BEYOND QUERY-MEMORY MATCHING: PATH-AWARE ANCHOR-GUIDED RETRIEVAL FOR LONG-TERM AGENT MEMORY

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-based agents often rely on Retrieval-Augmented Generation (RAG)-style memory systems to leverage historical experiences. However, existing methods typically rely on direct query–memory similarity, implicitly assuming that gold evidence memories are close to the query in embedding space. This assumption overlooks the semantic gap between user queries and fine-grained agent memories, making query-memory similarity alone insufficient for reliable evidence recall. To address this limitation, we propose PAGR, a **P**ath-Aware **A**nchor-**G**uided Evidence **R**etrieval framework. Inspired by the spreading activation mechanism in human associative memory, PAGR first retrieves a small set of seed memories as initial anchors. It then expands from these anchors through both global semantic associations and local temporal neighbors. A path-aware cross-encoder selects candidate memories that complement the current retrieval path. The selected memories are accumulated as evidence and reused as anchors for subsequent hops, enabling iterative memory exploration. Extensive experiments on the LoCoMo benchmark demonstrate that PAGR consistently outperforms all baselines, achieving the highest overall F1 scores of 45.2 with GPT-4o-mini and 41.8 with Llama-3.1-8B-Instruct, while maintaining token efficiency.

1 INTRODUCTION

Large language model (LLM)-based agents perform well across diverse tasks Park et al. (2023) Wang et al. (2023), yet they still struggle with long-term interactions that require coherence, continuity, and personalization Packer et al. (2023). A key limitation is that current LLMs condition on a bounded context window, making it difficult to reliably access past experiences, user preferences, and task-specific details across previous sessions Zhang et al. (2025). Although long-context LLMs can include more history, they are computationally expensive and may still be distracted by irrelevant information Liu et al. (2024). As a result, long-term agents require external memory mechanisms that can not only store past interactions, but also retrieve the right memories when needed.

Human memory, on the other hand, provides a natural mechanism for preserving and retrieving experiences accumulated over long periods of time Anderson (1983) Murre & Dros (2015). This motivates an important question for LLM-based agents: can agents be equipped with memory mechanisms that support long-term adaptation and personalization? Recent studies have explored external memory as a way to address this challenge. For example, Zhong et al. (2024) stores user-related information from past interactions and updates memory strength using the Ebbinghaus forgetting curve; Chhikara et al. (2025) extracts and consolidates salient information from conversations to support efficient personalized memory; and Xu et al. (2026) organizes memories as structured, dynamically updated notes linked to related historical memories. These methods demonstrate that effective memory mechanisms are critical for enabling agents to selectively preserve, retrieve, and utilize long-term information for coherent and personalized interactions.

Despite their effectiveness, most agent memory methods still rely on a retrieval-augmented generation (RAG) paradigm Lewis et al. (2020). As illustrated in Figure 1(left), they retrieve a fixed set of top-k memories based on query–memory embedding similarity, concatenate them into the context, and generate the response. However, this strategy is particularly problematic for agent memory retrieval.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

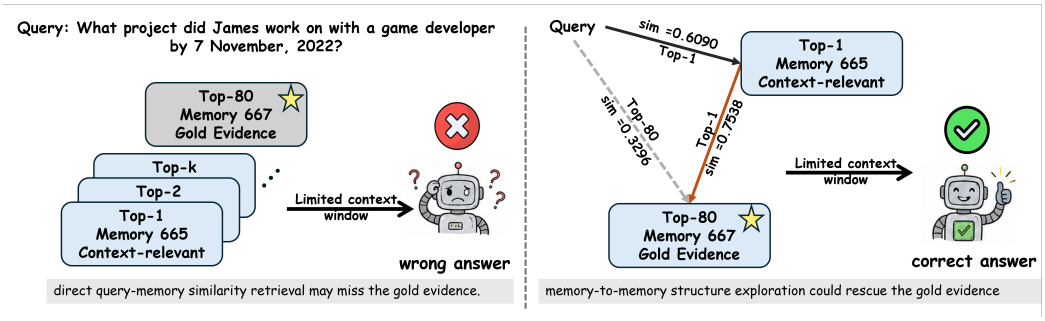


Figure 1: From direct query–memory matching to anchor-guided memory exploration. Direct top-k retrieval may miss the gold evidence memory when it has low similarity to the query. By retrieving a related seed memory as an intermediate anchor and expanding through memory–memory associations, the system can recover the missing evidence for answer generation.

Unlike conventional RAG corpora, where documents are often relatively heterogeneous and distinguishable, agent memories are usually fine-grained, episodic, temporally correlated, and semantically overlapping Hu et al. (2026). Moreover, we observed that a semantic gap exists between the query and the memory representation, making direct query-memory embedding similarity insufficient for reliable recall. Figure 2 illustrates this failure mode on LoCoMo. Across all four categories of the LoCoMo dataset Maharana et al. (2024), many gold memories are missing from the top-10 results, and even increasing the retrieval budget to top-50 does not fully resolve the problem. This suggests that similarity-based retrieval may either miss the gold memory or retrieve it together with many irrelevant and redundant memories. Such noisy context can distract the agent and degrade response quality (Liu et al., 2024).

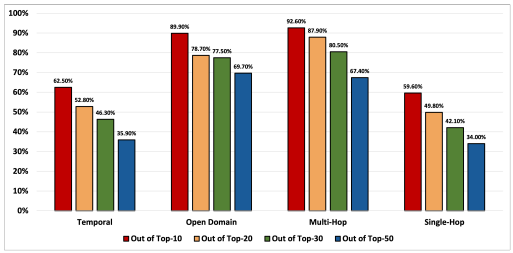


Figure 2: Percentage of queries whose mean rank of gold memories falls outside Top-K under direct query-memory embedding retrieval on LoCoMo. Higher values indicate weaker evidence recall, highlighting the semantic gap between user queries and fine-grained agent memories.

To address this limitation, as illustrated in Figure 1(right), we move beyond one-shot query-memory matching and model memory retrieval as a path-aware anchor-guided expansion process over the memory-memory structure. As discussed in cognitive psychology and associative memory studies, human memory retrieval is often cue-dependent and associative rather than purely direct: a person may fail to recall a target event from the initial cue, but recover it after first recalling a related contextual event Tulving & Thomson (1973). This is consistent with the spreading activation mechanism in human associative memory, where an activated memory can serve as an intermediate cue and trigger related memories through associative links (Collins & Loftus, 1975; Anderson, 1983). Inspired by this mechanism, our method first retrieves a set of query-relevant memories as contextual anchors. Rather than treating these anchors as final evidence, we use them as entry points to explore semantically and temporally related memories, recovering evidence that is missed by direct query-memory similarity. To improve precision, we further design a path-aware selection mechanism. Each candidate memory is evaluated not only by its relevance to the query, but also by its connection to the parent memory that activates it. In this way, a neighbor memory is selected for subsequent expansion by its contribution along the retrieved memory path, rather than treating each memory as an isolated retrieval unit.

Our contributions are:

- We identify a key retrieval limitation in long-term agent memory: direct query-memory similarity often fails to recover gold evidence due to the semantic gap between high-level queries and fine-grained, overlapping episodic memories.
- We propose PAGR, a path-aware anchor-guided retrieval framework that moves beyond one-shot query-memory matching by iteratively exploring the memory-memory structure through

global semantic memory-memory associations and local temporal continuity, enabling both global associative exploration and local episodic context recovery.

- We introduce a path-aware reranking mechanism that selects memories based on their relevance to the query and their contribution along the current retrieval path, reducing redundancy and noise.

2 Related Work

2.1 Long-term agent memory

Recent studies have explored various mechanisms for equipping LLM agents with long-term memory. Some works are inspired by human cognition. For example, MemoryBank extracts event summary and personality summary and updates memory using Ebbinghaus Forgetting Curve theory Zhong et al. (2024). A-MEM adopts a Zettelkasten-inspired design to construct structured memory notes and dynamically link related historical memories Xu et al. (2026). Nemori is inspired by Event Segmentation Theory and Free-energy Principle to autonomously organize and evolve memory Nan et al. (2025). Another line of work focuses on system-oriented memory management. Systems such as Mem0 Chhikara et al. (2025), MemoryOS Kang et al. (2025), LangMem LangChain (2025), and Zep Rasmussen et al. (2025), focus on scalable memory extraction, consolidation, hierarchical storage, modular memory management, or temporal knowledge graph construction. In addition, learning-based methods such as Memory-R1 use reinforcement learning to train agents to manage and utilize memory Yan et al. (2025). Despite their effectiveness, these methods mainly focus on memory construction, organization, updating, or management, while the retrieval failure caused by the semantic gap between query representations and memory representations remains underexplored.

2.2 Retrieval-Augmented Generation

Many agent memory systems adopt RAG-style retrieval to ground answer generation Lewis et al. (2020) Karpukhin et al. (2020). Recent variants move beyond single-step similarity retrieval through graph structures Edge et al. (2024) Gutiérrez et al. (2024), iterative retrieval Trivedi et al. (2023), or agentic workflows Singh et al. (2025). While these RAG variants share our motivation of moving beyond single-step similarity retrieval, they are designed mainly for heterogeneous external corpora, which differs from long-term agent memory. Agent memories are typically fine-grained, temporally dependent, and highly overlapping Hu et al. (2026). Moreover, agent memory exhibits a severe query-memory semantic gap: queries may ask about high-level user states, preferences, or attributes, while the supporting evidence is stored as concrete episodic interactions with limited lexical or semantic overlap. To address this issue, our method proposes a contextual-anchor-guided expansion process over the memory-memory structure instead of constructing an explicit knowledge graph or relying solely on query-memory similarity.

3 Method

Our method is inspired by spreading activation in human associative memory. Instead of relying solely on direct query-memory matching, we explore the memory-memory structure to bridge the semantic gap between user queries and gold memories. The proposed framework consists of three components: **structured memory construction**, **dual neighborhood expansion**, and **path-aware reranking**.

3.1 Structured Memory Construction

To construct structured memory notes, we integrate explicit memory content with LLM-generated metadata. For each memory m_i, we represent it as:

$$m_i = \{c_i, t_i, K_i, G_i, X_i, e_i\}, \tag{1}$$

where c_i represents the original content, t_i is the timestamp of the interaction, K_i denotes LLM-generated keywords, G_i contains LLM-generated tags for categorization, X_i represents the LLM-generated contextual description that provides high-level semantic understanding. e_i denotes the

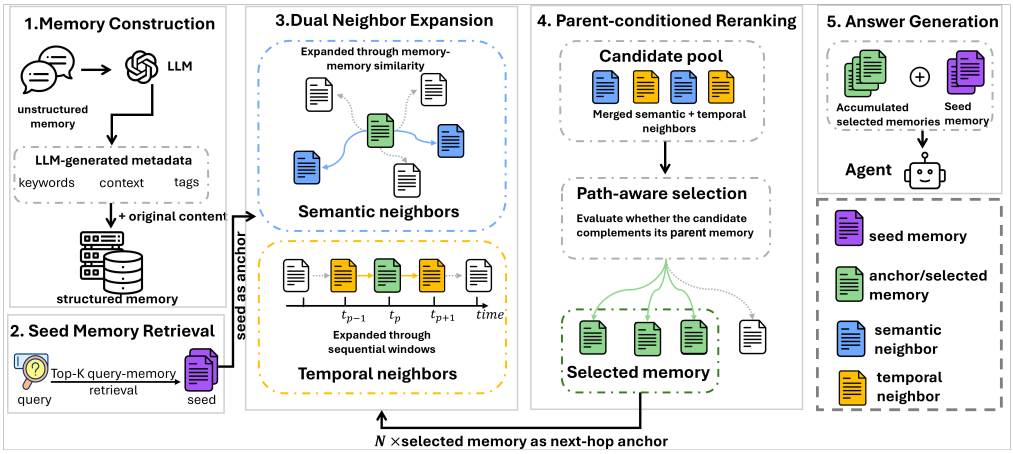


Figure 3: Overall workflow of the proposed method. The framework first constructs structured memory representations and retrieves seed memories using query–memory similarity, preserving them as evidence and using them as initial anchors. It then expands semantic and temporal neighborhoods for each anchor and applies path-aware selection to identify promising memories. The selected memories are accumulated for final answer generation and reused as next-hop anchors for iterative memory exploration.

embedding representation of the memory note and we compute e_i by:

$$e_i = f_{\text{enc}}\left(\text{concat}(c_i, K_i, G_i, X_i)\right) \tag{2}$$

In this way, the memory embedding contains rich semantic information and multi-level aspects of the memory.

3.2 Dual Neighborhood Expansion

Given a query q, we first retrieve the top-K memories from the memory collection based on cosine similarity between query and memory embeddings. These memories form the initial seed set $\mathcal{S}_{\text{seed}}$. In our framework, we maintain two memory sets during exploration: $\mathcal{F}_t$, the current anchor set at hop t, which contains the memories to be expanded in the current round; and $\mathcal{R}_t$, the accumulated evidence set after hop t, which stores the memories preserved for final answer generation.

The initially retrieved memories are referred to as *seed memories*. They play two roles in our framework. First, they are preserved as part of the retrieved evidence. Second, they initialize the first-hop expansion anchor:

$$\mathcal{F}_0 = \mathcal{S}_{\text{seed}}, \qquad \mathcal{R}_0 = \mathcal{S}_{\text{seed}}. \tag{3}$$

Each memory $p \in \mathcal{F}_0$ is treated as a contextual anchor for the first round of memory expansion. Unlike standard RAG, we do not stop after retrieving these top-K seed memories or treat them as the complete final context. Instead, we use them as entry points to further explore the memory-memory structure, allowing the retriever to discover evidence that may be missed by direct query-memory similarity.

For each anchor memory $p \in \mathcal{F}_t$, we perform dual neighborhood expansion to discover potential evidence memories. Specifically, we consider two complementary types of neighbors: semantic neighbors and temporal neighbors.

Semantic neighbors are selected using memory–memory embedding similarity. Specifically, we retrieve the top-L memories with the highest cosine similarity to the anchor:

$$\mathcal{N}_{\text{sem}}(p) = \underset{m_c \in \mathcal{M} \setminus \{p\}}{\arg\,\text{top-L}} \cos(\mathbf{e}_p, \mathbf{e}_c). \tag{4}$$

Here, $\mathbf{e}_p$ and $\mathbf{e}_c$ denote the embeddings of anchor memory p and candidate memory m_c, respectively, and $\mathcal{M}$ is the memory collection. This semantic expansion enables global exploration in the memory

space, allowing the agent to find memories that are strongly connected to the anchor even if they are weakly similar to the original query.

Since conversational memories are often short and context-dependent, an isolated memory turn may be insufficient for interpretation. We therefore complement semantic expansion with a temporal neighborhood defined by a sequential window around each anchor.

$$\mathcal{N}_{\text{temp}}(p) = \{m_j \in \mathcal{M} \mid |t_j - t_p| \leq w\}, \tag{5}$$

where t_p and t_j denote the temporal indices of memories p and m_j, and w is the window size. The temporal neighborhood preserves local episodic continuity and helps recover contextual scaffolding around the anchor memory.

Finally, we combine both neighborhoods into the candidate set $\mathcal{N}(p) = \mathcal{N}_{\text{sem}}(p) \cup \mathcal{N}_{\text{temp}}(p)$ for each anchor memory. By combining semantic and temporal expansion, our method explores both the global associative structure and the local sequential structure of agent memory, discovering semantically linked but temporally distant evidence while preserving the context needed to interpret short, context-dependent turns.

3.3 Path-aware Reranking

After each anchor memory is expanded to its semantic and temporal neighbors, all expanded neighbor memories are merged into a candidate pool. We then select candidate memories, which are added to the retrieved evidence set and reused as anchors for next-hop expansion. Since a cross-encoder captures fine-grained token-level interactions between the query and candidate text, which are not directly modeled by embedding-space similarity, we use the cross-encoder(CE) reranker to score and select the next-hop anchor.

A standard cross-encoder reranker scores each candidate independently as $\mathrm{CE}(q, c)$, where q is the query and c is a candidate memory, treating each candidate in isolation. However, this isolated scoring is insufficient for agent memory retrieval. First, individual memory units are often short and context-dependent, and a single memory may not contain enough information to determine its usefulness. Second, whether a candidate memory is useful depends not only on its direct relevance to the query, but also on how well it complements the parent memory from which it is expanded. In other words, a candidate should be evaluated according to its contribution along the current retrieval path. To address this limitation, we introduce path-aware reranking. Rather than scoring each candidate solely by its direct relevance to the query, we condition on the parent memory from which the candidate is expanded. This allows the model to evaluate whether the candidate contributes complementary evidence along the current retrieval path. For each candidate memory c expanded from a parent memory p, we compute:

$$\mathrm{score}(q, p, c) = \mathrm{CE}\big(q,\ p \oplus c\big). \tag{6}$$

where $\oplus$ denotes text concatenation. This mechanism allows the reranker to evaluate the candidate memory in the context of its parent memory, rather than treating it as an isolated unit. Intuitively, the score estimates whether adding c after p provides useful complementary evidence for answering the query.

At expansion step t, we collect all parent-candidate pairs into a candidate pool $\mathcal{P}_t$. We then select the top-B candidates according to the path-aware score:

$$\mathcal{W}_t = \underset{(p,c)\in\mathcal{P}_t}{\arg\,\text{top-B}}\ \mathrm{score}(q, p, c). \tag{7}$$

The unique selected memories at hop t, $\mathcal{C}_t = \{c \mid (p, c) \in \mathcal{W}_t\}$, are added to the accumulated retrieved evidence set, $\mathcal{R}_{t+1} = \mathcal{R}_t \cup \mathcal{C}_t$, and also reused as anchors for the next-hop expansion, $\mathcal{F}_{t+1} = \mathcal{C}_t$.

After T expansion steps, the final memory context provided to the agent is the union of the initial seed memories and all accumulated selected memories: $\mathcal{R}_{\text{final}} = \mathcal{R}_T = \mathcal{S}_{\text{seed}} \cup \bigcup_{t=0}^{T-1} \mathcal{C}_t$. The agent then generates the final response conditioned on the query and the final retrieved memory context.

Through this retrieval-path-aware frontier selection, our method filters out noisy neighbors while preserving memories that contribute useful evidence along the current retrieval path. At the same time, by using the selected memories as next-hop anchors, retrieval can iteratively propagate over the memory-memory structure and recover evidence that is missed by direct query-memory matching.

Table 1: Performance comparison across different question categories. Within each model group, best results in each column are shown in bold, and second-best results are underlined. Improvement is computed as Ours minus the best non-Ours baseline.

Model	Method	Temporal		Open Domain		Multi-hop		Single-hop		Overall	
		F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1	F1	BLEU-1
GPT-4o-mini	LangMem	31.9	26.2	**29.4**	**23.5**	33.5	23.9	38.8	33.1	35.8	29.4
	RAG	19.5	15.7	19.0	13.5	18.6	11.7	22.2	18.6	20.8	16.4
	Zep	44.8	38.1	22.9	15.7	27.5	19.3	39.7	33.7	37.5	30.9
	A-MEM	43.5	37.9	12.6	12.0	27.3	20.7	44.6	37.7	39.2	33.0
	Mem0	34.5	26.3	22.5	16.5	**36.8**	**26.5**	46.8	37.6	40.9	31.9
	HippoRAG	41.4	29.8	19.4	15.4	21.3	15.0	35.8	25.4	33.3	23.8
	Ours	**45.1**	**38.6**	25.9	19.3	32.0	24.3	**51.9**	**43.3**	**45.2**	**37.3**
	Improvement	+0.3	+0.5	-3.5	-4.2	-4.8	-2.2	+5.1	+5.6	+4.3	+4.3
Llama-3.1-8B Instruct	LangMem	27.7	21.5	18.6	**16.0**	22.4	15.2	31.6	23.8	28.3	21.3
	RAG	18.0	15.3	13.1	9.0	19.8	14.5	25.6	22.4	22.2	18.6
	Zep	3.5	2.6	15.0	11.6	30.2	17.2	26.7	18.4	21.8	14.5
	A-MEM	32.8	25.2	13.8	11.4	26.9	18.9	36.6	27.3	32.6	24.3
	Mem0	26.9	21.1	18.6	13.8	27.3	18.6	34.0	24.8	30.3	22.2
	HippoRAG	34.1	25.6	**21.4**	13.8	23.0	15.5	33.8	23.2	31.1	21.7
	Ours	**35.9**	**27.3**	15.0	11.3	**35.6**	**25.1**	**49.2**	**36.4**	**41.8**	**30.9**
	Improvement	+1.8	+1.7	-6.4	-4.7	+5.4	+6.2	+12.6	+9.1	+9.2	+6.6

4 EXPERIMENT

4.1 EXPERIMENTAL SETTINGS

Datasets and metrics. We evaluate our method on LoCoMo, a long-term conversational memory benchmark. LoCoMo consists of 10 dialogues with an average of 24K tokens, featuring 1,540 questions across four categories: temporal questions that require time-sensitive reasoning, single-hop questions that can be answered from a single session, multi-hop questions that require evidence across multiple sessions, and open-domain questions that combine conversational context with external knowledge. We report F1 and BLEU-1.

GVD Zhong et al. (2024) contains simulated multi-turn interactions between 15 virtual users and an assistant over 10 days, covering at least two topics per day. Following MemoryOS Kang et al. (2025), we report Memory Retrieval Accuracy (Acc.), Response Correctness (Corr.), and Contextual Coherence (Cohe.), scored by GPT-4o-mini. We additionally evaluate our method using F1 on PerLTQA Du et al. (2024), a personalized long-term memory benchmark with contexts averaging approximately 25K tokens and answers often expressed as sentences rather than short spans.

Implementation Details. We use `text-embedding-3-small` to compute memory and query embeddings. For backbone and response generation models, we evaluate with Llama-3.1-8B-Instruct Grattafiori et al. (2024) and GPT-4o-mini OpenAI (2024). For our method, we retrieve $K = 10$ seed memories for each query. At each hop, we expand each anchor memory with $M = 2$ semantic neighbors and a temporal window of $W = 2$ memories on each side within the same dialogue session. The candidate pool size is set to 48, from which the top $B = 4$ memories are selected as the next-hop anchor. The retrieval process is limited to at most $T = 3$ hops.

4.2 MAIN RESULTS

Table 1 shows the main results on LoCoMo with GPT-4o-mini and Llama-3.1-8B-Instruct. PAGR achieves the best overall performance under both backbones, demonstrating the benefit of memory–memory exploration beyond direct query–memory retrieval.

With GPT-4o-mini, memory-augmented methods generally outperform standard RAG in overall performance, highlighting the need to explicitly model long-term conversational memory. PAGR achieves the best overall results, with **45.2** F1 and **37.3** BLEU-1. It improves over the strongest F1 baseline, Mem0, by 4.3 points and over standard RAG by 24.4 points. With Llama-3.1-8B-Instruct, PAGR again ranks first, reaching **41.8** F1 and **30.9** BLEU-1, outperforming A-MEM by 9.2 F1 points

Table 2: Performance comparison on GVD and PerLTQA using Llama-3.1-8B-Instruct

	GVD			PerLTQA (F1)				
Method	ACC	Corr.	Cohe.	Profile	Social Relationship	Dialogues	Event	Overall
A-MEM	0.78	0.795	0.81	28.3	18.2	11.7	10.4	12.4
PAGR (Ours)	**0.88**	**0.92**	**0.92**	**48.3**	**47.2**	**29.5**	**46.0**	**40.1**

Example 1: Gold memory is in the temporal neighborhood

Question (Temporal, Category 2): "When did Maria adopt Shadow?"

Gold answer: "The week before 13 August 2023."

Step 0 — Seed retrieval using question–memory semantic similarity.

- **Seed 1: similarity = 0.598**
  3:14 pm on 13 August 2023, Maria: "Her name is **Shadow**! She's . . . "
- **Seed 2: similarity = 0.480**
  8:55 pm on 25 February 2023, John: "Wow, **Maria**! How did she . . . "

The gold evidence is absent from the initial top-10 seeds.

Step 1 — Temporal-neighbor expansion.

The gold memory is recovered as a temporal neighbor of Seed 1, two turns earlier in the same session:

- **Gold:**
  3:14 pm on 13 August 2023, Maria: ". . . I just adopted this cute pup from a shelter **last week** . . . "

Example 2: Gold memory is in semantic expansion

Question (Temporal, Category 2): "When did Jon host a dance competition?"

Gold answer: "May 2023."

Step 0 — Seed retrieval using question–memory semantic similarity.

- **Seed 1: similarity = 0.650**
  4:04 pm on 20 January 2023, Jon: "Woah, that pic's from when my dance crew took home first in a local **competition** last year."
- **Seed 7: similarity = 0.545**
  1:26 pm on 3 April 2023, Jon: "My **dance** studio and some other schools are . . . "

The gold evidence is absent from the initial top-10 seeds.

Step 1 — Semantic-neighbor expansion.

The gold memory is recovered as a semantic neighbor of Seed 7 through high anchor–memory similarity.

- **Gold:**
  1:26 pm on 3 April 2023, Jon: "I'm also hosting a dance competition **next month** to showcase . . . "

Figure 4: Qualitative examples illustrating complementary evidence recovery through temporal and semantic expansion.

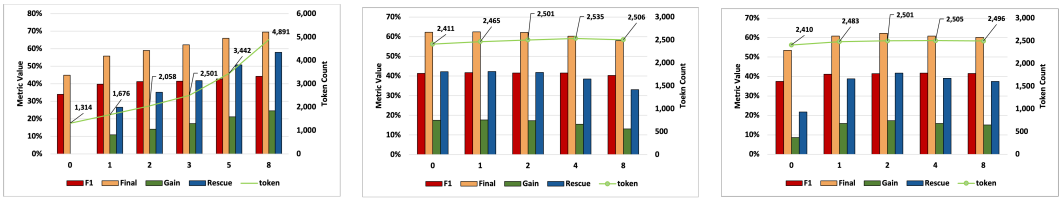
and 6.6 BLEU-1 points. These results show that our retrieval strategy is effective across different backbone models.

The gains are particularly clear on temporal and single-hop questions. For temporal questions which evaluate the understanding of time-related information, our method achieves the best performance on both backbone models, suggesting that temporal neighbor expansion helps recover evidence from temporally related memories. For single-hop questions which require information from a single session, our method achieves 51.9 F1 on GPT-4o-mini and 49.2 F1 on Llama-3.1-8B, outperforming A-MEM by 7.3 and 12.6 points, respectively. This indicates that even when only one key evidence memory is needed, direct query-memory matching may still fail to retrieve the gold memory. Although A-MEM incorporates linked memories into the context, these additional memories may introduce redundant or distracting information. In contrast, our dual-neighbor exploration recovers necessary evidence through memory–memory associations, and path-aware reranking further filters out redundant or noisy memories.

Across both GVD and PerLTQA as shown in Table 2, PAGR consistently outperforms A-MEM. On PerLTQA, PAGR achieves an overall F1 score of 40.1, compared with 12.4 for A-MEM. On GVD, PAGR consistently outperforms A-MEM across all three metrics. These results provide additional evidence for the robustness and generalizability of our method beyond LoCoMo.

4.3 Analysis of Open-domain Performance

To understand the weaker Open-domain performance, we examine evidence availability and answer quality, using Single-hop questions for comparison in Table 3. Expansion recovers gold evidence for an additional 179 Single-hop and 15 Open-domain questions missed by the initial seeds. However, 50 of the 96 Open-domain questions still lack gold evidence in the final context.



(a) hop depth (b) semantic neighbor (c) temporal neighbor

Figure 5: Impact of path-aware memory exploration parameters. Increasing hop depth improves retrieval and answer quality, but substantially increases the number of tokens in the final context. Adding semantic or temporal neighbors brings clear gains over no expansion, while further increasing their sizes provides limited improvement and may introduce noisy memory.

Table 3: Evidence availability and conditional F1 on Single-hop and Open-domain questions.

	Single-hop ($N = 841$)		**Open-domain** ($N = 96$)	
Evidence condition	# Questions	F1	# Questions	F1
Gold in seed	479	61.4	31	16.2
Gold in final context	658	59.4	46	16.3
Gold not in final context	183	12.2	50	13.2

Besides, retrieving at least one gold memory only slightly improves F1 from 13.2 to 16.3 for Open-domain questions. In contrast, for single-hop, F1 increases substantially from 12.2 to 59.4. This suggests that retrieving at least one gold memory is insufficient for accurate Open-domain answers. This is because approximately half of Open-domain questions require multiple gold memories, whereas over 94% of Single-hop questions require only one. Open-domain questions also require external knowledge, and difficulties in evidence integration and reasoning may further limit answer quality.

4.4 Case Study

Figure 4 illustrates the complementary roles of temporal and semantic expansion. In both cases, query–memory embedding similarity ranks seed memories above the gold evidence, leaving it absent from the initial top-10 retrieval, thus highlighting the need to go beyond query–memory embedding matching.

This first example shows that temporal expansion can recover locally adjacent evidence that is difficult to retrieve through direct embedding similarity alone. In the second example, the gold memory actually falls outside the temporal neighborhood of all seed memories but is rescued through semantic expansion. Together, these retrieval traces show that temporal expansion recovers locally distributed episodic evidence, whereas semantic expansion retrieves related evidence outside the local sequential neighborhood.

4.5 Ablation Study

We conduct ablations on LoCoMo by separately removing semantic neighbor expansion (Sem), temporal neighbor expansion (Temp), and cross-encoder reranking (CE). When CE is removed, candidates are selected using embedding similarity. To further isolate the contribution of parent context, we also replace the path-aware scorer $\mathrm{CE}(q, p \oplus c)$ with a standard query–candidate scorer $\mathrm{CE}(q, c)$. In Table 4, we report results on LoCoMo across four complementary metrics:

F1: Overall answer F1. **Initial Cover**: Percentage of questions whose initial top-K seeds $\mathcal{F}_0$ contain *all* gold evidence memories. **Final Cover**: Percentage of questions whose final retrieved set $\mathcal{R}_T$ contains *all* gold evidence memories. **Gain**: Final Cover minus Initial Cover. **Rescue Rate**: Among questions with *no* gold evidence in the initial seeds, the percentage whose final retrieved set contains *at least one* gold memory.

As shown in Table 4, the full model achieves the best performance across all evaluation metrics. Removing either temporal sequential expansion or semantic neighbor expansion degrades the perfor-

Table 4: Ablation study of retrieval components.

Method	F1	Initial Cover	Final Cover	Gain	Rescue Rate
w/o TEMP.	40.2	60.3	65.4	+5.1	17.6
w/o SEM.	42.6	60.4	70.3	+9.9	36.5
w/o CE	40.9	60.4	65.3	+4.9	18.2
CE(q,c)	41.2	60.3	68.0	+7.6	28.6
Ours	**45.2**	**60.4**	**75.8**	**+15.4**	**52.6**

Table 5: Efficiency analysis.

Method	Tokens	Avg. F1	Avg. BLEU-1
LangMem	125	35.8	29.4
RAG	3525	20.8	16.4
Zep	2247	37.5	30.9
A-MEM	6962	39.2	33.0
Mem0	3538	40.9	31.9
Ours	**2716**	**45.2**	**37.3**

mance. Without sequential expansion, Final Cover drops from 75.8% to 65.4%, and Gain decreases from 15.4% to 5.1%. Without semantic expansion, Final Cover drops to 70.3%, and Gain decreases to 9.9%. The results highlight the structure of agent memory: neighboring turns within a session frequently provide the contextual scaffolding required to interpret short, context-dependent memories. Semantic neighbors instead contribute long-range globally related memories across sessions.

Path-aware selection is also essential. When the parent-conditioned cross-encoder is replaced with a standard cross-encoder, both Rescue Rate and Gain are nearly halved. This indicates that modeling only the relationship between query and candidate memory is insufficient. To select useful next-hop evidence, the model must also evaluate how the candidate complements its parent memory along the current retrieval path.

Overall, the ablation results validate the effectiveness of semantic-temporal expansion and path-aware memory selection in recovering gold memories that are missed by the initial query-memory retrieval.

4.6 Efficiency Analysis

Table 5 reports the efficiency comparison. Our method achieves the best overall F1 and BLEU-1 scores while using only about one third of the tokens required by A-MEM. This shows that our approach not only improves answer quality, but also maintains strong token efficiency by selecting a compact and informative memory context.

4.7 Hyperparameter Analysis

We examine hop depth T, semantic neighbor count M, and temporal window size W. As shown in Figure 5, increasing hop depth improves both answer accuracy and retrieval coverage but increases the number of context tokens.

Temporal expansion plays a key role in PAGR. Increasing W from 0 to 1 raises F1 from 37.5 to 41.2 and the rescue rate from 21.8% to 38.8%, showing that nearby conversational turns often contain useful supporting evidence. Gains beyond $W = 2$ are marginal, suggesting that useful temporal evidence is largely local and larger windows may introduce redundant candidates.

Semantic expansion provides complementary benefits. Its gains are small with $W = 2$, where temporal expansion already recovers much of the missing evidence. When temporal expansion is disabled ($W = 0$), however, increasing semantic neighbors from 0 to 2 improves final evidence coverage from 44.9% to 53.5% and the rescue rate from 0.0% to 21.8%, confirming the independent value of semantic memory–memory exploration.

Overall, moderate exploration provides the best trade-off between retrieval quality and context cost. We therefore set $T = 3$, $M = 2$, and $W = 2$ as the default configuration.

5 Conclusion

In this paper, we propose PAGR, a path-aware anchor-guided evidence retrieval framework for long-term agent memory. Motivated by the semantic gap between queries and memories, PAGR uses initially retrieved seed memories as anchors and expands from them through both semantic associations and temporal neighborhoods. A path-aware cross-encoder further selects memories that complement the current retrieval path, reducing redundancy and noise during expansion. Experiments on LoCoMo demonstrate that PAGR improves long-term memory retrieval and answer generation while maintaining token efficiency.

AI USE STATEMENT

The authors used ChatGPT 5.6 for language editing, including formatting, grammar correction, and limited code assistance. The entire structure of the paper, including the initial and subsequent versions, was created by the authors. All ideas, analyses, and conclusions are solely those of the authors. The authors take full responsibility for the content of this publication.

REFERENCES

John R Anderson. A spreading activation theory of memory. *Journal of verbal learning and verbal behavior*, 22(3):261–295, 1983.

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

Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407, 1975.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. Perltqa: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pp. 152–164, 2024.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Bernal J Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. *Advances in neural information processing systems*, 37:59532–59569, 2024.

Zhanghao Hu, Qinglin Zhu, Di Liang, Hanqi Yan, Yulan He, and Lin Gui. Beyond rag for agent memory: Retrieval by decoupling and aggregation. *arXiv preprint arXiv:2602.02007*, 2026.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. Memory os of ai agent. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 25972–25981, 2025.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

LangChain. LangMem SDK for Agent Long-Term Memory. `https://www.langchain.com/blog/langmem-sdk-launch`, 2025. Accessed: 2026-03-25.

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

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

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

Jaap MJ Murre and Joeri Dros. Replication and analysis of ebbinghaus' forgetting curve. *PloS one*, 10(7):e0120644, 2015.

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

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`, 2024. Accessed: 2026-03-25.

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

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

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

Aditi Singh, Abul Ehtesham, Saket Kumar, Tala Talaei Khoei, and Athanasios V Vasilakos. Agentic retrieval-augmented generation: A survey on agentic rag. *arXiv preprint arXiv:2501.09136*, 2025.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 10014–10037, 2023.

Endel Tulving and Donald M Thomson. Encoding specificity and retrieval processes in episodic memory. *Psychological review*, 80(5):352, 1973.

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

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *Advances in Neural Information Processing Systems*, 38:17577–17604, 2026.

Sikuan Yan, Xiufeng Yang, Zuchao Huang, Ercong Nie, Zifeng Ding, Zonggen Li, Xiaowen Ma, Jinhe Bi, Kristian Kersting, Jeff Z Pan, et al. Memory-r1: Enhancing large language model agents to manage and utilize memories via reinforcement learning. *arXiv preprint arXiv:2508.19828*, 2025.

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

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