ActMem: Bridging the Gap Between Memory Retrieval and Reasoning in LLM Agents

Anonymous ACL submission

Abstract

Memory management is essential for LLM agents in long-term interactions. Current memory frameworks typically treat agents as passive "recorders" and retrieve information without understanding its deeper implications. They may fail in scenarios requiring reasoning and complex decision-making. To bridge this critical gap, we propose a novel actionable memory framework called ActMem that integrates memory retrieval with active causal reasoning. ActMem transforms unstructured dialogue history into a structured causal and semantic graph. By leveraging counterfactual reasoning and commonsense completion, it enables agents to deduce implicit constraints and resolve potential conflicts between past states and current intentions. Furthermore, we introduce a comprehensive dataset ActMemEval to evaluate agent reasoning capabilities in logic-driven scenarios, moving beyond the fact-retrieval focus of existing memory benchmarks. Experiments demonstrate that ActMem significantly outperforms baselines in handling complex, memory-dependent tasks, paving the way for more consistent and reliable intelligent assistants.

1 Introduction

Large language models (LLMs) have shown remarkable capabilities in natural language understanding and generation, serving as the backbone for various agents. To enable these agents to handle long-term interactions, effective memory management is essential (Zhang et al., 2025c). Conventionally, memory frameworks in LLM agents focus on expanding context windows or utilizing retrieval-augmented generation (RAG) to store, update and recall historical interactions. They aim to improve *the capacity of memory recall*, ensuring that the agent can retrieve specific information from a vast repository of past dialogues (Hu et al., 2025).

However, a fundamental gap remains between simply remembering the past and effectively using it. Existing memory management methods typically serve as a passive "recorder," where the goal is to compress, summarize, and retrieve historical text (Fang et al., 2026; Yang et al., 2026). While RAG-based methods can fetch relevant dialogue snippets (Gao et al., 2023), they often fail to understand the deeper implications of that memory for current decision-making. In real-world applications, an agent acts as a human assistant, not just an "archivist" (He et al., 2025). For instance, when a user proposes a new schedule, the agent should not only recall previous similar ones but also find event conflicts. This requires the agent to go beyond memory retrieval and perform active reasoning based on historical states. We define this capability as *the utility of memory for action*.

This limitation is also reflected in current memory benchmarks. Existing datasets, such as LoCoMo (Maharana et al., 2024), LongMemEval (Wu et al., 2025), and HaluMem (Chen et al., 2025), primarily evaluate the accuracy of memory retrieval through QA tasks. They assess whether the agent can "find" the answer but overlook whether the agent can "reason" upon it. They lack scenarios involving complex causal reasoning where the answer is not explicitly stated in the history but must be deduced. Their evaluations fail to measure how well an agent can utilize memory to assist in current planning that may conflict with past state.

To show the limitations of current benchmarks and the motivation of our work, we present a comparison in Figure 1. Subfigure (a) shows an example from LongMemEval. The user's query explicitly matches a record in the past memory summaries. The agent only needs to perform semantic or vectorized matching to retrieve the fact. This is a task that evaluates the capacity of memory recall.

In contrast, Subfigure (b) shows a more realistic and challenging scenario that necessitates memory reasoning. The user asks where to buy "Sago Palms." A retrieval-only agent may simply return

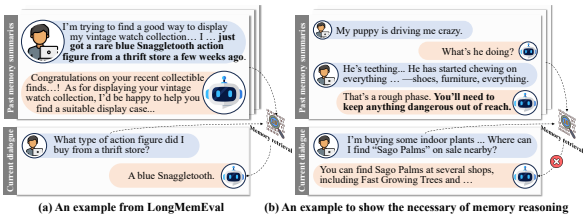


Figure 1: A motivating comparison between memory retrieval and our memory reasoning. (a) A memory retrieval example from LongMemEval. (b) A memory reasoning example. The current query (buying Sago Palms) has little semantic overlap with the past memory (a teething puppy), but the agent should still infer the hidden conflict because Sago Palms are toxic to dogs.

shopping information. However, the dialogue history contains a critical information: the user has a puppy that is "teething" and "chewing on everything." Given that Sago Palms are highly toxic to dogs, a responsible agent should infer the potential danger and intervene with a warning. We find that LLMs generally possess this commonsense knowledge, but current memory systems often fail to connect it with the stored history. The inability to reason over memory can lead to conflicting or even dangerous advice. This is the gap we aim to bridge. We want to turn the agent from a passive retriever into an active reasoner.

To bridge this gap, we propose ActMem, a memory framework that combines retrieval with causal reasoning. It first extracts memory facts from raw dialogue history. Then it clusters related facts and adds semantic edges between similar fact pairs. After that, it mines causal relations of facts. Furthermore, it performs counterfactual reasoning, commonsense completion and PMI-based filtering to refine cause edges. These facts and edges form a memory knowledge graph (KG). The KG helps the agent find associations, detect conflicts, and infer hidden constraints for current actions. We also introduce a new dataset, ActMemEval, to evaluate this setting. Unlike traditional retrieval benchmarks, it focuses on active reasoning and causal deduction. In summary, the contributions of this paper are as follows:

- We identify the critical limitation of current agent memory systems: the disconnect between memory retrieval and reasoning. We propose a shift in focus from the "capacity of recall" to the "utility of memory for action."
- We introduce ActMem, an actionable memory management framework. By integrating a memory KG with causal and semantic edges, ActMem enables agents to utilize past experiences for robust decision-making.
- We construct a comprehensive dataset ActMemEval to evaluate the reasoning capabilities of agents in long-term interactions, covering complex tasks that require logic-driven memory utilization.
- Extensive experiments demonstrate that ActMem significantly outperforms existing baselines in handling complex, logic-driven tasks. It also achieves competitive and even SOTA performance on conventional datasets.

2 Related Work

2.1 Agent Memory Management

Memory management is a core capability for intelligent agents in long-term interactions (Zhang et al., 2025c; Hu et al., 2025). Existing studies mainly fall into three groups: retrieval-centric memory, structured memory, and lightweight memory systems.

Retrieval-centric memory management. These methods mainly improve access to relevant history in long contexts. Methods such as Mem0 (Chhikara et al., 2025) and MemoryBank (Zhong et al., 2024) follow a RAG-style paradigm (Gao et al., 2023). They focus on storing, retrieving, and updating memories. Their main goal is efficient and accurate recall from large memory banks.

Structured memory management. To overcome the limits of flat text, recent work explores structured memory formats. Examples include tree-based methods like MemTree (Rezazadeh et al., 2025) and graph-based methods like A-Mem (Xu et al., 2025). Other studies use hierarchical or compositional designs (Rasmussen et al., 2025; Zhang et al., 2025a) to model dependencies across turns. These structures often improve retrieval precision over plain text chunks.

Lightweight memory systems. Another line of work focuses on memory efficiency under limited token budgets. LightMem (Fang et al., 2026) uses hierarchical compression to summarize dialogue history and reduce storage and retrieval cost. SimpleMem (Liu et al., 2026) further adds semantic lossless compression, structured indexing, and adaptive retrieval to improve token usage. These methods emphasize compact memory and efficient

access. They are suitable for long-term interaction under limited computation.

Discussions. Despite this progress, a key gap remains. Recent surveys (Hu et al., 2025) show that most methods still treat memory as passive storage with fixed representations. Whether they use RAG or trajectory replay, retrieval still depends heavily on surface semantic similarity. This makes it hard to detect hidden conflicts or adapt past experience to changing situations. As a result, these methods may suffer from negative transfer or hallucination. In contrast, ActMem uses an event-centric design that explicitly models logical structure and causal dependency. It turns memory from passive storage into an active reasoning engine.

2.2 Synthetic Evaluation Benchmark

Synthetic benchmarks are widely used to evaluate LLMs in controlled settings. They have moved from simple information retention to more complex reasoning. Recent datasets like RULER (Hsieh et al., 2024) and BABILong (Kuratov et al., 2024) test variable tracking and logical chaining. Datasets like LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025) focus on long-term dialogue consistency and memory retrieval. However, these datasets are mostly QA-oriented. They test passive retrieval more than active use of memory for decision-making. We address this issue with ActMemEval. It evaluates the utility of memory for action via implicit constraints and causal conflicts. We use human verification to improve reliability.

3 The ActMem Framework

We present ActMem to bridge the gap between passive retrieval and active reasoning in LLM agents. As illustrated in Figure 2, the framework consists of four modules: (1) memory fact extraction, (2) fact clustering, (3) memory KG construction, and (4) counterfactual-based retrieval and reasoning. Formally, we define the raw dialogue history as $\mathcal{D} = \{u_1, a_1, u_2, a_2, \ldots, u_t, a_t, \ldots, u_N, a_N\}$, where u_i and a_i denote the user and agent utterances at turn i, respectively. Our goal is to transform $\mathcal{D}$ into a structured memory graph $\mathcal{G}$ to support complex decision-making.

3.1 Memory Fact Extraction

Raw dialogue history often contains redundancy and noise that hinder effective reasoning. To address this, we first compress the raw history $\mathcal{D}$ into a set of atomic memory facts $\mathcal{F}$. We use an LLM-based information extractor to parse the dialogue and generate declarative sentences containing key entities and facts following (Fang et al., 2026). During extraction, we resolve pronouns in person or place mentions to explicit entities, and normalize relative time expressions into absolute timestamps based on the dialogue time. We define an extraction function $f_{\text{ext}}(\cdot)$ powered by an LLM to map raw dialogues into a set of atomic facts $\mathcal{F}$. For each dialogue turn (u_t, a_t), we extract a subset of facts:

$$\mathcal{F}_t = \text{LLM}(p_{\text{ext}}, u_t, a_t) = \{f_{t,1}, f_{t,2}, \ldots, f_{t,k}\}, \quad (1)$$

where p_{ext} denotes the extraction prompt designed to identify entities and events. The global fact set is the union of all turns: $\mathcal{F} = \bigcup_{t=1}^{N} \mathcal{F}_t$. Unlike traditional summarization which compresses the entire context into a single paragraph, our method focuses on *atomicity*. This granular extraction serves as the foundation for subsequent causal mining.

3.2 Fact Clustering

To model long-term interactions, we need to find relations between facts from different sessions. Besides, mining causal edges by computing relations across all pairs of facts in the global history is time-consuming. To mitigate this, we introduce a fact clustering module. Let $\phi(\cdot)$ be an embedding function, e.g., Qwen3-Embedding-8B (Zhang et al., 2025b), that maps a textual fact f_i to a d-dimensional vector $\mathbf{h}_i \in \mathbb{R}^d$. We compute semantic similarity with cosine similarity:

$$\mathbf{h}_i = \phi(f_i), \quad \text{sim}(f_i, f_j) = \frac{\mathbf{h}_i^\top \mathbf{h}_j}{\|\mathbf{h}_i\| \|\mathbf{h}_j\|}. \quad (2)$$

To partition $\mathcal{F}$ into disjoint clusters $\mathcal{C} = \{C_1, \ldots, C_K\}$ efficiently, we employ a single-pass incremental clustering algorithm. We maintain a centroid vector $\mathbf{c}_k$ for each cluster. For each incoming fact f_i, we calculate its similarity with existing centroids. If the maximum similarity exceeds a clustering threshold δ, the fact is assigned to the corresponding cluster; otherwise, a new cluster is initialized with f_i as its centroid. This strategy ensures that facts related to the same topic are consolidated dynamically. It lets us mine causal relations inside local clusters, which reduces computation while keeping topical coherence.

3.3 Memory KG Construction

The core of ActMem is the construction of the memory KG, denoted as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where nodes

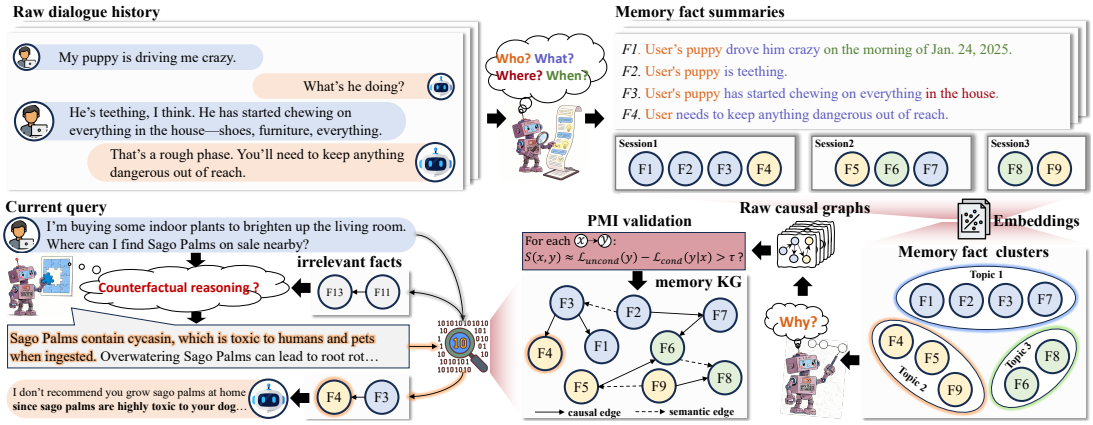


Figure 2: The overview of our proposed framework.

$\mathcal{V}$ correspond to memory facts and edges $\mathcal{E}$ represent causal or semantic relations.

Semantics edge generation. Following clustering, for each fact cluster, we calculate the pairwise similarity between facts and add a semantic edge between any two nodes whose similarity exceeds a semantic threshold τ.

Raw causal edge generation. Within each cluster C_k and each dialogue session, we prompt the LLM to identify potential causal edges. Let $\mathcal{H}_k$ be the set of candidate pairs (f_i, f_j) identified by the model as potentially causal (i.e., $f_i \rightarrow f_j$):

$$\mathcal{H}_k = \text{LLM}(p_{\text{cause}}, C_k). \quad (3)$$

PMI-based hallucination filtering. A major challenge in using LLMs for causal mining is "causal hallucination" where the model identifies spurious correlations as causation. To verify the candidate edges in $\mathcal{H}_k$, we employ a verification mechanism based on Pointwise Mutual Information (PMI). It quantifies the information gain provided by the antecedent fact f_i for generating the consequent fact f_j. We define the verification score S_{PMI} as the reduction in the negative log-likelihood of generating f_j when conditioned on f_i:

$$S_{\text{PMI}}(f_i, f_j) = \mathcal{L}_{\mathcal{M}}(f_j \mid T_{\text{uncond}}) - \mathcal{L}_{\mathcal{M}}(f_j \mid T_{\text{cond}}), \quad (4)$$

where $\mathcal{L}_{\mathcal{M}}(\cdot)$ denotes the cross-entropy loss of the target sequence f_j computed by a pre-trained language model $\mathcal{M}$ (e.g., GPT2-Large). To model the causal dependency, we design specific prompt templates for the input contexts: (1) The *conditional context* T_{cond} constructs the sequence "f_i. *As a result*, f_j"; (2) The *unconditional context* T_{uncond} uses a neutral prefix "*The fact is that* f_j". A positive S_{PMI} means that f_i clearly reduces the uncertainty of f_j. We keep only pairs whose PMI score is above the threshold γ:

$$\mathcal{E}_{\text{causal}} = \{(f_i, f_j) \in \bigcup \mathcal{H}_k \mid S_{\text{PMI}}(f_i, f_j) > \gamma\}. \quad (5)$$

Finally, we merge these validated causal edges with semantic edges to form the final memory graph $\mathcal{G}$.

3.4 Counterfactual Retrieval and Reasoning

Traditional RAG systems retrieve memories solely based on the semantic similarity to the user's current query q, often missing implicit constraints. We propose a retrieval-reasoning-refinement loop to utilize the memory KG effectively.

Step 1: Initial retrieval. We retrieve an initial fact set $\mathcal{V}_{\text{init}} \subset \mathcal{V}$ by vector similarity to the query q:

$$\mathcal{V}_{\text{init}} = \text{TopK}(\text{sim}(q, f_i), f_i \in \mathcal{V}). \quad (6)$$

Step 2: Counterfactual reasoning. We then generate commonsense knowledge with a counterfactual question:

$$k_{cs} = \text{LLM}(p_{\text{counter}}, q, \mathcal{V}_{\text{init}}). \quad (7)$$

p_{counter} can be briefly summarized as "If the user does q, what negative consequences might occur considering $\mathcal{V}_{\text{init}}$ or general knowledge?"

Step 3: Refinement and response. We use k_{cs} as an enhanced query for secondary retrieval over the memory graph. First, we retrieve a seed set of nodes similar to k_{cs}:

$$\mathcal{V}_{\text{seed}} = \text{TopK}(\text{sim}(k_{cs}, f_i), f_i \in \mathcal{V}), \quad (8)$$

We then expand them with their structured graph information. Here, $\mathcal{N}_{\mathcal{G}}(v)$ denotes the neighbor nodes of v in $\mathcal{G}$. The refined set is

$$\mathcal{V}_{\text{refine}} = \mathcal{V}_{\text{seed}} \cup \bigcup_{v \in \mathcal{V}_{\text{seed}}} \mathcal{N}_{\mathcal{G}}(v). \quad (9)$$

The final context is $\mathcal{C}_{\text{final}} = \mathcal{V}_{\text{init}} \cup \{k_{cs}\} \cup \mathcal{V}_{\text{refine}}$. The agent generates ans given this context:

$$ans = \text{LLM}(p_{\text{answer}}, q, \mathcal{C}_{\text{final}}). \tag{10}$$

With this mechanism, ActMem turns the agent from a passive retriever to an active reasoner, ensuring the utility of memory for decision-making.

4 The ActMemEval Benchmark

In this section, we introduce the pipeline to construct our ActMemEval and the evaluation results.

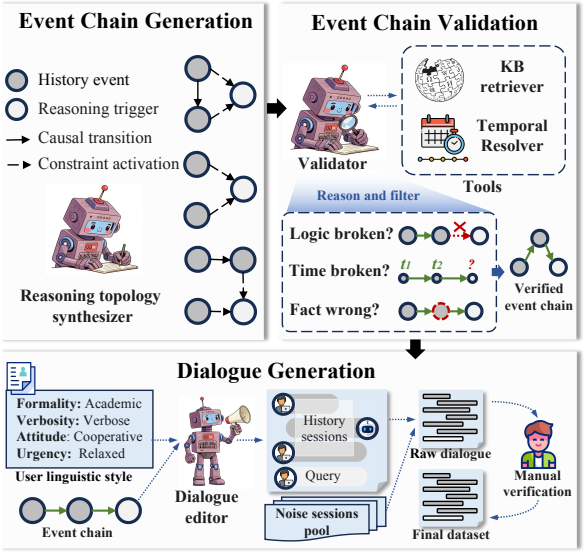


Figure 3: Pipeline to construct our ActMemEval dataset.

4.1 Dataset Construction

Figure 3 shows the construction pipeline, which has three steps. To avoid model self-bias and samples that are easy for the same model to solve, we use different models for data generation and validation. The generation model is also stronger than the model used in our method, which helps improve data quality. Specifically, we use Gemini 3.1 Pro for generation and Claude Sonnet 4.5 for validation. To improve diversity, we synthesize six categories: safety-health risks, feasibility limits, time-space-procedure mismatches, access or availability gaps, preference incompatibility, and benefit-reuse opportunities. Detailed definitions are in Appendix E.

Step 1: Event Chain Generation. Unlike previous datasets that generate dialogues directly, we first synthesize the underlying event logic. We employ an advanced LLM (Gemini 3.1 Pro) as the event chain generator. It generates two types of nodes: history events and reasoning triggers. For each sample, we randomly choose one category and generate the event chain accordingly. History events describe the background facts, such as "User adopted a puppy" and "Puppy is teething". Reasoning triggers represent the user's current intent, such as "User wants to buy Sago Palms". The reasoning trigger links the history events and forms a complete reasoning chain. In later dialogue generation, history events are converted into historical dialogues, while the reasoning trigger is converted into the current user query.

Step 2: Event Chain Validation. We then validate each chain from three aspects: logic, time, and knowledge. Claude Sonnet 4.5 checks logical consistency directly. For time and knowledge validation, we provide two external tools: *Temporal Resolver* and *KB retriever*. The validator calls them when needed. *Temporal Resolver* normalizes time expressions into ISO dates and checks temporal order. *KB retriever* queries Wikidata and Wikipedia and returns evidence snippets. The validator keeps only chains that are coherent, temporally valid, and knowledge grounded.

Step 3: Dialogue Generation. We convert the validated topologies into natural-language dialogues using a dialogue editor. To reflect real-world diversity, we define a set of user linguistic styles along four dimensions: formality, verbosity, attitude, and urgency. For each sample, we instantiate every node in the event chain as an anchor session. The user in each anchor session follows one randomly sampled style combination. Each anchor session reflects the facts in its corresponding node. We then randomly sample noise sessions from UltraChat (Ding et al., 2023) and insert them between the anchor sessions to form a raw dialogue. Finally, three NLP researchers manually verify each sample. We keep only the samples with no disagreement.

4.2 Dataset Evaluation

We evaluate the dataset from two aspects: context structure and query-history similarity. We compare our ActMemEval with the retrieval-based benchmarks LongMemEval (Wu et al., 2025) and LoCoMo. Table 1 reports the basic statistics, and Figure 4 shows the similarity distribution between each question and its supporting history context.

Table 1 shows that ActMemEval is structurally denser. It has the largest average number of sessions and much longer contexts. Its evidence sessions are more numerous and appear later in

Dataset	#Q	Avg. Sess.	Avg. Tok.	Avg. Evid. Sess.	Avg. Evid. Pos.
ActMemEval	327	67.25	120k	2.48	34.53
LongMemEval	500	47.73	115k/1.5M	1.90	25.16
LoCoMo	7412	27.20	10k	1.29	14.14

Table 1: Dataset statistics of ActMemEval and two related benchmarks. #Q is the number of questions. Avg. Sess. is the average number of sessions. Avg. Tok. is the average number of tokens. Avg. Evid. Sess. is the average number of evidence sessions. Avg. Evid. Pos. is the average position of evidence sessions.

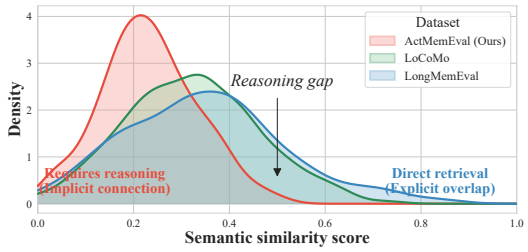


Figure 4: Distribution of semantic similarity scores between each question and its supporting history context in ActMemEval, LoCoMo, and LongMemEval.

the dialogue. Compared with LongMemEval and LoCoMo, ActMemEval spreads supporting clues across more sessions.

Figure 4 compares semantic similarity distributions. ActMemEval has the lowest average similarity (0.232) and LongMemEval is highest at 0.347. The ActMemEval curve is shifted left, which means the question is often semantically far from the key history. In other words, the evidence cannot be found by simple similarity retrieval. Thus, ActMemEval is less retrieval-oriented and better tests memory reasoning.

5 Experiments

5.1 Baselines

We compare ActMem with six representative baselines: NaiveRAG, MemoryBank (Zhong et al., 2024), Mem0 (Chhikara et al., 2025), A-Mem (Xu et al., 2025), SimpleMem (Liu et al., 2026), and LightMem (Fang et al., 2026). Detailed descriptions are provided in Appendix A.

5.2 Evaluation Metrics

We employ two key metrics to assess both the *precision of recall* and the *effectiveness of reasoning*:

- **Retrieval accuracy:** It tests whether the agent can recall the memory facts needed to answer the query. Following (Chen et al., 2025), retrieval accuracy is the recall rate of ground-truth evidence facts in our dataset.
- **Question-answering accuracy:** This metric measures the final utility of the memory. It assesses whether the generated response correctly addresses the user intent while adhering to historical constraints.

5.3 Implementation Details

We employ two LLMs as backbones for reasoning and generation: DeepSeek-V3, representing high-performance open-source models, and GPT-4o-mini, representing efficient lightweight models. For memory representation, we utilize Qwen3-Embedding-8B (Zhang et al., 2025b) to generate semantic vectors. For hyperparameter settings, in the event clustering module, we set the clustering threshold δ to 0.2. The semantic-edge threshold τ is set to 0.3. For causal edge construction, the PMI validation threshold γ is empirically set to 0.2. During the retrieval phase, we set the number of retrieved memory facts to 20 for the initial retrieval step. The number of counterfactual facts retrieved is set to 10. In the refinement step, we dynamically retrieve additional facts based on the generated KG. For the baseline configuration, we preserve each method's original final QA prompt and make minimal incremental modifications. Specifically, we append brief instructions on reasoning to better adapt the baselines to our dataset. This setting keeps the original prompting design intact while making the required reasoning process more explicit. We use DeepSeek-V3 as the judge to evaluate retrieval results and generated responses, and we further verify its judgments through an agreement check with human annotation; see Appendix F.

5.4 Main Results

Table 2 presents the results on ActMemEval. Although NaiveRAG achieves the highest retrieval accuracy by recalling extensive surface-level semantic matches, its inferior QA performance indicates a failure to synthesize these facts with noise to deduce the implicit logical constraints required for QA. Compared to LightMem, ActMem achieves substantial improvements. Specifically, when using DeepSeek-V3, it substantially improves QA accuracy over LightMem while also delivering consistently stronger retrieval quality across different categories. A similar trend is observed with GPT-4o-mini. In contrast, retrieval-based baselines exhibit relatively poor performance. We attribute their failure to the reasoning gap. These methods

LLM	Method	Overall Ret / QA	SHR Ret / QA	FL Ret / QA	TSP Ret / QA	AA Ret / QA	PI Ret / QA	BRO Ret / QA
DeepSeek-V3	NativeRAG	**85.57** / 61.79	**79.89** / 63.22	**80.85** / 55.32	**92.86** / 66.67	**95.45** / 68.18	**91.67** / 53.33	**86.11** / 66.67
	MemoryBank (Zhong et al., 2024)	44.37 / 44.59	46.25 / 47.50	33.72 / 34.88	53.57 / 57.14	50.00 / 42.86	35.19 / 33.33	47.22 / 44.44
	A-Mem (Xu et al., 2025)	46.30 / 32.27	48.15 / 41.10	39.29 / 18.92	52.70 / 29.41	44.44 / 35.29	51.79 / 37.04	40.54 / 25.00
	Mem0 (Chhikara et al., 2025)	78.07 / 41.80	73.84 / 44.19	78.72 / 34.04	88.37 / 51.16	75.00 / 45.45	82.76 / 48.28	67.65 / 11.76
	SimpleMem (Liu et al., 2026)	52.63 / 61.54	49.43 / 60.92	48.94 / 53.19	56.98 / 65.12	50.00 / 59.09	51.67 / 60.00	72.22 / **83.33**
	LightMem (Fang et al., 2026)	57.29 / 62.35	60.34 / 67.82	53.19 / 53.19	55.81 / 65.12	54.55 / 50.00	55.00 / 63.33	63.89 / 66.67
	ActMem	71.66 / **76.52**	71.51 / **82.56**	71.28 / **72.34**	75.58 / **86.05**	68.18 / **72.73**	66.67 / **66.67**	80.56 / 61.11
GPT-4o-mini	NativeRAG	**86.03** / 34.01	**80.46** / 35.63	79.79 / 19.15	**94.19** / 39.53	**93.18** / 45.45	**96.67** / 40.00	**83.33** / 27.78
	MemoryBank (Zhong et al., 2024)	32.32 / 33.49	25.00 / 31.58	25.00 / 31.71	33.82 / 42.86	41.18 / 33.33	24.07 / 30.00	63.57 / 33.33
	A-Mem (Xu et al., 2025)	45.29 / 18.42	38.06 / 22.73	34.72 / 13.89	48.61 / 13.89	45.83 / 33.33	58.33 / 12.50	71.88 / 18.75
	Mem0 (Chhikara et al., 2025)	78.95 / 31.10	79.58 / 38.03	**81.71** / 29.75	80.77 / 30.77	72.50 / 25.00	88.10 / 33.33	61.76 / 5.88
	SimpleMem (Liu et al., 2026)	54.93 / 40.36	52.44 / 48.78	46.34 / 26.83	60.53 / 42.11	57.50 / 45.00	51.79 / 32.14	82.14 / **35.71**
	LightMem (Fang et al., 2026)	63.36 / 40.49	63.79 / 44.83	58.51 / 27.66	67.44 / 41.86	63.64 / **54.55**	63.33 / 40.00	63.89 / 29.11
	ActMem	76.92 / **66.80**	75.29 / **79.31**	77.66 / **65.96**	81.40 / **74.42**	77.27 / 50.00	73.33 / **53.33**	77.78 / 33.33

Table 2: Performance comparison on ActMemEval. We report retrieval accuracy (Ret) and QA accuracy (QA) for the overall dataset and for each category: safety-health risks (SHR), feasibility limits (FL), time-space-procedure mismatches (TSP), access or availability gaps (AA), preference incompatibility (PI), and benefit-reuse opportunities (BRO). Best results in each backbone block are highlighted in **bold**, and second-best results are underlined.

rely primarily on surface-level semantic matching, which struggles to retrieve historical events that are logically related but semantically distant from the current query. For structured memory baselines, A-Mem underperforms ActMem because it uses the graph merely as a retrieval aid, while ActMem employs it as a reasoning framework with explicit causal and semantic edges. In summary, the ability of ActMem to "reason before retrieving" allows it to identify implicit constraints that other methods miss, thereby achieving state-of-the-art performance in complex, action-oriented scenarios. A qualitative case study is provided in Appendix B.

We also observe an interesting gap across different LLM backbones. GPT-4o-mini attains high retrieval accuracy, yet its QA accuracy is markedly lower. In contrast, DeepSeek-V3 delivers balanced and consistently superior results on both metrics. We attribute this discrepancy to a reasoning bottleneck inherent in lightweight LLMs. The lower QA accuracy suggests that GPT-4o-mini struggles to integrate and reason over these memory clues. This phenomenon also reflects the difficulty of ActMemEval that requires reasoning. With stronger reasoning ability, DeepSeek-V3 can use the retrieved context to resolve conflicts, while weaker LLMs may still fail even with correct evidence.

5.5 Ablation Study

To verify the contribution of each module, we design three variants by removing key modules:

- **w/o Reasoning:** This variant removes the *counterfactual reasoning* module. The agent retrieves memory facts directly from the memory KG without intermediate reasoning.
- **w/o Causal:** It constructs the memory KG excluding *causal edges*. The graph uses only semantic similarity to connect nodes.
- **w/o Semantic:** This variant constructs the memory KG excluding *semantic edges*. The graph consists only of extracted causal chains, making it a sparse causal network.

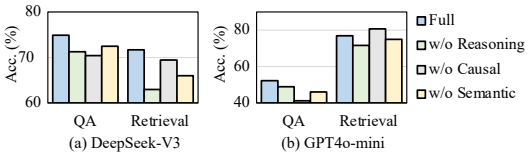
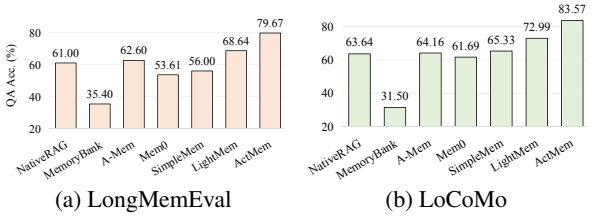


Figure 5: Ablation study on counterfactual reasoning, causal and semantic edge construction.

Figure 5 presents the ablation results. First, removing the reasoning module (*w/o Reasoning*) with DeepSeek-V3 causes the largest drop in retrieval accuracy (about 9%), confirming our core motivation. In complex scenarios, relevant memories (e.g., "owning a puppy") may be semantically distant from the query (e.g., "buying plants"). Without counterfactual reasoning and commonsense completion, the agent fails to uncover such "hidden" evidence. Second, removing causal edges (*w/o Causal*) yields the lowest QA accuracy. Even when some facts are retrieved, the absence of explicit causal links prevents the agent from forming coherent reasoning chains, leading to generic rather than conflict-aware decisions. We observe that when



(a) LongMemEval (b) LoCoMo

Figure 6: Performance on conventional datasets.

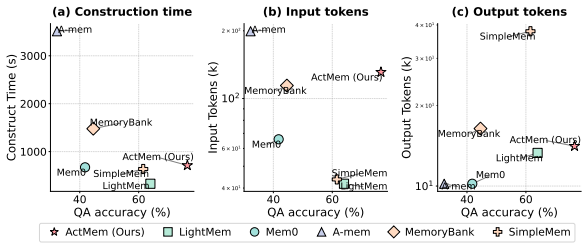


Figure 7: Efficiency comparison using GPT-4o-mini.

using GPT-4o-mini, removing causal edges actually yields better retrieval performance than the full model. We attribute this to the weaker reasoning capability of GPT-4o-mini, which may introduce errors during world-knowledge completion, adding more noise and degrading the retrieval quality of the full model. Third, the *w/o Semantic* variant also reduces retrieval accuracy, indicating that semantic edges serve as crucial connectors. They ensure that topic-related but non-causal facts remain reachable, preserving the completeness of the retrieved context. In summary, causal edges provide logical backbone, semantic edges ensure coverage, and the reasoning module guides inference. Their combination gives ActMem strong performance. Additional analyses of PMI validation and semantic threshold sensitivity are in Appendices C and D, respectively.

5.6 Results on Conventional Datasets

To test the generalization ability of ActMem, we run additional experiments on LongMemEval (Wu et al., 2025) (LongMemEval-S) and LoCoMo (Maharana et al., 2024). For fairness, we use GPT-4o-mini as the backbone LLM for all methods and keep the evaluation setting consistent. ActMem uses the same parameters on all three datasets, without dataset-specific tuning. As shown in Figure 6(a), ActMem achieves the best QA accuracy on LongMemEval. It outperforms the strongest baseline LightMem, and also exceeds NativeRAG and A-Mem. Figure 6(b) shows a similar trend on LoCoMo. ActMem reaches 83.57%, again ranking first among all methods. These consistent gains on both datasets indicate that our structured memory KG is also effective in conventional retrieval-oriented settings. By organizing scattered memories into connected fact structures, ActMem reduces retrieval noise and helps the model produce more accurate final answers.

5.7 Efficiency Comparison

Figure 7 compares efficiency and QA accuracy from multiple views. Subfigure (a) shows that heavier frameworks such as A-Mem and MemoryBank have much higher construction cost, but still do not achieve competitive performance. In contrast, ActMem achieves the best accuracy with moderate extra overhead. Although it needs more construction time than lightweight methods such as LightMem and SimpleMem, it is still much more efficient than the most expensive baselines. The second and third subfigures show the same trade-off in token usage. ActMem uses more input tokens than compression-based methods because building the causal graph needs more context. However, this extra cost brings clear gains in answer quality. At the same time, its output token usage stays close to LightMem and is much lower than SimpleMem. This suggests that the gain does not come from longer responses. Overall, ActMem gives a good balance between reasoning performance and computational cost. It adds moderate overhead for causal and semantic reasoning, and converts that cost into strong performance.

6 Conclusion and Future Work

In this paper, we identify a key limitation of current agent memory systems: the gap between memory retrieval and reasoning. We argue that high retrieval accuracy is not enough. Agents must not only recall past information, but also understand what it means for the present. To address this issue, we propose ActMem, which transforms passive dialogue history into an active memory KG with causal and semantic edges. We also introduce ActMemEval to evaluate the utility of memory for action. Experiments on both our dataset and standard benchmarks show that ActMem outperforms baselines.

In future work, we plan to explore lightweight LLMs to improve KG construction efficiency without hurting reasoning quality. We also plan to extend ActMem to multimodal scenarios and study fine-tuning strategies to improve causal reasoning.

Limitations

Our method still has several limitations. First, ActMem relies on LLM-based fact extraction, commonsense completion, and causal analysis, so errors in these stages may propagate to the final graph and response. Second, constructing and refining the memory KG introduces additional time and token cost compared with simpler compression-based methods. Third, ActMemEval focused on six predefined categories and cannot fully cover the diversity of real-world long-term interactions.

References

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. 2025. HaluMem: Evaluating hallucinations in memory systems of agents. *CoRR*, abs/2511.03506.

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

Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 3029–3051.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2026. Lightmem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Gaole He, Gianluca Demartini, and Ujwal Gadiraju. 2025. *Plan-Then-Execute: An Empirical Study of User Trust and Team Performance When Using LLM Agents As A Daily Assistant*.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? In *CoLM*.

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

Yuri Kuratov, Aydar Bulatov, Petr Anokhin, Dmitry Sorokin, Artyom Y. Sorokin, and Mikhail Burtsev. 2024. In search of needles in a 11M haystack: Recurrent memory finds what LLMs miss. *CoRR*, abs/2402.10790.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. 2026. Simplemem: Efficient lifelong memory for LLM agents. *CoRR*, abs/2601.02553.

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*.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: A temporal knowledge graph architecture for agent memory. *CoRR*, abs/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 *ICLR*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. LongMemEval: Benchmarking chat assistants on long-term interactive memory. In *ICLR*.

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

Chengyuan Yang, Zequn Sun, Wei Wei, and Wei Hu. 2026. Beyond static summarization: Proactive memory extraction for llm agents. *CoRR*, abs/2601.04463.

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

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025c. A survey on the memory mechanism of large language model-based agents. *ACM Trans. Inf. Syst.*, 43(6):155:1–155:47.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. In *AAAI*, pages 19724–19731.

A Baseline Details

We compare ActMem with six representative baselines:

- **NaiveRAG:** The standard baseline that retrieves the top-k relevant dialogue snippets based on vector similarity and concatenates them into the context window.
- **MemoryBank** (Zhong et al., 2024): A summary-based memory framework that stores historical interactions as evolving summaries with a forgetting-aware update mechanism.
- **Mem0** (Chhikara et al., 2025): A fact-level memory framework for personalized agents, focusing on continuous memory storage, updating, and retrieval.
- **A-Mem** (Xu et al., 2025): An agentic memory framework that models relations between memories and emphasizes association-driven retrieval.
- **SimpleMem** (Liu et al., 2026): A lightweight lifelong memory framework based on semantic lossless compression, structured indexing, and adaptive retrieval.
- **LightMem** (Fang et al., 2026): A lightweight memory framework that uses hierarchical compression and summary-based retrieval to reduce token cost.

B ActMem Case Study

Table 3 presents a case study from our dataset. The user asks for a location to buy "Sago Palms." Although the query appears simple, it contains a critical implicit constraint in the dialogue history: the user owns a puppy that is "teething" and "chewing on everything." As Sago Palms are toxic to dogs, an intelligent agent should detect this risk and issue a warning. LightMem fails to retrieve the relevant evidence about the puppy, instead surfacing irrelevant noise such as "high-end hotels" and "cooking meals." It treats the query as a straightforward information-seeking task and directly recommends where to buy the plants. In contrast, ActMem successfully bridges this reasoning gap. Through the counterfactual reasoning, the agent first infers the potential consequence of the user's intent (i.e., that Sago Palms are toxic). This knowledge becomes a semantic anchor that guides retrieval of the specific memory that the user's puppy is chewing on everything. By integrating this retrieved history with world knowledge, ActMem identifies the conflict and produces a safety-aware response, advising against the purchase. This case highlights our ability to transform memory from passive storage into an active reasoning engine for robust decision-making.

C PMI Verification Case Study

To further illustrate the effect of PMI validation, we randomly sampled five positive examples and five negative examples from candidate causal pairs. Tables 4 and 5 report their PMI scores. The positive examples all exceed the threshold $\gamma = 0.2$ and correspond to clear causal relations, while the negative examples remain near zero or below zero and are filtered out. This case study shows that PMI helps suppress weak or spurious causal pairs.

D Analysis of Semantic Similarity Thresholds

We investigate the impact of the semantic similarity threshold on QA accuracy and graph density metrics. As presented in Table 6, we observe a clear correlation between graph connectivity and QA accuracy. A lower threshold (e.g., 0.3) yields the optimal QA accuracy of 76.52%. This improvement stems from the richer retrieved context: a relaxed threshold preserves more semantic edges, leading to a substantially higher average node degree (20.48) and more retrieved KG facts (5.21). The resulting dense connectivity provides the agent with broader semantically relevant background information, reducing the risk of missing critical clues. In contrast, a stricter threshold (e.g., 0.7) produces a much sparser graph, with the average degree dropping to 2.34. Such aggressive pruning removes potential semantic associations, resulting in insufficient context retrieval (only 0.29 facts on average) and a corresponding decline in QA accuracy. These results highlight the trade-off between graph sparsity and information recall.

E Dataset Categories

This section defines the six categories used in ActMemEval. Each category reflects a different hidden constraint in the user's current request.

- **Safety-Health Risks:** the action may cause health, safety, toxic, material, or environmental

	Query	History	ActMem	Retrieved memory	Reasoning
LightMem	I'm buying some indoor plants to brighten up the living room. Where can I find "Sago Palms" on sale nearby?	[User]: My puppy is driving me crazy. [Assistant]: What's he doing? [User]: He's teething, I think. He has started chewing on everything in the house—shoes, furniture, everything. [Assistant]: That's a rough phase. You'll need to keep anything dangerous out of reach.	I wouldn't recommend a Sago Palm... You mentioned last week that your new puppy is chewing on everything, and Sago Palms are extremely toxic—often fatal—to dogs if ingested. Can I suggest a pet-safe alternative ... ?	User is looking for high-end hotels in Narva that offer spa services or other luxury amenities. User wants to make the most of their time in Narva and is seeking suggestions ...	... This is a great way to brighten up your living room and improve your environment! Given that you've been ... planning a trip to Narva and cooking healthier meals ... Here are some steps you can take to find Sago Palms nearby: ...
ActMem	I'm buying some indoor plants to brighten up the living room. Where can I find "Sago Palms" on sale nearby?	[User]: My puppy is driving me crazy. [Assistant]: What's he doing? [User]: He's teething, I think. He has started chewing on everything in the house—shoes, furniture, everything. [Assistant]: That's a rough phase. You'll need to keep anything dangerous out of reach.	I wouldn't recommend a Sago Palm... You mentioned last week that your new puppy is chewing on everything, and Sago Palms are extremely toxic—often fatal—to dogs if ingested. Can I suggest a pet-safe alternative ... ?	User is interested in creative ways to use zucchini in cooking. User's puppy has started chewing on everything in the house, including shoes and furniture. ...	... The user's request to buy Sago Palms directly conflicts with their current situation: The user has a puppy. World knowledge indicates that Sago Palms are highly toxic to dogs and cats ...

Table 3: A case study from our ActMemEval benchmark.

PMI	Cause	Effect
0.9163	User had a scary reaction where their hands swelled up like balloons after cleaning the bathroom yesterday.	User's doctor advised them to banish all natural rubber from the house due to a reaction.
0.4278	User's iron levels are critically low, and their supplements are not absorbing effectively.	User's doctor advised them to eat steak twice a week to avoid needing hospital infusions.
1.4725	User's grandmother's mobility has worsened after a fall, and she is now permanently using a wheelchair.	User's family is adjusting travel arrangements for their grandmother due to her permanent use of a wheelchair.
1.3756	The railway workers' union announced a total 24-hour strike for tomorrow.	No trains will be running tomorrow due to the railway workers' union strike.
1.2329	User's professor emailed that all proctored exams must go to the North Campus facility during the downtown closure.	User needs to book a slot at the North Campus facility for May 18th at 9 AM.

Table 4: Randomly sampled positive examples retained by PMI validation.

risk.

- **Feasibility Limits:** the action is impractical or cannot be carried out in reality.
- **Time-Space-Procedure Mismatches:** the action conflicts with time, place, procedure, permission, policy, or order.
- **Access or Availability Gaps:** a required resource, account, file, tool, item, or system is unavailable.
- **Preference Incompatibility:** the plan conflicts with stable preferences, habits, soft constraints, or relationship context.
- **Benefit-Reuse Opportunities:** an existing resource, right, reusable asset, or better option already exists.

Representative cases are shown below.

Safety-Health Risks

Question. It's cold in here. Search for tips on how to build a long-lasting wood fire in the fireplace.

History events.
- Vet warned me that smoke can kill the cockatiel instantly.
- Threw out all my non-stick pans and scented candles.
- Checking the air quality monitor constantly for the bird's safety.

Gold answer. Urgent Stop; warn that smoke is fatal to the new parrot; suggest checking the thermostat instead.

Feasibility Limits

Question. My nephew's birthday is coming up. Can you recommend some popular Lego sets suitable for a 12-year-old?

History events.
- Nephew broke his arm and is in a cast.
- He is bored because he cannot use his right hand well.

Gold answer. Legos might be a bit frustrating for him right now. Maybe a digital gift card, an audiobook, or a strategy board game would be better.

Time-Space-Procedure Mismatches

Question. Schedule a team brainstorming workshop in Conference Room B for next Tuesday at 2 PM.

History events.
- IT announced projector replacement next Tuesday afternoon.
- The workshop requires a projector.

Gold answer. That time slot will not work for a presentation. Conference Room B is scheduled for projector replacement maintenance.

Access or Availability Gaps

Question. I'm locked out of my Apple ID. Guide me on how to use the "Recovery Key" method to reset my password.

History events.

PMI	Cause	Effect
-0.0655	User used 2 tablespoons of vegetable oil.	User thinks the vegetable stir-fry recipe sounds delicious and is seeking a protein option to pair with it, as well as tips for making the veggies really crispy.
0.0738	The application will have a smooth, modern user interface.	User is seeking more information regarding the customizable intervals in the time-lapse photography application.
0.0859	User found the comprehensive guide on how to make a room diffuser very helpful.	User wants warm and cozy essential oil combinations to help create a relaxing atmosphere in their home.
0.0646	User's doctor said they have to eat a steak twice a week or they will need hospital infusions.	User felt like a traitor to the cause on March 22, 2023.
0.0923	User is really excited to get involved in reforestation efforts.	User is going to start researching local organizations and events to make a difference in their community.

Table 5: Randomly sampled negative examples filtered out by PMI validation.

Threshold	QA acc.	Avg. node degree	Avg. retrieved KG facts
0.3	76.52	20.48	5.21
0.5	74.89	5.07	1.24
0.7	75.30	2.34	0.29

Table 6: QA accuracy and graph density metrics with different semantic similarity thresholds.

Metric	κ_{hh}	κ_{hl}
Retrieval accuracy	0.8956	0.8336
QA accuracy	0.8780	0.8410

Table 7: Cohen's Kappa scores for human–human and human–LLM agreement.

- User generated a Recovery Key and printed it on paper.
- User shredded all old loose papers during an office cleanup.

Gold answer. That might be a problem. Unless you saved a digital copy, you might have destroyed the key.

Preference Incompatibility

Question. Book a table for two at "The Butcher's Block" steakhouse.

History events.
- New partner is strict vegetarian.
- Meat-only venues should be avoided.

Gold answer. Suggest an alternative restaurant with good vegetarian options for the new partner.

Benefit-Reuse Opportunities

Question. I'm going to buy the "Shadows of Rose" DLC for Resident Evil Village on the PS5 store right now.

History events.
- User bought the Gold Edition disc.
- User redeemed the code inside the box.

Gold answer. You do not need to buy that. The Gold Edition already includes the DLC.

F Human–LLM Agreement

For both retrieval accuracy and QA accuracy, we conduct a stratified sampling evaluation and randomly select 80 samples for each metric. We then report Cohen's Kappa between two human annotators (κ_{hh}) and between the final human labels and the LLM judge (κ_{hl}), as shown in Table 7.

G Artifact License and Usage

The ActMemEval dataset will be released under CC BY 4.0 and is intended for academic research use only.