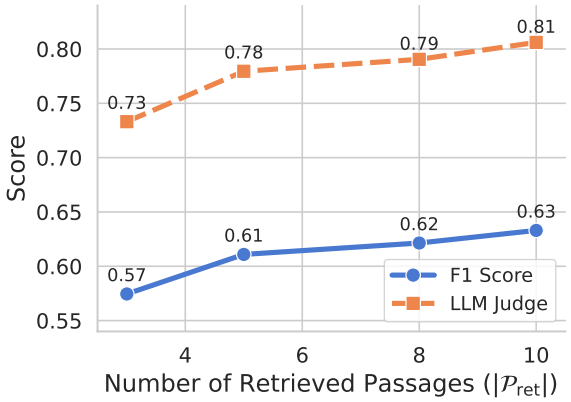

Figure 3: Impact of the initial retrieval size ($|\mathcal{P}_{ret}|$).

Configuration	LoCoMo BLEU-1	LoCoMo F1	LoCoMo J
SEEM (Full Model)	**56.1**	**61.1**	**78.0**
w/o Fact Provisioning ($\mathcal{K}_{top}$)	55.2	60.4	77.7
w/o Relational Propagation	54.5	59.6	76.3
w/o RPE	55.1	60.2	77.1
w/o EEF ($\mathcal{E}_{ret}$)	53.5	58.5	75.0

Table 4: Ablation study of key components in the SEEM framework on the LoCoMo benchmark.

Figure 3 shows the performance trends for F1 and J as $|\mathcal{P}_{ret}|$ varies from 3 to 10.

We observe a consistent improvement in both metrics as the initial retrieval window expands. Specifically, increasing $|\mathcal{P}_{ret}|$ from 3 to 10 results in a 5.9% gain in F1. Notably, SEEM does not exhibit the typical performance degradation often seen in traditional RAG systems when the context window grows. This positive correlation suggests that our framework can effectively leverage a broader range of initial evidence to refine its final answer without being overwhelmed by the additional potential noise in the retrieved passages.

5.3 Ablation Study

We conduct an ablation study to evaluate the individual contributions of the core components. We compare the full framework against four variants: (1) *w/o Fact Provisioning*, which excludes the injection of relational quadruples; (2) *w/o Relational Propagation*, which replaces the graph-based seed set expansion with direct lexical retrieval; (3) *w/o RPE*, which disables the Reverse Provenance Expansion mechanism; and (4) *w/o EEF*, which removes the structured episodic event frames.

Contribution of System Components. As shown in Table 4, the removal of any component leads to a measurable decrease across all evaluation metrics, confirming their synergy. The relational

propagation mechanism serves as the foundation for identifying relevant historical entries; its absence results in a notable decline in the LLM Judge score, as the system struggles to navigate the global graph topology to locate non-contiguous passages. The RPE mechanism plays a key role in enriching the retrieved context; its absence leads to fragmented evidence, which negatively impacts reasoning quality. The omission of fact provisioning primarily impacts the factual grounding of responses, as the LLM lacks the explicit logical constraints provided by the graph-based quadruples. Finally, the EEFs provide the necessary structure for event-centric synthesis. The omission of EEFs requires the model to rely on unstructured text, which may lead to a decrease in the coherence of the generated responses.

Architectural Robustness. Experimental results indicate that SEEM maintains a consistent performance threshold even under ablated configurations. Observationally, the core hierarchical architecture yields reasoning scores that exceed those of established baselines, even when specific modules are deactivated. This suggests that the fundamental separation of episodic and relational information provides a structurally effective foundation for managing long-term context. These findings imply that the performance gains of SEEM are derived not only from auxiliary components but also from the underlying organization of its memory layers.

The results demonstrate that while the hierarchical architecture ensures a strong performance baseline, the integration of EEFs, RPE, fact provisioning, and relational propagation is essential to achieve optimal reasoning accuracy.

5.4 Case Study

To qualitatively evaluate SEEM, we compare it against the gold standard and HippoRAG 2 on the LoCoMo (see Table 5). Our analysis focuses on three critical dimensions of agentic memory.

Multi-attribute Grounding. Unlike raw text snippets, the EEF explicitly decomposes each interaction into granular roles such as *Reason* and *Method*. This structural decomposition allows the agent to distinguish between the intent and the action, which facilitates deeper social and causal reasoning across extended interaction histories.

Narrative Synthesis. The framework achieves narrative synthesis through the *Associative Fusion* of conversational turns. By merging an inquiry and its corresponding response into a single cohesive unit, the system effectively preserves the logical flow of the interaction. This consolidation approach also significantly reduces retrieval redundancy by avoiding the storage of fragmented conversational turns.

Temporal Resolution. The frame exhibits sophisticated temporal grounding by processing reference dates alongside relative durations. For instance, by analyzing a reference date of January 23, 2022 in conjunction with a duration of three years, the system implicitly resolves the event's origin to January 2019. Such precise resolution ensures chronological consistency and factual accuracy within the EML.

In summary, SEEM ensures more grounded and logically consistent responses by transforming disparate interactions into a structured, coherent agentic memory.

6 Conclusion

We proposed SEEM, a hierarchical framework addressing scattered retrieval in long-term interactions. By integrating episodic event frames with an associative fusion mechanism, the system synthesizes coherent narratives from fragmented observations, outperforming traditional RAG and graph-based baselines. Our method effectively maintains global context and provides a scalable approach for enhancing the long-term reasoning capabilities of LLM-based agents in complex environments.

Limitations

Despite its effectiveness, the framework faces limitations regarding computational efficiency, as the heavy reliance on LLMs for extracting frames and performing associative fusion increases latency and token costs compared to standard vector retrieval. Additionally, the system is susceptible to error propagation, where inaccuracies in the initial LLM-based extraction or fusion phases can permanently corrupt the structured memory store. Finally, the reliance on predefined semantic slots within EEFs may limit the ability to capture abstract information that does not fit neatly into standard cognitive frame definitions.

Ethical Considerations

The development of SEEM introduces considerations regarding the management and persistence

of long-term interaction data. Unlike standard retrieval augmented generation which primarily accesses external corpora , SEEM transforms interaction streams into persistent episodic event frames and relational quadruples. While our experiments are conducted on publicly available benchmarks , real-world deployment of such a memory framework involves the retention of user information over extended periods. It is essential that future applications implement data anonymization protocols and provide users with explicit control over their stored interaction histories, including the right to modify or delete specific memory frames.

The framework is also subject to algorithmic bias and safety. Since SEEM relies on large language models for both episodic frame extraction and final response generation , it may inherit or amplify social biases present in these underlying models. The structured nature of event frames could potentially solidify these biases within the agent's long-term memory, leading to biased reasoning in subsequent interactions. We recommend that developers implement content filtering and auditing mechanisms during the memory consolidation phase to mitigate these risks.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62406088, 62422603) and the Guangdong Basic and Applied Basic Research Foundation (Grant No. 2025A1515011376, 2024B0101050003).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Hanting Chen, Yasheng Wang, Kai Han, Dong Li, Lin Li, Zhenni Bi, Jinpeng Li, Haoyu Wang, Fei Mi, Mingjian Zhu, Bin Wang, Kaikai Song, Yifei Fu, Xu He, Yu Luo, Chong Zhu, Quan He, Xueyu Wu, Wei He, and 5 others. 2025. Pangu embedded: An efficient dual-system llm reasoner with metacognition. *Preprint*, arXiv:2505.22375.

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*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2024. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Charles J Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences*, 280(1):20–32.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1).

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. In *Forty-second International Conference on Machine Learning*.

Taher H Haveliwala. 2002. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick SH Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. NV-embed: Improved techniques for training LLMs as generalist embedding models. In *The Thirteenth International Conference on Learning Representations*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen,

and 244 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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.

M Minsky. 1975. A framework for representing knowledge. *The psychology of computer vision*.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. 2024. RAPTOR: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*.

Endel Tulving and 1 others. 1972. Episodic and semantic memory. *Organization of memory*, 1(381-403):1.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025a. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 2025b. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of rag noise in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5019–5039.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, and 1 others. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

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*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Xinping Zhao, Xinshuo Hu, Zifei Shan, Shouzheng Huang, Yao Zhou, Xin Zhang, Zetian Sun, Zhenyu Liu, Dongfang Li, Xinyuan Wei, Youcheng Pan, Yang Xiang, Meishan Zhang, Haofen Wang, Jun Yu, Baotian Hu, and Min Zhang. 2025. Kalm-embedding-v2: Superior training techniques and data inspire a versatile embedding model. *Preprint*, arXiv:2506.20923.

A Supplemental Experimental Results

A.1 Cross-Model Generalization and Architectural Robustness

To evaluate whether the performance gains of SEEM are model-dependent or derive from its underlying architecture, we conduct supplemental experiments on the LoCoMo benchmark by replacing the primary Qwen3-Next-80B-A3B-Instruct backbone with **GPT-OSS-120B**. This cross-model validation serves as a controlled comparison, ensuring that the observed improvements are attributable to our hierarchical memory mechanisms rather than the inherent capabilities of a specific LLM.

Analysis of Results. As demonstrated in Table 6, SEEM maintains its performance leadership when integrated with the GPT-OSS-120B backbone, mirroring the trends observed with the Qwen3-Next-80B-A3B-Instruct model. The consistent performance gains across these distinct large language models reinforce the conclusion that the advantages of our hierarchical episodic architecture are model-agnostic. By decoupling the memory organization mechanism from the specific underlying LLM, SEEM demonstrates robust generalization capabilities in narrative consistency and retrieval precision. These results confirm that the framework serves as a versatile enhancement for various long-context reasoning agents regardless of their specific architectural implementations.

A.2 Granular Category-wise Evaluation

To further investigate the performance characteristics of SEEM across diverse reasoning challenges, we present a granular analysis of the results on the **LoCoMo** and **LongMemEval** benchmarks, categorized by specific task dimensions.

Analysis of LoCoMo Categories. As illustrated in Table 7, SEEM achieves superior performance across four out of five reasoning categories. The framework demonstrates significant advantages in Temporal and Multi-hop reasoning, outperforming competitive baselines by a notable margin. These results suggest that the structured EEFs effectively capture chronological dependencies that are often overlooked by dense retrieval or static graph-based approaches. While HippoRAG 2 maintains competitive performance in Open-domain queries due to its focus on static entity indexing, SEEM prioritizes the reconstruction of complex narrative chains. This architectural focus is further evidenced by SEEM's higher resilience to adversarial distractors, indicating lower vulnerability to hallucinations compared to traditional retrieval-based systems.

Analysis of LongMemEval Categories. The evaluation encompasses six distinct reasoning categories: Speaker-Specific (S-S) tasks focused on the user, assistant, or preferences; Multi-Session (Multi-S) interaction; Temporal reasoning; and Knowledge Update (K-Update). This comprehensive assessment further reinforces the efficacy of the SEEM architecture. As shown in Table 8, SEEM achieves the highest average accuracy, driven primarily by its strong performance in the Knowledge Update and Temporal reasoning categories. The framework's capacity to resolve user-specific information highlights its effectiveness in grounding queries to the appropriate episodic context. While certain baselines demonstrate specialized strengths in preference-based retrieval, SEEM provides a more balanced performance profile. This equilibrium is achieved by bridging high-level semantic abstractions with the granular requirements of long-term interaction history, ensuring consistent reasoning across diverse and evolving query types.

A.3 Evaluation of Incremental Memory Construction

To assess the practical applicability of SEEM in streaming interaction scenarios, we conduct an evaluation under an incremental construction setting. In this configuration, the complete sequence of interaction passages is partitioned into four chronological segments, which are processed by the memory system sequentially rather than in a single batch.

The results, summarized in Table 9, demonstrate that SEEM maintains highly stable performance across all evaluation metrics. The marginal discrepancy between the batch and incremental modes suggests that the associative fusion mechanism effectively preserves narrative coherence and structural integrity, even when information is presented in fragments. This minimal performance trade-off indicates the framework's robustness for real-world deployment, where memory evolves continuously in response to sequential updates without significant loss in reasoning integrity.

Query	Gold Answer	HippoRAG 2	SEEM (Ours)
Q1: What book did Melanie read from Caroline's suggestion? (Multi-hop)	"Becoming Nicole"	The book's title is not specified.	"Becoming Nicole" by Amy Ellis Nutt
Q2: How did John describe his kids' reaction at the military memorial? (Single-hop)	Awestruck and humbled.	John said the experience made an impact on his kids, but did not describe their specific reaction.	They were awestruck and humbled.
Q3: What day did Tim get into his study abroad program? (Temporal)	January 5, 2024	January 7, 2024	January 5, 2024

Table 5: Case study comparison between the gold answer and different memory frameworks.

Method	LoCoMo		
	BLEU-1	F1	*J*
Dense Retrieval			
KaLM-Embedding-V2.5 (Zhao et al., 2025)	38.7	42.8	63.2
NV-Embed-v2 (Lee et al., 2025)	44.1	49.2	75.5
Memory-based Frameworks			
A-MEM (Xu et al., 2025)	42.4	47.3	63.0
HippoRAG 2 (Gutiérrez et al., 2025)	44.6	50.2	73.6
SEEM (Ours)	**50.7**	**55.7**	**77.1**
Backbone LLM: GPT-OSS-120B			

Table 6: Performance comparison on LoCoMo based on GPT-OSS-120B. The best results are highlighted in **bold**.

B Analysis

B.1 Structural Analysis of the Graph Memory Layer

The GML provides the static factual foundation of the SEEM framework, complementing the dynamic nature of the EML. As summarized in Table 10, the structural statistics across various narrative partitions reflect a high density of relational knowledge and entity connectivity.

The internal composition of the graph highlights two critical capabilities of the system. The prevalence of temporal anchors indicates that a vast majority of the extracted facts are grounded in specific temporal contexts, which is essential for resolving chronological dependencies in long-term reasoning. This structural density ensures that the GML can serve as a reliable foundation for relational propagation, providing the necessary factual context for hybrid retrieval.

B.2 Qualitative Analysis of Episodic Event Frames

Figure 4 provides a representative instance of a consolidated EEF, illustrating the framework's capacity for high-fidelity narrative synthesis. Several key advantages of the SEEM architecture are evident in this structured representation:

Multi-attribute Grounding. Unlike raw text snippets, the EEF explicitly decomposes the interaction into fine-grained roles such as Reason and Method. This decomposition allows the agent to distinguish between intent and action, which facilitates deeper social and causal reasoning across extended interaction histories.

Narrative Synthesis. The SEEM framework achieves narrative synthesis through the associative fusion of interaction pairs. By merging a conversational inquiry and its corresponding response into a single, cohesive episodic unit, the system preserves the logical continuity of the dialogue. This consolidation mechanism effectively captures the functional relationship between speaker turns while significantly reducing retrieval redundancy in the memory store.

Temporal Resolution. The EEF exhibits sophisticated temporal grounding by processing the reference date alongside the relative duration. For instance, the system implicitly resolves an event's origin to "January 2019" by analyzing the reference date in conjunction with a three-year duration.

Method	Multi-hop (Count: 282)		Temporal (Count: 321)		Open-domain (Count: 96)		Single-hop (Count: 841)		Adversarial (Count: 446)	
	Correct	Acc.	Correct	Acc.	Correct	Acc.	Correct	Acc.	Correct	Acc.
A-MEM	154	54.61%	90	28.04%	45	46.88%	496	58.98%	430	96.41%
HippoRAG 2	173	61.35%	203	63.24%	**59**	**61.46%**	659	78.36%	416	93.27%
NV-Embed-v2	148	52.48%	205	63.86%	56	58.33%	647	76.93%	417	93.50%
SEEM (Ours)	**177**	**62.77%**	**219**	**68.22%**	52	54.17%	**668**	**79.43%**	**432**	**96.86%**

Table 7: Category-specific performance on the LoCoMo dataset. Sample counts for each reasoning category are provided in parentheses. The best results are highlighted in **bold**.

Method	S-S (User) (Count: 70)	S-S (Asst.) (Count: 56)	S-S (Pref.) (Count: 30)	Multi-S (Count: 133)	Temporal (Count: 133)	K-Update (Count: 78)	Mean
HippoRAG 2	82.86	**94.64**	20.00	**58.65**	48.12	56.41	60.11
NV-Embed-v2	80.00	**94.64**	**33.33**	48.12	43.61	65.38	60.85
SEEM (Ours)	**91.43**	**94.64**	30.00	54.89	**53.38**	**70.51**	**65.81**

Table 8: Detailed performance comparison on the LongMemEval benchmark. Accuracy (%) is reported across six reasoning categories, with sample counts for each category provided in parentheses. The best results are highlighted in **bold**.

Method	BLEU-1	F1	J
SEEM (Batch)	**56.1**	**61.1**	**78.0**
SEEM (Incremental)	55.6	60.6	77.6

Table 9: Comparison between Batch and Incremental Memory Construction in SEEM.

This precise resolution ensures chronological consistency and factual integrity within the EML.

By transforming ambiguous pronouns into structured attributes while maintaining strict textual grounding via provenance pointers, the EEF provides a high-density semantic anchor. This structured representation ensures that retrieved context is not only chronologically accurate but also logically complete for downstream reasoning.

B.3 Analysis of Associative Fusion

We evaluate the structural impact of the associative fusion mechanism by analyzing the distribution of consolidated frames relative to the original interaction turns. As demonstrated in Table 11, SEEM reduces the total number of memory units by synthesizing fragmented turns into unified episodic frames. This consolidation mitigates semantic redundancy and improves retrieval density by grouping chronologically and logically linked interactions. The presence of multi-turn fusions indicates that the framework can bridge narrative sequences, transforming discrete conversational segments into more compact semantic representations. This structural efficiency ensures a logically continuous memory state, which is essential for maintaining context during long-horizon agentic reasoning.

Passages per Memory	Number of Memory Frames
1	371
2	79
3	20
4	3
5	4
8	1
Total Memory Frames	**478**
Total Passages	**629**
Consolidation Ratio	**1.32:1**

Table 11: Distribution of consolidated episodic memory frames across constituent interaction passages in a LoCoMo narrative partition.

B.4 Redundancy Analysis of Dual-Layer Retrieval

To verify the necessity of the dual-layer architecture, we analyze the global distribution of semantic redundancy between the GML and the EML. For each query in the LoCoMo dataset, we retrieve the corresponding structural quadruples from the GML and EEFs from the EML. We apply an LLM-based filter to the GML outputs to ensure precision, resulting in 1,282 valid retrieval pairs from the original 1,986 queries. The aggregate semantic overlap is quantified by computing the cosine similarity between their respective embeddings, with the overall distribution detailed in Table 12.

Metric	h_1	h_2	h_3	h_4	h_5	h_6	h_7	h_8	h_9	h_{10}	Average
Entities	1,242	902	1,845	1,486	1,820	1,692	1,745	1,665	1,286	1,575	**1,525.8**
Facts	1,749	1,320	2,534	2,194	2,673	2,699	2,557	2,395	1,868	2,348	**2,233.7**
Temporal Anchors	1,557	1,213	2,294	1,948	2,363	2,385	2,258	2,070	1,694	2,056	**1,983.8**
Synonymy Edges	11,732	5,439	19,963	14,178	16,433	14,344	15,904	15,402	10,670	12,459	**13,652.4**

Table 10: Structural statistics of the GML across 10 Narrative Partitions (h_1–h_{10}) in the LoCoMo dataset. The metrics quantify the internal density of the GML, representing the static knowledge foundation of the SEEM framework.

Example: Consolidated Episodic Memory Frame

Summary: On January 23, 2022, at 2:01 pm, Joanna asked Nate how long he had had "them," prompting Nate to respond that he had owned them for three years—since approximately January 2019—and that they brought him significant joy.

Events (Structured EEF):

Event 1:

Participants	Joanna
Action	Joanna asked how long Nate had had "them"
Time	2:01 pm on 23 January, 2022
Reason	Expressing affectionate curiosity about an object Nate possesses
Method	Through verbal inquiry

Event 2:

Participants	Nate
Action	Nate stated he had owned "them" for three years, and that they brought him tons of joy
Time	From approx. January 2019 to January 23, 2022
Reason	Responding to Joanna's question about the duration of ownership
Method	Through verbal response

Figure 4: An illustrative example of a consolidated Episodic Event Frame (EEF) in the SEEM framework. This structured representation demonstrates how the associative fusion mechanism synthesizes multi-turn interactions into coherent, attribute-rich episodic units.

Similarity Range	Count	Prop. (%)
$[0.25, 0.30)$	1	0.08
$[0.30, 0.35)$	12	0.94
$[0.35, 0.40)$	106	8.27
$[0.40, 0.45)$	398	31.05
$[0.45, 0.50)$	497	38.77
$[0.50, 0.55)$	224	17.47
$[0.55, 0.60)$	40	3.12
$[0.60, 0.65)$	4	0.31
Total Valid Pairs		**1282**
Mean Similarity		**0.46**

Table 12: Distribution of cosine similarity between retrieved quadruples (GML) and EEFs (EML) on the LoCoMo dataset.

The mean similarity of 0.46 suggests that the GML and EML capture complementary semantic dimensions. This divergence confirms that the structural extraction and narrative synthesis capture distinct information even when grounded in the same interaction context, justifying the use of a dual-layer architecture.

C Prompt Templates and Agent Instructions

In this section, we provide the detailed prompt templates used in the SEEM framework. These prompts are designed to implement the formal functions defined in Section 3, specifically the extraction function $\mathcal{F}_{ext}$, the consolidation function $\mathcal{F}_{fuse}$, and the final generation function G.

Prompt 1: Episodic Event Frame Extraction ($\mathcal{F}_{\text{ext}}$)

You are an expert at extracting episodic memories from conversation turns. Your task is to analyze a single conversation turn (which may contain time information, speaker information, and text content), identify distinct events mentioned in the turn, and extract structured event details for each event.

The input format is a single conversation turn that may include:

- Time information
- Image descriptions in the format: `[Image: <description>]`
- The actual text content of the turn

For each event, extract the following Structured Event Attributes (use `null` if not reliably determined):

- **Participants:** List of actors. Replace pronouns with specific names; include full names/roles.
- **Action:** List of substantive actions. *CRITICAL*: Each action MUST include the subject/actor. Format: "Subject verb object".
- **Time:** Time, date, or duration. For continuous events, explicitly specify the range. Specify the event's actual time, not just the conversation time.
- **Location/Reason/Method:** The venue, purpose, or means if explicitly stated.

Guidelines: 1) **Event Definition:** Define an "event" as an occurrence with a clear subject, action, and temporal context. 2) **Coreference Resolution:** Resolve pronouns to specific entities. 3) **No Redundancy:** Do not extract the act of "speaking" as a separate event. 4) **Output strict JSON.** No markdown formatting.

Output format (strict JSON):

```
{
  "summary": "1-3 concise sentences...",
  "events": [{ "participants": [], "action": [], "time": "", ... }]
}
```

Figure 5: The structured prompt for Episodic Event Frame Extraction ($\mathcal{F}_{ext}$). This initial stage of the SEEM pipeline converts unstructured interaction logs into discrete, attribute-rich event units, providing the grounded anchors necessary for long-term temporal and multi-hop reasoning.

Prompt 2: Associative Consolidation and Fusion ($\mathcal{F}_{\text{fuse}}$)

You are an expert at integrating episodic memories. Your task is to combine two episodic memories into a single, comprehensive memory that preserves all important information from both memories.

IMPORTANT: Event Integration Strategy

- **LESS:** When events from both memories describe the same occurrence (merge).
- **EQUAL:** When all events are distinct and unrelated.
- **GREATER:** When integration reveals new event relationships.

Guidelines:

1. **Conservative Merge:** Only merge events that clearly describe the EXACT SAME occurrence. If events are part of a sequence (Plan $\rightarrow$ Execute), DO NOT merge them.
2. **Entity Alignment:** Unify participant names.
3. **Conflict Resolution:** Highest Priority: Evidence found in "Original Passages". Use these original sources as the primary reference to cross-verify all episodic attributes and mitigate error propagation.
4. **Summary Synthesis:** Do NOT simply concatenate. Rewrite a single, cohesive narrative describing progression or causal relationships.

CRITICAL: Time Information Handling Rules

1. **Same Event, Different Time:** If one is more specific, prefer it; if complementary, combine into a range.
2. **Sequential Events:** Events at different times MUST be kept separate.
3. **Temporal Ordering:** Arrange events in chronological order.

Output Format: Strict JSON ONLY.

Figure 6: The structured prompt for associative consolidation and fusion, designed to synthesize fragmented interaction logs into coherent Episodic Event Frames (EEFs).

Prompt 3: Memory-Augmented Question Answering (G)

You are a reading-comprehension QA assistant operating in episodic-memory mode.

Each query provides:

(A) an "Original Passages (Grounded Evidence)" section (most trusted evidence);
(B) an "Episodic Memory Summary" section (high-signal reference distilled from retrieved passages; use as a guide and as supplemental evidence when not contradicted);
(C) optionally, a "Relevant Facts" section (high-signal reference quadruples; use to locate/verify key entities/relations and as supplemental evidence when not contradicted).

Evidence Policy:

1. You MUST read (B) and (C) (if present). Treat them as high-signal hints to guide what to look for in (A).
2. If (A) is incomplete, you MAY answer using explicit (B)/(C) as supplemental references ONLY if they do not contradict (A).
3. If (B)/(C) conflicts with (A), trust (A) and ignore the conflicting parts of (B)/(C).

Output Format:
Thought: <reasoning>
Answer: <answer only>
(Constraint: The answer must be concise, definitive, and devoid of additional elaborations)

Figure 7: The Inference Prompt for SEEM's Memory-Augmented Question Answering. By providing the model with distilled episodic summaries and graph-based facts alongside raw evidence, the system effectively mitigates the "scattered retrieval" problem in long-context interactions.