H-Mem: Hybrid Multi-Dimensional Memory Management for Long-Context Conversational Agents

Zihe Ye[1], Jingyuan Huang[1], Weixin Chen[1,2], Yongfeng Zhang[1]
[1]Rutgers University
[2]Hong Kong Baptist University
{zihe.ye, chy.huang, weixin.chen, yongfeng.zhang}@rutgers.edu

Abstract

Long-context conversational agents require robust memory, but existing frameworks struggle to organize information effectively across dimensions like time and topic, leading to poor retrieval. To address this, we introduce H-Mem, a novel Hybrid Multi-Dimensional Memory architecture. H-Mem stores conversational facts in two parallel, hierarchical data structures: a temporal tree that organizes information chronologically and a semantic tree that organizes it conceptually. This dual-tree design enables a hybrid retrieval mechanism managed by an intelligent Mode Controller. Based on the query, the controller dynamically chooses between a sequential search using semantic anchors and an intersective search combining both hierarchies. Our experiments on long-context QA datasets demonstrate that H-Mem provides a more flexible approach to memory management, leading to significant improvements of over 8.4% compared to other state-of-the-art systems. Our code is available at: https://github.com/rutgerswiselab/H-mem.

1 Introduction

The proliferation of Large Language Models (LLMs) has catalyzed a new frontier in agentic artificial intelligence, where systems are expected to maintain context, recall past interactions, and reason over long periods. A robust and scalable memory mechanism is the cornerstone of such advanced agents (Achiam et al., 2023; OpenAI, 2024; LangChain, 2025; OpenAI, 2024; Wu et al., 2025b; Du et al., 2025). However, existing memory systems, including retrieval-augmented generation (RAG) frameworks leveraging vector databases (Lewis et al., 2020; Karpukhin et al., 2020; Qian et al., 2025; Gutiérrez et al., 2024) and early hierarchical models (Rezazadeh et al., 2025; He et al., 2024; Sun and Zeng, 2025), often operate on a single dimension. They typically retrieve information based solely on semantic similarity or, less commonly, chronological proximity (Ong et al., 2025; Maharana et al., 2024; Wang et al., 2025), failing to capture the multifaceted, associative nature of human memory (Squire et al., 2015; Anderson, 1983). While some recent works have explored hierarchical memory structures (Rezazadeh et al., 2025; Pan et al., 2025; Ong et al., 2025) or memory organization inspired by operating systems (Li et al., 2025; Kang et al., 2025; Packer et al., 2023), a critical gap remains: no existing architecture cohesively integrates and queries memory across both temporal and semantic dimensions in a structured, hierarchical manner (Jiang et al., 2024; Zhang et al., 2025a; Xu et al., 2025a; Chen et al., 2026).

To address this limitation, we introduce a Hybrid Multi-Dimensional Memory (H-Mem) system designed to enhance long-context conversational question-answering performance. The core of our contribution is a novel dual-tree architecture that simultaneously organizes atomic pieces of information, or "Fact Nodes," into two parallel hierarchical structures: a Time-Dimension Tree that indexes facts chronologically and a Semantic-Dimension Tree that organizes the same facts based on conceptual relatedness. This dual representation allows for more flexible and robust information retrieval. Our primary contributions can be listed as follows:

- We propose the novel dual-tree architecture for parallel temporal and semantic memory organization.
- We introduce a dynamic retrieval mechanism managed by an LLM-based mode controller, which intelligently switches between two distinct search strategies, which are a sequential, context-expanding search and a parallel, intersective search, based on the nature of the user's query.
- We demonstrate that this multi-dimensional

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 7756–7775

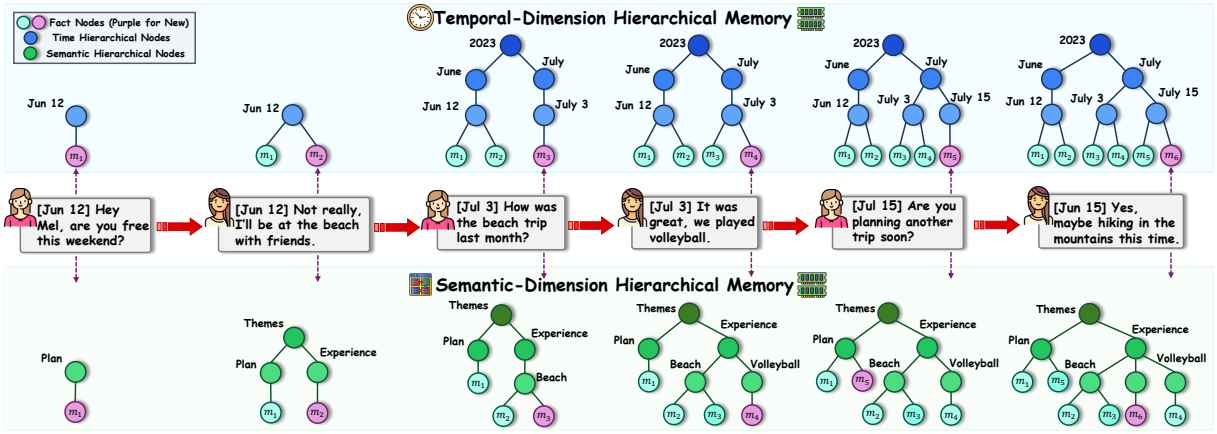

Figure 1: The H-Mem dual-tree construction process. Each new FactNode (m_6) is simultaneously indexed in both the Temporal Tree (by timestamp) and the Semantic Tree (by conceptual topic), creating a multi-dimensional memory structure.

approach, which more closely mimics the associative and hierarchical properties of human memory, is a promising direction for building more capable agentic LLMs.

In this paper, we detail the architecture of the H-Mem system, its ingestion and retrieval pipelines, and set the stage for its evaluation on long-context conversational QA benchmarks (Maharana et al., 2024; Wu et al., 2025a; Kamradt, 2023; Li et al., 2023; Bianchi et al., 2025).

2 Related Works

Long-Context Modeling for Dialogue. Transformer models can process long sequences, but their attention mechanism imposes a fixed context window limit (Vaswani et al., 2017; Devlin et al., 2019). Even with extended windows, models often fail to exploit dialogue structure effectively, leading to degraded reasoning and coherence (Li et al., 2023; He et al., 2024; Tan et al., 2025). In multi-session dialogue, models must maintain both character consistency and episode continuity (Rasmussen et al., 2025; Li et al., 2025), which motivates memory-based architectures (Zhong et al., 2024; Packer et al., 2023). Thus, relying solely on Transformers for long-context dialogue remains inadequate. When key information is distant or scattered across turns, models often fail to retrieve and integrate it coherently (Liu et al., 2023; Maharana et al., 2024; Wu et al., 2025a).

To address this, Retrieval-Augmented Generation (RAG) augments models with external knowledge, improving dialogue quality (Lewis et al., 2020). Recent extensions further adapt RAG to long-context settings (Qian et al., 2025; Gutiérrez et al., 2024; Xu et al., 2025a). However, RAG is primarily designed for large-scale document retrieval, often returning semantically related but fragmented snippets (Zhang et al., 2025b), which is insufficient to maintain the conversation with information forgetting (Zhong et al., 2024; Das et al., 2024) and updating (Wang et al., 2025; Xu et al., 2025b; Yan et al., 2025) over time.

From Dynamic RAG to Agentic Memory. Critical evidence is often obscured by irrelevant content, and fragmented knowledge makes retrieval noisy and incoherent (Liu et al., 2023; Bianchi et al., 2025; Kamradt, 2023). Neither enlarging context windows nor enhancing retrieval is sufficient, as models still suffer from incomplete reasoning and disrupted dialogue coherence (Xu et al., 2021; Maharana et al., 2024; Wu et al., 2025a). This limitation has led to dynamic RAG (Su et al., 2025; Jeong et al., 2024; Sarthi et al., 2024), which extends static retrieval by continuously updating and reorganizing information during interactions. In this view, retrieval evolves toward memory: past information is dynamically compressed (Yu et al., 2023; Chung et al., 2024), rewritten (Xu et al., 2025b; Chhikara et al., 2025), and prioritized for future use (Yan et al., 2025; Xu et al., 2025a).

Recent work further positions memory as an independent research domain. For single agents, memory architectures enhance long-term reasoning (Packer et al., 2023; Zhong et al., 2024). In multi-agent settings, shared memory supports state persistence and coordination (Zhang et al., 2025a; Hong et al., 2024; Yuen et al., 2025). Other approaches organize knowledge into structured rep-

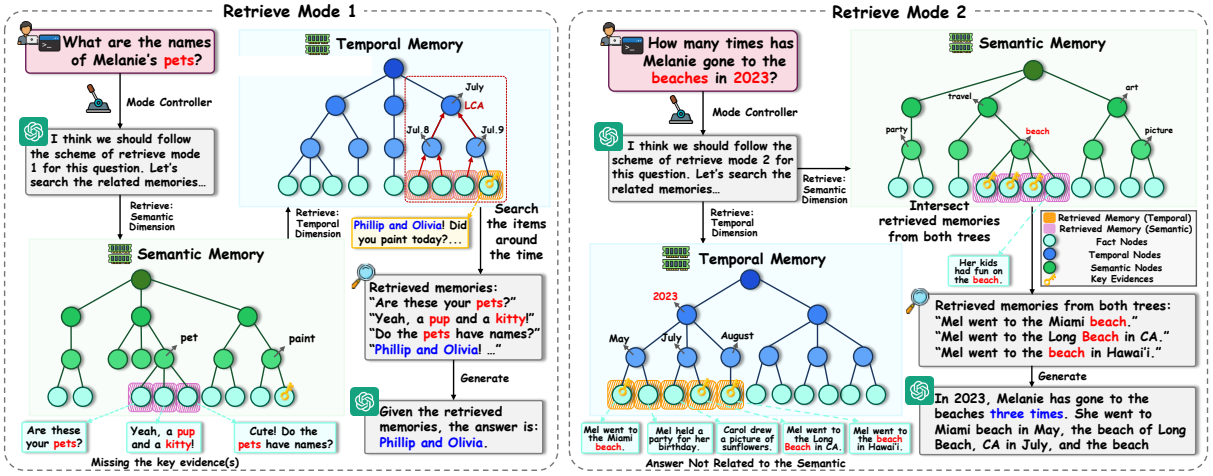

Figure 2: The two retrieval modes of H-Mem. **(Left) Mode 1 (Sequential Search)**: A semantic search for a topic is followed by a localized temporal search to find specific details. **(Right) Mode 2 (Intersective Search)**: Parallel searches on temporal and semantic dimensions are performed, and the context is derived from their intersection.

resentations, such as graphs, to capture entities, relations, and evolving context (Xu et al., 2025b; Rasmussen et al., 2025; Rezazadeh et al., 2025). Together, these directions suggest that memory extends beyond retrieval, forming a core infrastructure for building agentic systems (Huang et al., 2026; Guo et al., 2025). Yet unlike most current designs, human memory is inherently hierarchical, spanning from fine-grained events to abstract schemas (Squire and Alvarez, 1995; Van Kesteren et al., 2012; Squire et al., 2015; Janik, 2023).

Recent studies have attempted to model such hierarchy for LLM agents. Some draw inspiration from hippocampal processing (Gutiérrez et al., 2024; Das et al., 2024), some organize memory into semantic granularities but remain coarse in evidence localization (Xu et al., 2025a; Rasmussen et al., 2025; Sun and Zeng, 2025; Rezazadeh et al., 2025), others distinguish short- from long-term memory, though cross-session recall remains weak (Li et al., 2025; Ong et al., 2025; Pan et al., 2025). Despite these advances, current approaches fall short of capturing multi-dimensional hierarchy and of integrating different levels into a unified agent memory system.

3 Methodology

The H-Mem architecture is designed to address the challenges of long-context memory management by organizing information across two distinct dimensions: time and semantics. Our methodology is divided into two primary pipelines: (1) a Memory Ingestion and Formation pipeline that processes and stores conversational data, and (2) a Hybrid Retrieval pipeline that dynamically selects the optimal strategy to recall information.

3.1 Memory Ingestion and Formation

The ingestion pipeline is responsible for converting unstructured conversational turns into a structured, dual-hierarchy memory. This process involves three key steps for each turn.

3.1.1 Memory Curation

Not all parts of a conversation are equally important. To prevent the memory from being cluttered with conversational filler, each turn is first processed by a Memory Curation Agent (C_{agent}) powered by a large language model (LLM). This agent makes a binary decision: it classifies the turn as either "memorable" (containing specific facts, events, plans, or opinions) or “not_memorable” (containing simple greetings, acknowledgements, or filler). Only "memorable" turns proceed to the next stage, ensuring that the memory stores high-signal information. The turn j is then packed with prompt template $\mathcal{P}_1$ and passed to C_{agent} to get the decision on importance.

$$Decision_j = C_{agent}(turn_j|\mathcal{P}_1) \quad (1)$$

3.1.2 Dual-Tree Construction

Each memorable turn is stored as a single FactNode, m, a data object containing its timestamp, speaker, original source text, and both parents on dual-tree structures.

$$m = (t, speaker, text, parent_t, parent_s) \quad (2)$$

This FactNode becomes a leaf node in two parallel hierarchical trees: the Temporal Tree and the Semantic Tree, as illustrated in Figure 1.

1. The Temporal Tree: This tree provides a deterministic chronological index. A FactNode m is placed into this hierarchy based on its timestamp (for example, 2023 -> July -> 08 -> FactNode). This structure is created algorithmically by parsing the date and traversing the tree, creating new date nodes as needed.

$$P_t(m) = Path(Y(t), M(t), D(t)) \quad (3)$$

2. The Semantic Tree: This tree organizes facts conceptually. A FactNode m is placed via top-down search guided by vector similarity on embeddings. For a parent node N_p with child category nodes, the traversal proceeds to the child with the highest cosine similarity

$$\mathcal{N}_{next} = \text{argmax}\{\text{sim}(\vec{e}_m, \vec{e}_{c_i})\} \quad (4)$$

over all child nodes c_i.

The cosine similarity is calculated as:

$$sim(a, b) = \frac{a \cdot b}{||a|| * ||b||} \quad (5)$$

If the maximum similarity is below a threshold θ, we set $\theta = 0.8$ for the rest of our paper. A new conceptual node is created and named by an LLM, allowing the memory's ontology to grow dynamically.

3.1.3 Vector Indexing

To enable efficient global similarity search, the embedding for every ingested FactNode is also added to a FAISS index. This index I maintains a flat, searchable vector store of all memories, decoupled from the hierarchical structures, which is critical for the first step of our sequential retrieval mode:

$$I = \{E(m_i) \mid \forall m_i \in Memory\} \quad (6)$$

3.2 Hybrid Retrieval Mechanism

To answer a user's query, q, H-Mem employs a two-mode retrieval system managed by an LLM-based Mode Controller, as shown in Figure 3.

3.2.1 Mode Controller

The Mode Controller analyzes the user's query to determine the most efficient retrieval strategy, distinguishing between queries with (Mode 2) and without (Mode 1) absolute timeframes as hints. The two retrieval modes are visualized in Figure 2.

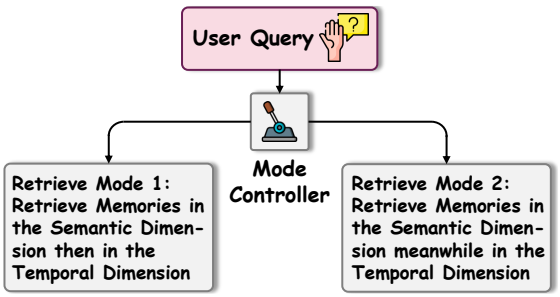

Figure 3: The H-Mem Mode Controller. The controller analyzes the user query to determine the optimal retrieval strategy, dynamically switching between Mode 1 and Mode 2 for different tasks.

3.2.2 Mode 1: Sequential Search

This mode is designed for detail-oriented and context-dependent questions. It follows a two-step process:

1. Global Vector Search: The query q is embedded and searched against the FAISS index I to retrieve the top-k most semantically relevant FactNodes from anywhere in the conversation, which serve as a set of high-quality "seed facts" M_{seed}:

$$M_{seed} = TopK_{similar}(E(q), I) \quad (7)$$

2. Temporal Expansion: The system then uses the temporal_parent pointers of these seed facts, $\{p_1, \ldots, p_k\}$ to jump to their locations in the temporal tree. It calculates the Lowest Common Ancestor (LCA) of these temporal nodes, defining the smallest relevant time window N_{ctx}:

$$N_{ctx} = LCA(p_1, p_2, \ldots, p_k) \quad (8)$$

The final set of retrieved facts, M_{ret}, comprises all facts within this temporal context window:

$$M_{ret} = FactsUnder(N_{ctx}) \quad (9)$$

3.2.3 Mode 2: Intersective Search

This mode is designed for filtering and summarization queries. It also follows a two-step process:

1. Entity-Driven Parallel Search: An LLM extracts the key topic and timeframe entities from the query. The system then performs two independent searches: a semantic similarity search in the semantic tree to find all facts related to the topic, and a robust, F1-score based search in the temporal tree to find all facts related to the timeframe.

2. Intersection: The final set of retrieved facts is the intersection of the results from the semantic and temporal searches.

3.3 Answer Generation

Finally, the list of facts retrieved from either Mode 1 or Mode 2 is compiled into a context block, including speaker information. This context, along with the original query, is passed to a final LLM with prepared prompt template for direct and concise answer synthesis.

4 Experiments

To evaluate the effectiveness of H-Mem, we conduct a comprehensive set of experiments on a long-context conversational question-answering benchmark. Our evaluation is designed to assess the model's ability to accurately recall information over long dialogues.

4.1 Dataset

We evaluate H-Mem on the LoCoMo dataset, a challenging public benchmark designed to test long-context memory in conversational agents. LoCoMo (Maharana et al., 2024) is a multi-session dialogue dataset featuring conversations between two human users. It includes a variety of question types that require recalling specific details, events across different sessions, making it an ideal testbed for our memory architecture.

We pick the following strong baselines over other solutions, including RAG and other hierarchical memory systems, as these picked baselines are specifically designed and tailored for the best performance in a conversational memory scenario and suited for LoCoMo-like long multi-session multi-turn conversations benchmarking.

4.2 Baselines

We pick the following strong baselines over other solutions, including RAG and other hierarchical memory systems, as these picked baselines are specifically designed and tailored for the best performance in a conversational memory scenario and suited for LoCoMo-like long multi-session multi-turn conversations benchmarking.

- **Mem0** (Chhikara et al., 2025): An intelligent memory system that uses a two-phase pipeline to extract, consolidate, and retrieve salient facts from conversations.
- **Mem0^g** (Chhikara et al., 2025): An enhanced variant of Mem0 that incorporates a graph-based store to capture the relationships between entities in dialogue turn using LLM.
- **Zep** (Rasmussen et al., 2025): A memory system that uses a temporally-aware knowledge graph to synthesize conversational data and maintain historical relationships.
- **OpenAI Memory** (OpenAI, 2024): The native memory feature provided by OpenAI, designed to retain information across interactions and sessions within its ecosystem.
- **LangMem** (LangChain, 2025): A summarization-centric memory framework that provides tools to manage episodic, and procedural memory for AI agents.

4.3 Evaluation Metrics

Our evaluation is conducted using both standard NLP metrics and a qualitative, LLM-based judgment score for robust semantic comparison.

- **F1 Score**: Measures the harmonic mean of precision and recall on a token level, providing an assessment of content overlap between the generated and ground-truth answers.
- **BLEU-1 (B1)**: Calculates unigram overlap to measure the fluency and adequacy of the generated answer compared to the reference.
- **J-Score (Judge Score)**: To capture semantic correctness beyond lexical overlap, we employ an LLM-as-a-Judge. A separate LLM is prompted with the question, the gold answer, and the generated answer, and is tasked with providing a binary label of TRUE or FALSE based on semantic equivalence. The J-Score is the percentage of TRUE labels.

4.4 Implementation Details

The H-Mem system is implemented in Python. The Memory Curation Agent and the final Answer Generation Agent are powered by Azure's gpt-4o-mini deployment. LLM is also used to polish writing. For all embedding tasks, we utilize BAAI/bge-large-en-v1.5 model, which has an embedding dimension of 1024. We use GPT-4o-mini (OpenAI, 2024) for our experiments. The vector index for our Global Vector Search is managed by FAISS using an IndexFlatL2 configuration. For our ablation studies, we will test variations of the retrieval mechanism, including disabling the LCA-based temporal expansion.

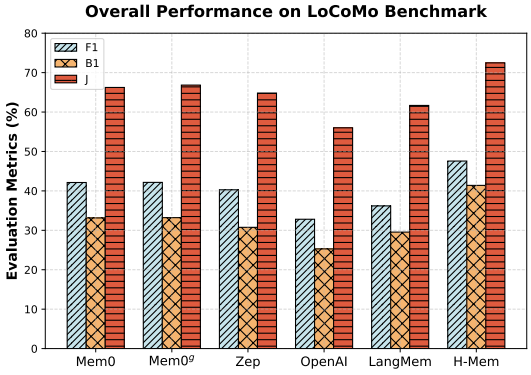

Figure 4: Overall performance comparison on the **LoCoMo** benchmark. The bars represent three evaluation metrics (**F1**, **BLEU-1**, and **J-Score**) across different memory architectures. H-Mem achieves the highest overall performance, showing consistent improvements across all metrics.

4.4.1 Results

We present our experimental results in three parts. First, we show the main comparison of H-Mem against the baselines. Second, we conduct a series of ablation studies to analyze the contribution of each component of our architecture. Finally, we provide a qualitative case study to illustrate the behavior of our model.

4.4.2 Main Results

Our main results on the LoCoMo benchmark are presented in Table 1 and Figure 4. The evaluation demonstrates that H-Mem substantially outperforms all baseline methods across every metric, highlighting the effectiveness of its hybrid multi-dimensional architecture.

As shown in Table 1, H-Mem achieves the highest overall F1, B1, and J-Score. Most notably, our model achieves a J-Score of 72.47, representing a significant improvement of over 5.6 points compared to the strongest baseline, Mem0^g. This indicates that H-Mem is better at retrieving correct information both in lexical and semantic dimension, as judged by the LLM. This improvement is statistically significant at $p < 0.05$.

The overall performance comparison, visualized in Figure 4, further illustrates H-Mem's consistent superiority. The bar chart clearly shows that H-Mem sets a new state-of-the-art on this benchmark, with its J-Score bar towering over the others.

When breaking down the performance by question category, H-Mem shows its most significant gains in the "Single-Hop", "Multi-Hop", and "Temporal" categories. The strong performance on temporal questions, where H-Mem achieves a J-Score of 57.63, underscores the value of the dedicated temporal tree. The high scores in single and multi-hop questions demonstrate the power of our hybrid retrieval mechanism, which combines a precise global vector search for seed facts with a robust hierarchical expansion to gather the context for complex reasoning.

4.4.3 Ablation Studies

To understand the impact of H-Mem's key components, we perform several ablation studies. We first analyze the effect of our LLM-based memory curation filter and then explore the contributions of our hybrid retrieval architecture.

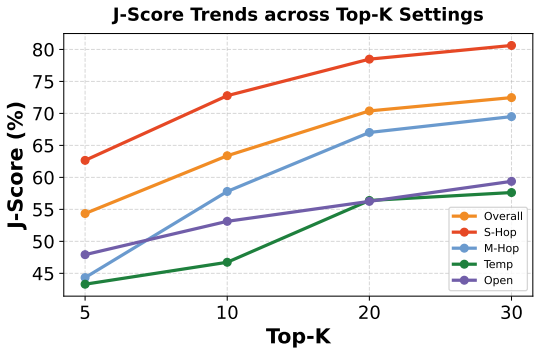

Figure 5: H-Mem performance across Top-K settings. Scores improve with increasing Top-K, shown across Overall, Single-Hop, Multi-Hop, Temporal, and Open-Domain categories.

Memory Curation Filter Our ingestion pipeline includes a Memory Curation Agent that filters out non-memorable conversational turns (simple greetings for example). To quantify the benefit of this approach, we compare the full H-Mem model against

Method	Single-Hop			Multi-Hop			Temporal			Open-Domain			Overall		
	F1(↑)	B1(↑)	J(↑)	F1(↑)	B1(↑)	J(↑)	F1(↑)	B1(↑)	J(↑)	F1(↑)	B1(↑)	J(↑)	F1(↑)	B1(↑)	J(↑)
Mem0	47.65	38.72	72.93	**38.72**	27.13	67.13	28.64	21.58	51.15	48.93	**40.51**	55.51	42.13	33.14	66.24
Mem0^g	49.27	40.30	75.71	38.09	26.03	65.71	24.32	18.82	47.19	**51.55**	40.28	58.13	42.16	33.21	66.84
Zep	49.56	38.92	76.60	35.74	23.30	61.70	19.37	14.82	41.35	42.00	34.53	49.31	40.27	30.76	64.82
OpenAI	39.31	31.16	62.29	34.30	23.72	63.79	20.09	15.42	42.92	14.04	11.25	21.71	32.81	25.28	56.00
LangMem	40.91	33.63	71.12	35.51	26.86	62.23	26.04	22.32	47.92	30.75	25.84	23.43	36.19	29.55	61.68
H-Mem	**54.53**	**48.90**	**80.62**	37.95	**29.62**	**69.50**	**45.61**	**39.16**	**57.63**	21.36	17.70	**59.38**	**47.57**	**41.40**	**72.47**

Table 1: Main results on the LoCoMo benchmark. The three metrics including F1, BLEU-1(B1), J, where J represents LLM-as-a-Judge (J-score). H-Mem outperforms all baselines in all categories in J-score comparison. The best result is significant at $p < 0.05$.

a variant where this filter is disabled (w/o Curation Filter), meaning every single turn is ingested into the memory trees. The results are shown in Table 2.

Method	F1(↑)	B1(↑)	J(↑)
H-Mem (with Filter)	47.57	41.40	72.47
w/o Curation Filter	45.63	39.82	67.34

Table 2: Ablation study of the Memory Curation Agent on the LoCoMo benchmark. The result shows the Curation Filter improves the overall correctness due to a more concise and denser memory structure.

Retrieval Architecture Next, we analyze the components of the retrieval mechanism itself. The results are presented in Table 3. We study the effect of removing the temporal hierarchy, disabling the adaptive LCA expansion, and forcing the system to use only the sequential retrieval mode.

Our study includes the following variants:

1. **w/o Temporal Tree**: In this variant, we remove the temporal tree and its expansion logic entirely. Retrieval is based solely on the globally relevant "seed facts" found by the initial FAISS search. This effectively makes the model a sophisticated RAG system without our memory curation frontend.

2. **w/o LCA Expansion**: We disable the Lowest Common Ancestor (LCA) logic in the Mode 1 search, reverting to a simpler fixed-window context. This tests the value of our adaptive temporal context window.

3. **Mode 1 Only**: We disable the Mode Controller and force the system to use only Mode 1 (Sequential Search) for all queries (as Mode 2 does not fit most queries, as there is no clear absolute timestamp, but Mode 1 fits every query). This evaluates the importance of the dynamic retrieval strategy.

The results clearly quantify the contribution of each architectural decision. Removing the temporal tree (w/o Temporal Tree) causes a noticeable drop in all scores, particularly the J-score (from 72.47 to 71.04), confirming that the temporal context provided by the hierarchy is crucial for semantic correctness. Similarly, disabling the adaptive context window (w/o LCA Expansion) also harms performance, with the J-score falling to 70.71. This demonstrates the value of the LCA method in gathering a broader, more relevant temporal context compared to a simple fixed window. Finally, forcing the system to use only Mode 1 leads to a drop in performance, showing the Mode Controller is a valuable component of the H-Mem architecture due to the ability of dynamic mode selection.

Method	F1(↑)	B1(↑)	J(↑)
H-Mem	47.57	41.40	72.47
w/o Temporal Tree	46.81	40.93	71.04
w/o LCA Expansion	48.14	41.91	70.71
Mode 1 Only	47.18	41.32	71.30

Table 3: Ablation study of H-Mem components on the LoCoMo benchmark. The full model outperforms all ablated versions, quantifying the contribution of the temporal tree, LCA-based expansion, and the dynamic mode controller.

Sensitivity to Top-K Seed Facts The initial global vector search retrieves the top-k most relevant facts to "seed" the subsequent hierarchical expansion. To analyze the model's sensitivity to this hyperparameter, we evaluated the J-Score across different values of k, from 5 to 30. The results

are shown in Figure 5. The performance improves as k increases across all question categories. This suggests that a larger set of initial seed facts provides a more robust starting point for hierarchical expansion, increasing the likelihood of capturing the necessary information to answer the query.

However, the gains begin to diminish as k approaches 30, indicating a point of diminishing returns. Based on the analysis, we set k to 30 for our main experiments and not increasing k further as it offers a strong balance between performance and computational efficiency and cost.

4.4.4 Qualitative Analysis

To provide a more intuitive understanding of H-Mem's behavior, we present two case studies that highlight the specific advantages of our hybrid retrieval mechanism.

Case 1: Sequential Search for Temporal Reasoning (Mode 1)

- **Query**: "When did Caroline mention she was going to the LGBTQ support group?"
- **Mode Selected**: The Mode Controller correctly identifies there is no absolute timeframe in the query and choose Mode 1.
- **Standard RAG Failure**: A standard RAG retrieves facts about "support groups" in general, but misses the specific temporal context.
- **H-Mem's Success**: In contrast, H-Mem's Mode 1 search first identifies the key FactNode about the LGBTQ support group via vector search, then uses its temporal parent to perform an LCA expansion, correctly retrieving nearby facts that specify the event happened "yesterday". This allows the final generation agent to correctly answer the question.

Consider the query: "When did Caroline mention she was going to the LGBTQ support group?". A standard RAG baseline retrieves facts about "support groups" in general but misses the specific temporal context. In contrast, H-Mem's Mode 1 search first identifies the key FactNode about the LGBTQ support group via vector search, then uses its temporal parent to perform an LCA expansion, correctly retrieving nearby facts that specify the event happened "yesterday". This allows the final generation agent to correctly answer the question.

Case 2: Intersective Search for Robust Reasoning (Mode 2)

- **Query**: "How many times has Melanie gone to the beach in 2023?"
- **Mode Selected**: The Mode Controller correctly identifies the absolute timeframe ("2023") and choose Mode 2.
- **Standard RAG Failure**: A standard RAG system fails. It retrieves memories about "beach" from other years and general memories about "2023" that have nothing to do with the beach, leading to the incorrect count.
- **H-Mem's Success**: H-Mem performs a precise, two-part search:
  1. It finds all memories related to the "beach" topic in the **semantic tree**.
  2. It finds all memories from the "2023" node in the **temporal tree**.
  3. It then takes the intersection of these two sets, resulting in a clean, accurate list of only the beach trips that happened in 2023, which allows for the synthesis of a correct final answer.

5 Conclusions

In this work, we introduced H-Mem, a novel hybrid multi-dimensional memory architecture specifically designed to overcome the critical limitations of single-axis memory systems, which often leads to context collapse or catastrophic forgetting in long-context conversational agents. By organizing memories into parallel temporal and semantic hierarchies, our solution effectively captures the rich, interwoven nature of dialogue. It employs a dynamic hybrid retrieval mechanism that flexibly combines a global vector search for key "seed" memories with a rich hierarchical expansion. This ensures both speed in identifying relevant information and depth in reconstructing the necessary conversational context.

Our comprehensive experiments on LoCoMo benchmark demonstrate that H-Mem significantly outperforms state-of-the-art baselines, showing particularly strong gains in complex temporal and multi-hop reasoning tasks. These results validate our multi-dimensional approach as a more robust foundation for building intelligent, coherent, and

truly stateful conversational agents that can maintain consistency over long interactions. Future work could enrich the memory with new dimensions (like user sentiment and conversational goals) and more sophisticated graph-based methods that integrate external knowledge. A more fundamental vision aims to transcend simple retrieval, developing a system capable of proactive memory management to anticipate user needs and meta-cognition to self-correct its own knowledge gaps.

Limitations

Current agent memory storage and retrieval heavily rely on powerful LLM systems. In the future, we will attempt to further decompose challenging tasks to enable dialog question-answering tasks using smaller language models. Plus, current modeling of semantic and temporal trees remains relatively rudimentary. In the future, we will leverage LLM to construct more dynamic trees featuring seamless memory management between storage and retrieval.

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*.

John R Anderson. 1983. A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior*, 22(3):261–295.

Owen Bianchi, Mathew J Koretsky, Maya Willey, Chelsea X Alvarado, Tanay Nayak, Adi Asija, Nicole Kuznetsov, Mike A Nalls, Faraz Faghri, and Daniel Khashabi. 2025. Lost in the haystack: Smaller needles are more difficult for llms to find. *arXiv preprint arXiv:2505.18148*.

Weixin Chen, Yuhan Zhao, Jingyuan Huang, Zihe Ye, Clark Mingxuan Ju, Tong Zhao, Neil Shah, Li Chen, and Yongfeng Zhang. 2026. Memrec: Collaborative memory-augmented agentic recommender system. *arXiv preprint arXiv:2601.08816*.

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*.

Tsz Ting Chung, Leyang Cui, Lemao Liu, Xinting Huang, Shuming Shi, and Dit-Yan Yeung. 2024. Selection-p: Self-supervised task-agnostic prompt compression for faithfulness and transferability. *arXiv preprint arXiv:2410.11786*.

Payel Das, Subhajit Chaudhury, Elliot Nelson, Igor Melnyk, Sarath 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. *Preprint*, arXiv:2403.11901.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z Pan. 2025. Rethinking memory in ai: Taxonomy, operations, topics, and future directions. *arXiv preprint arXiv:2505.00675*.

Minghao Guo, Xi Zhu, Haochen Xue, Chong Zhang, Shuhang Lin, Jingyuan Huang, Ziyi Ye, and Yongfeng Zhang. 2025. Reagan: Node-as-agent-reasoning graph agentic network. *Preprint*, arXiv:2508.00429.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Zifan He, Yingqi Cao, Zongyue Qin, Neha Prakriya, Yizhou Sun, and Jason Cong. 2024. Hmt: Hierarchical memory transformer for efficient long context language processing. *arXiv preprint arXiv:2405.06067*.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2024. Metagpt: Meta programming for a multi-agent collaborative framework. International Conference on Learning Representations, ICLR.

Jingyuan Huang, Dan Luo, Zihe Ye, Weixin Chen, Minghao Guo, and Yongfeng Zhang. 2026. From aggregation to selection: User-validated distributed social recommendation. *Preprint*, arXiv:2505.21388.

Romuald A Janik. 2023. Aspects of human memory and large language models. *arXiv preprint arXiv:2311.03839*.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. *arXiv preprint arXiv:2403.14403*.

Xia Jiang, Wei Xu, Ming Zhao, Li Chen, and Zhicheng Dou. 2024. Graphrag: Unlocking llms for graph-structured reasoning. *arXiv preprint arXiv:2403.09236*.

Greg Kamradt. 2023. Llmtest_needleinahaystack: Doing simple retrieval from llm models at various context lengths to measure accuracy.

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

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.

LangChain. 2025. Langmem. https://langchain-ai.github.io/langmem/.

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.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2023. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*.

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

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

OpenAI. 2024. Hello gpt-4o. *OpenAI Blog*.

OpenAI. 2024. Memory and new controls for chatgpt. https://openai.com/index/memory-and-new-controls-for-chatgpt/.

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

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, and 1 others. 2025. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. Memorag: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of the ACM on Web Conference 2025*, pages 2366–2377.

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*.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. 2025. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for LLMs. In *The Thirteenth International Conference on Learning Representations*.

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*.

Larry R Squire and Pablo Alvarez. 1995. Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current opinion in neurobiology*, 5(2):169–177.

Larry R Squire, Lisa Genzel, John T Wixted, and Richard G Morris. 2015. Memory consolidation. *Cold Spring Harbor perspectives in biology*, 7(8):a021766.

Weihang Su, Qingyao Ai, Jingtao Zhan, Qian Dong, and Yiqun Liu. 2025. Dynamic and parametric retrieval-augmented generation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 4118–4121.

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

Zhen Tan, Jun Yan, I Hsu, Rujun Han, Zifeng Wang, Long T 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. *arXiv preprint arXiv:2503.08026*.

Marlieke TR Van Kesteren, Dirk J Ruiter, Guillén Fernández, and Richard N Henson. 2012. How schema and novelty augment memory formation. *Trends in neurosciences*, 35(4):211–219.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. 2025. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193.

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*.

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

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, Tong Xu, and 1 others. 2025a. Towards multi-granularity memory association and selection for long-term conversational agents. *arXiv preprint arXiv:2505.19549*.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. 2025b. A-mem: Agentic memory for llm agents. *Preprint*, arXiv:2502.12110.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Hinrich Schütze, Volker Tresp, and Yunpu Ma. 2025. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*.

Haofei Yu, Cunxiang Wang, Yue Zhang, and Wei Bi. 2023. Trams: Training-free memory selection for long-range language modeling. *arXiv preprint arXiv:2310.15494*.

Sizhe Yuen, Francisco Gomez Medina, Ting Su, Yali Du, and Adam J Sobey. 2025. Intrinsic memory agents: Heterogeneous multi-agent llm systems through structured contextual memory. *arXiv preprint arXiv:2508.08997*.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025a. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*.

Qinggang Zhang, Shengyuan Chen, Yuanchen Bei, Zheng Yuan, Huachi Zhou, Zijin Hong, Junnan Dong, Hao Chen, Yi Chang, and Xiao Huang. 2025b. A survey of graph retrieval-augmented generation for customized large language models. *arXiv preprint arXiv:2501.13958*.

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

A LoCoMo Categories Classification

A notable ambiguity exists in the original LoCoMo dataset and its accompanying paper concerning the mapping between category indices and their corresponding names. The benchmark paper presents the results for Conversational QA in Table 2, listing the categories in the following order:

1. **Single Hop**
2. **Multi Hop**
3. **Temporal**
4. **Open Domain**
5. **Adversarial**

This presentation has led subsequent researchers to assume a direct correspondence between the numerical indices (1-5) and this list of category names, often without further inspection of the dataset's contents.

To clarify this discrepancy, we performed a detailed analysis of the dataset, exemplified here by the first conversation (sample_id: conv-26), whose question distribution is representative of the entire dataset. Before presenting our findings, we briefly summarize the category definitions from the original paper:

- **Single Hop vs. Multi Hop**: Single Hop questions can be answered using evidence from a single session, whereas Multi Hop questions can require integrating information from multiple sessions.
- **Temporal**: These questions pertain to time-related information (e.g., when an event occurred).
- **Open Domain**: These questions may span multiple sessions but are distinct from Multi Hop as they typically involve hypotheticals (using terms like "would" or "will") rather than established facts.

- **Adversarial**: These questions are designed to have no answer within the conversational context, expecting a "No answer" response from the QA system.

Based on our analysis of the QA pairs and the question distribution statistics, we have determined the correct mapping. We propose that the following mapping be adopted by future researchers to ensure fair comparisons and enhance the reliability of experimental results:

- **Category 1**: "Multi Hop"
- **Category 2**: "Temporal"
- **Category 3**: "Open Domain"
- **Category 4**: "Single Hop"
- **Category 5**: "Adversarial"

A.1 Reasoning from Content

Previous works have typically misidentified Category 1 as Single Hop. However, our analysis, supported by the examples in Table 4, invalidates this assumption. For instance, the second and third example questions for Category 1 draw evidence from multiple dialogue sessions, which is the defining characteristic of Multi Hop questions. In the provided evidence notation Dn:m, n refers to the dialogue/session number and m refers to the turn number. Accordingly, Question #2 requires evidence from both the second ($D2$) and third ($D3$) sessions, while Question #3 synthesizes an answer from three distinct sessions. This cross-session evidence requirement definitively classifies them as Multi Hop.

The classification of Category 2 as Temporal is straightforward. As shown in Table 4, the example questions consistently pertain to time-related queries, such as specific times or the duration of events or vice versa, aligning perfectly with the Temporal definition.

For Category 3, we identify the questions as Open Domain. The original LoCoMo paper defines this category as focusing on hypothetical scenarios and reasoning about future possibilities. This is consistent with the examples in Table 4, which prominently feature subjunctive words like "would."

Finally, we classify Category 4 as Single Hop. Although these questions share a similar thematic scope with Multi Hop questions, their evidentiary basis is fundamentally different. For all Category 4 examples, the required evidence, even if composed of multiple pieces, is contained entirely within a single session (e.g., the third example draws multiple facts, but all are from the same session).

A.2 Reasoning from Distribution

We further corroborate our proposed mapping by analyzing the question distribution statistics provided in the original LoCoMo paper (Table 5). A minor limitation of this analysis is the discrepancy between the dataset referenced in the paper (50 conversations) and the publicly released version (10 conversations, locomo10.json). Consequently, our analysis relies on the distribution ratios of each question category rather than absolute counts.

While this prevents a definitive proof from statistics alone, the observed ratios in the public dataset align with those reported for our proposed mapping. It is also worth noting that the correct classification of these four categories implies, by process of elimination, that Category 5 correctly corresponds to Adversarial questions.

To quantitatively validate our hypothesis, we compare the question distribution in the public LoCoMo10 dataset against the statistics reported for the full LoCoMo50 dataset in the paper. The results, presented in Table 5, show a remarkable consistency when our proposed mapping is used.

Crucially, the frequency ranking of the categories is identical across both datasets: Category 4 (Single Hop) is the most common, followed by Category 2 (Temporal), Category 1 (Multi-Hop), and finally Category 3 (Open Domain). While the precise distribution percentages show minor variations, likely due to the difference in dataset sizes, this perfect alignment in the ordinal ranking strongly corroborates our findings.

Thus, our conclusion is supported by two independent and converging forms of evidence. The initial reclassification based on the definitional characteristics of the questions is validated by this subsequent statistical analysis, confirming our proposed mapping is correct.

B Human-Inspired Memory Reasoning

B.1 Tutorial

Prior studies on memory modeling like MemTree (Rezazadeh et al., 2025), SeCom (Pan et al., 2025), MemGAS (Xu et al., 2025a), MemoryOS (Kang et al., 2025), and MemOS (Li et al., 2025) have been made from two major

Category Index	Question ID	Question	Evidence
1	1	What did Caroline research?	"D2:8"
	2	What is Caroline's relationship status?	"D3:13", "D2:14"
	3	What activities does Melanie partake in?	"D5:4", "D9:1", "D1:12", "D1:18"
2	1	When did Melanie paint a sunrise?	"D1:12"
	2	When did Caroline meet up with her friends, family, and mentors?	"D3:11"
	3	How long has Caroline had her current group of friends for?	"D3:13"
3	1	What fields would Caroline be likely to pursue in her education?	"D1:9", "D1:11"
	2	Would Melanie be more interested in going to a national park or a theme park?	"D10:12", "D10:14"
	3	Would Caroline be considered religious?	"D14:19", "D12:1"
4	1	What country is Caroline's grandma from?	"D4:3"
	2	Did Melanie make the black and white bowl in the photo?	"D5:8"
	3	How did Melanie's son handle the accident?	"D18:6", "D18:7"

Table 4: Examples of Questions and Evidence by Category.

Category Index	LoCoMo10 Count	Category Names	LoCoMo50 Paper Count
1	282 (18.31%)	Multi-Hop	1,104 (19.47%)
2	321 (20.84%)	Temporal	1,547 (27.28%)
3	96 (6.23%)	Open Domain	285 (5.03%)
4	841 (54.61%)	Single Hop	2,705 (47.70%)
Total (w.o. adversarial)	**1,540**	**Total (w.o. adversarial)**	**5,671**

Table 5: LoCoMo Dataset Counts by Category.

perspectives: the temporal dimension (capturing aspects of sequence and recency) and the semantic dimension (capturing aspects of meaning and topicality). However, we see much less research on multi-dimensional hierarchical memory that jointly captures both axes.

To give more motivations, we first studied the long-context conversational QA benchmark **LoCoMo** from a human perspective. The primary observation is that **LoCoMo** exhibits a large performance gap between human and agents, the reason is that we carefully observed and found that humans are not reasoning over a large span of text in a linear fashion. In particular, human take a well-structured, two-stage reasoning process:

- **Memory Construction:** Humans read session by session, summarizing each into concise bullet points. They track the frequency of concepts (semantically related events), and distinguish high-frequency concepts forming the main storyline from low-frequency ones tied to specific episodes. This process builds a hierarchical memory structure effectively, where high-level concepts anchor the main storyline and low-level details can be retrieved selectively.
- **Memory Reasoning and Retrieval:** Humans begin by interpreting the question, decomposing it into an adaptive plan guided by the question words. They parse high-frequency and low-frequency concepts: when high-frequency concepts dominate, they revisit the sessions containing them; when specific low-frequency concepts appear, they locate the few sessions where they occur. While rereading, explicit clues are extracted and organized into an answer. If no explicit evidence emerges, or if the question itself suggests uncertainty, humans rely on prior knowledge to infer a plausible response.

B.2 Performance of H-Mem VS Humans

In order to further compare our model's reasoning process with human performance, we asked 10 volunteers aged from 18 to 25 to answer all the questions in LoCoMo10 in the same manner as outlined in the tutorial. The answers from H-Mem and human participants were then compared on four

Category	F1(↑)		BLEU-1(↑)		J-Score(↑)	
	H-Mem	Human	H-Mem	Human	H-Mem	Human
All	**47.57**	36.84	**41.40**	37.00	72.47	**75.71**
M-Hop	**37.95**	20.25	**29.62**	23.51	**69.50**	58.16
S-Hop	**54.53**	46.45	**48.90**	45.71	80.62	**82.64**
Temp	**45.61**	33.99	**39.16**	33.47	57.63	**78.19**
Open	**21.36**	10.86	**17.70**	12.05	**59.38**	58.33

Table 6: Performance comparison between H-Mem and human evaluators across QA categories.

QA categories: Multi-Hop, Single-Hop, Temporal, and Open-Domain. The overall results, shown in Table 6, provide a glimpse that whether human evaluator group or H-Mem is good at and where each falls short. In total, H-Mem's results outperform those of human participants in terms of the surface metrics (F1: 47.57 vs. 36.84; BLEU-1: 41.40 vs. 37.00) indicating that, on average, H-Mem can generate more complete and factually consistent responses than humans. By contrast, human participants outperform H-Mem in terms of J-Score (75.71 vs. 72.47), suggesting that human-generated answers are more semantically aligned and exhibit a better holistic understanding of questions, and that's the reason why J-score is the primary evaluation metric.

A more fine-grained breakdown by QA category further demonstrates the strengths and weaknesses of both groups, as well as their potential for complementarity.

- **Multi-hop questions:** H-Mem performs significantly better than human participants, with gains of 17.7 and 6.1 in F1 and BLEU-1, respectively, when it comes to combining scattered clues from different parts of the conversation.

- **Single-hop questions:** The performance of H-Mem is close to that of human participants, with H-Mem being slightly better, indicating a similar capability to handle single-fact questions.

- **Temporal questions:** H-Mem's results are higher than human participants in both F1 and BLEU-1, while human participants outperform H-Mem in J-Score (78.19 vs. 57.63), suggesting human participants have a better temporal understanding of the questions and the ability to infer causal relations over long conversations.

- **Open-domain questions:** H-Mem achieves the best margin over human participants in all three metrics, suggesting a higher coverage of external knowledge and factual recall ability.

Overall, H-Mem outperforms humans in terms of multi-hop retrieval and fact composition, while humans remain more robust in scenarios that require temporal reasoning or capturing semantic subtleties of the conversation.

B.3 Human Evaluation Studies

To provide a more intuitive demonstration of human performance in long-context reasoning tasks, we invited multiple volunteers to manually complete all questions in the LoCoMo dataset without consulting the answers. After reading the full long-form dialog text, volunteers answered two main types of question: temporal reasoning based on the timeline, and content-based reasoning. Question types encompassed various forms found in LoCoMo, including single-hop, multi-hop, and open-domain reasoning. During human responses, we recorded behavioral pathways involving reading, locating, integrating, and reasoning. These findings provide valuable insights for comparing differences between humans and models in long-context information integration and temporal consistency.

Temporal Questions. Figures 6–9 illustrate representative temporal reasoning examples. Human evaluators performed remarkably well in temporal reasoning tasks, as they can intuitively anchor events to concrete timestamps or relative expressions (like "two weeks ago," "last month," "under a year"). Once the relevant time points are located, temporal inference becomes straightforward. This advantage demonstrates that human cognition naturally excels at constructing chronological narratives and maintaining temporal coherence across long contexts. Such observations highlight potential directions for improving memory architectures in long-context models. Future systems could incor-

Case Study 1: Human Answered Correctly in Temporal-related Question (Example 1).

Question: How many times has Jolene been to France?
Human:
Step 1: For the place name related to France, the conversation only mentioned . So I'll identify which sessions contain references to Paris.
Step 2: I reviewed my reading notes and found that **Session 1** and **Session 2** include scenes mentioning Paris. I'll examine them more closely.
Step 3: **Session 1** mentions that Jolene's mother gave her a pendant purchased during their trip to Paris in 2010.
Step 4: **Session 2** describes Jolene discussing the small snake Seraphim with Deborah, where Jolene mentions buying this snake in Paris a year ago.
Step 5: Therefore, Jolene has visited Paris twice in total.
Human Answer: Twice.
LoCoMo Answer: Twice.

Figure 6: An example for how human evaluators answer the temporal-related question correctly. This question requires counting the number of times specific episodes appear.

Case Study 2: Human Answered Correctly in Temporal-related Question (Example 2).

Question: How many weeks passed between Maria adopting Coco and Shadow?
Human:
Step 1: To find the answer to this question, one must revisit the episodes involving the adoption of Coco and the adoption of Shadow.
Step 2: Let's recap all the sessions mentioned Coco and Shadow, since these two keywords can help to locate more accurately.
Step 3: Only **Session 30** mentioned Coco, and only **Session 31** mentioned Shadow.
Step 4: According the episodes in **Session 30**, Maria got a puppy named Coco two weeks ago before August 11, 2023.
Step 5: According the episodes in **Session 31**, Maria just adopted Shadow on August 13, 2023.
Step 6: Since Aug 11 and Aug 13 are only two days apart, the duration between two adoptions can be considered as two weeks, which is the answer.
Human Answer: Two weeks.
LoCoMo Answer: Two weeks.

Figure 7: An example for how human evaluators answer the temporal-related question correctly. This question requires calculating the time duration between two moments.

porate human-like temporal indexing or timeline reconstruction mechanisms to better preserve and retrieve temporally dependent information.

Content-based Questions. Figures 10–15 provide several content-based reasoning examples. If the question keywords are exact matches to explicit or low-frequency entities (for example, a dog name "Max" or cities "Toronto" and "Vancouver"), humans can easily identify the corresponding passages and find the correct answer. In contrast, when clues are implicit or distributed (for example, for which the screenplay was rejected), mistakes are more likely to be caused by a failure to integrate different evidence than by misinterpretation. Moreover, there exist open-domain questions for which there is no single ground-truth answer. Human evaluators can have multiple reasonable interpretations. In such cases, the J-Score and other automatic metrics are less informative because they do not represent the semantic variety of human reasoning. Evaluating open-domain answers requires human judgment or a more flexible semantic matching method to approximate it.

Case Study 3: Human Answered Correctly in Temporal-related Question (Example 3).

Question: Which places in Canada was Evan visiting in July 2023?

Human:

Step 1: To find out where Evan traveled in Canada during July 2023, we need to search for Canadian place names while restricting the results to July. After reviewing the sessions containing Canadian place names, they are: **Session 1** (Rockies), **Session 2** (Jasper), **Session 8** (Banff), **Session 9** (Rocky Mountains), **Session 19** (Banff).

Step 2: Then, let's narrow the scope to July 2023. Actually, only **Session 4** took place in July 2023. Therefore, we need to adjust our search strategy to determine when Evan actually traveled to Canada. He likely recalled these trips during other sessions.

- **Session 1:** Evan talked about a recent family trip to the Rockies. (18 May 2023)
- **Session 2:** Evan mentioned the family trip to Jasper on the past weekend of 24 May 2023.
- **Session 8:** Evan implied that he traveled to Canada last month, since **Session 8** is in August 2023, the trip he referred is in July. In this session, Evan explicitly said he went to Banff.
- **Session 9:** Evan continued to talk about the July Canada trip and mentioned he went to Rocky Mountains.
- **Session 19:** Evan showed a photo from Banff and reminisced, a memory of the July trip.

Step 3: Since Jasper trip took place in May, it shouldn't be included in the answer. Therefore, Evan traveled to Banff and the Rocky Mountains in July 2023 (possibly a different location from Banff).

Human Answer: Banff, Rocky Mountains.

LoCoMo Answer: Banff, Rocky Mountains.

Figure 8: An example for how human evaluators answer the temporal-related question correctly. To answer this question correctly, one need to figure out the facts from July 2023 and facts about Evan's trip to Canada simultaneously.

Case Study 4: Human Answered Incorrectly in Temporal-related Question.

Question: When did John start playing professionally?

Human:

Step 1: This requires reviewing all episodes involving playing professionally. I revisited the scenes where John participated in professional ball games. Playing professionally is a high-frequency concept spanning **Sessions 1, 2, 3, 5, 6, 7, 8, 11, 13, 14, 16, 18, 19, 21, 22, 23, 24, 25, 26, 28, and 29.**

Step 2: Since start playing implies a temporal aspect, let's identify the sessions that mention starting or continuing over a period of time, from my note taken previously...

- **Session 1:** John just signed with the Minnesota Wolves.
- **Session 6:** John reminisced about getting drafted and how his pro career began.
- **Session 8:** John started a new workout regime to help his basketball game.
- **Session 14:** John began a new journey in life by starting to mentor younger players.
- **Session 20:** John started a new yoga practice to work on his strength and flexibility.
- **Session 21:** John expressed that he had been playing professionally for under a year.
- **Session 24:** John shared that his recovery was good and continuing to improve.
- **Session 25:** John reflected on his new endorsement and team's continued success.
- **Session 26:** John began offering seminars on sports and marketing.
- **Session 28:** John put together a benefit basketball game and was still involved.
- **Session 29:** John announced that he just signed a new endorsement for a beverage company.

Step 3: Since **Session 21** has mentioned about John had been playing professionally for under a year, therefore answer is not later than **Session 21** (Dec 6 2023).

Step 4: It seemed that John didn't mention when he started playing professionally, I can't find any time related words with numerical signs like "yesterday", "a month ago"....

Step 5: Therefore, the best answer I tried to give is "under a year from December 6, 2023".

Human Answer: Under a year from December 6 2023.

LoCoMo Answer: May 2023, since John signed Minnesota Wolves in Session 1.

Figure 9: An example for how human evaluators answer the temporal-related question incorrectly. The example didn't mention the timing explicitly, it paraphrased the "playing professionally" concept in Session 1 (May 21, 2023).

Case Study 5: Human Answered Correctly in Content-based Question (Explicit).

Question: What are the names of James's dogs?
Human:
Step 1: This is a content-based question that requires identifying all sessions where James's dogs are mentioned.
Step 2: By reviewing the sessions, the following references were found:
- **Session 5:** James shared about adopting a pup named Ned, saying it made his days happier.
- **Session 9:** James mentioned spending time with his Labrador Daisy and two loyal shepherds.
- **Session 14:** James talked about his dog Max, who enjoys swimming and playing fetch.
- **Session 15:** James introduced Max, Daisy, and the new pup Ned, noting that they were slowly bonding.

Step 3: By combining information from all relevant sessions, James owns three dogs: Max, Daisy, and Ned.
Human Answer: Max, Daisy, Ned.
LoCoMo Answer: Ned, Daisy, Max.

Figure 10: An example showing how human evaluators correctly answer a content-based question. This question requires integrating factual mentions of entities across multiple sessions rather than temporal reasoning.

Case Study 6: Human Answered Correctly in Content-based Question (Explicit).

Question: Which of Joanna's screenplays were rejected from production companies?
Human:
Step 1: I need to locate all sessions mentioning Joanna's screenplays and find where rejection from production companies is explicitly stated.
Step 2: Reviewing the sessions that mentioned screenplays.
- **Session 2:** Joanna completed her first screenplay and planned to submit it to film festivals.
- **Session 4:** Joanna discussed a new screenplay and finishing her second script, expressing mixed emotions for the critical comments.
- **Session 9:** She was working on a screenplay titled "Finding Home."
- **Session 14:** Joanna shared she had received a rejection letter for her screenplay from a production company and felt discouraged.
- **Session 24:** Joanna mentioned a recent setback with a production company rejection. Nate offered support.

Step 3: Based on the sessions, the screenplays explicitly rejected by production companies are those mentioned in **Session 14** and **Session 24**. Since I didn't find evidence for the acceptance of the second screenplay, the rejection mentioned in **Session 14** should be refer to the second, and **Session 24**'s refer to the third.
Human Answer: The second and the third.
LoCoMo Answer: The third screenplay.

Figure 11: An example showing how a human evaluator answers an explicit content-based question by identifying sessions where Joanna's screenplays were explicitly rejected by production companies. However, the answer was incorrect since the human evaluator didn't figure out which screenplay was exactly rejected.

Case Study 7: Human Answered Correctly in Content-based Question (Implicit).

Question: Which type of vacation would Evan prefer with his family, walking tours in metropolitan cities or camping trip in the outdoors?
Human:
Step 1: The question does not contain explicit matches to phrases like "metropolitan city" or "camping trip." Therefore, the answer must be inferred from contextual evidence across sessions mentioning travel or leisure activities.
Step 2: I reviewed the sessions that describe Evan's vacations or preferred activities:
- **Session 1:** Evan shared about a recent trip to the Rockies, emphasizing nature and relaxation.
- **Session 2:** Evan mentioned a road trip to Jasper, describing it as peaceful and rejuvenating.
- **Session 6:** Evan talked about a vacation in Canada filled with outdoor activities.
- **Session 14:** Evan suggested going on a hike with Sam to enjoy nature and relieve stress.
- **Session 20:** Evan shared that his son's bike accident happened during an outdoor activity, and he later showcased a painting inspired by nature.
Step 3: Across these sessions, Evan consistently preferred open-air, nature-based experiences, with no mention of enjoying crowded or urban sightseeing. Given the options "walking tours in metropolitan cities" and "camping trip in the outdoors", the reasoning aligns strongly with the latter.
Human Answer: Camping trip in the outdoors.
LoCoMo Answer: Camping trip in the outdoors.

Figure 12: An example for how human evaluators correctly answer an implicit content-based question. This question requires inference rather than direct keyword matching: the evidence comes from multiple sessions emphasizing outdoor and nature-related experiences rather than city tours.

Case Study 8: Human Answered Correctly in Content-based Question (Implicit).

Question: Which country did James book tickets for in July 2022?
Human:
Step 1: I need to find all place names mentioned in the conversation to infer which country they belong to.
Step 2: I will check each session containing geographic locations:
- **Session 6:** James mentioned he had visited Italy, Turkey, and Mexico, while John talked about his trip to Japan.
- **Session 16:** James said he planned to travel to Toronto and Vancouver, departing on July 11 and returning on July 20.
- **Session 17:** James added that during his trip to Canada, he also spent some time in Nuuk, the capital of Greenland.
Step 3: Since **Session 6** occurred in April, those countries (Italy, Turkey, Mexico) are from earlier trips and can be excluded.
Step 4: **Session 16** clearly takes place in July, and the destinations Toronto and Vancouver are in Canada.
Step 5: **Session 17** mentions Greenland, although Greenland is accessible by plane from Toronto and Vancouver, it is not a country. Therefore, the correct country is Canada.
Human Answer: Canada.
LoCoMo Answer: Canada.

Figure 13: An example showing how human evaluators correctly answer an implicit content-based question. The reasoning process requires identifying multiple place names across sessions and aligning them with the correct time frame.

Case Study 9: Human Answered Incorrectly in Content-based Question (Implicit).

Question: What areas of the U.S. has John been to or is planning to go to?

Human:

Step 1: I need to search for all sessions mentioning U.S. locations and determine whether John has been to or is planning to go to them.

Step 2:

- **Session 11:** John shared photos from his road trip to the Pacific Northwest, describing its peaceful scenery.
- **Session 18:** John went mountaineering and visiting a park with his family, while Maria mentioned a trip to Oregon, suggesting both were discussing the northwestern U.S..

Step 3: Therefore, John has been to the Pacific Northwest, and his recent outdoor activities also indicate travel within the northwestern U.S. region. No other regions are mentioned as destinations he plans to visit.

Human Answer: Pacific Northwest, national parks.

LoCoMo Answer: Pacific Northwest, east coast.

Figure 14: An example showing how humans derive an implicit content-based answer by inferring U.S. regions from scattered geographical clues across multiple sessions. Since this is an inference question with no definitive answer, using LLM-as-Judge to evaluate it may not yield accurate results.

Case Study 10: Human Answered Incorrectly in Content-based Question (Implicit).

Question: What are John's suspected health problems?

Step 1: The question has a keyword "suspected", implying that I should search for all sessions mentioning physical conditions, fatigue, or exercise. These may indicate potential health problems even if not stated directly.

Step 2: After checking my reading notes, I noticed that **Session 7, 10 and 13** contain descriptions of John's unhealthy lifestyle habits. I'll take a closer look at these sessions.

Step 3: According to **Session 7**, John mentioned feeling exhausted and suffering from tight shoulders due to long hours of coding, which may imply cervical strain.

Step 4: According to **Session 10**, John started doing yoga to feel more relaxed, suggesting previous tension and discomfort.

Step 5: According to **Session 13**, John reported spending long hours gaming and using computers. This could point to myopia or vision problems.

Step 6: Therefore, I think John likely suffers from eye strain (myopia) and neck problems (cervical spondylosis) due to his sedentary, screen-heavy lifestyle.

Human Answer: Myopia, cervical spondylosis.

LoCoMo Answer: Obesity.

Figure 15: An example showing how implicit clues about health need interpretation. The human inferred eye and neck strain from occupational details, rather than obesity.