Bridging the Semantic Gap in Long-Term Agent Memory Retrieval

Sixu Chen
Department of Computer Science
Kent State University
schen53@kent.edu

Qiang Guan
Department of Computer Science
Kent State University
qguan@kent.edu

Introduction

LLM-based agents perform well across many tasks[1] but still struggle with long-term interactions requiring coherence, continuity, and personalization[2]. LLM-based agents often use RAG-style memory systems to retrieve past experiences, but direct query–memory similarity can fail when user queries and fine-grained memories are semantically misaligned.

We address this semantic gap with an iterative, anchor-guided retrieval strategy inspired by spreading activation[3] in associative memory. Starting from a small set of seed memories, the method expands through global semantic associations and local temporal neighbors, while a path-aware cross-encoder selects complementary evidence across multiple hops. Retrieved memories are accumulated and reused as anchors for further exploration.

Background

Most agent memory methods still rely on a retrieval-augmented generation (RAG) paradigm. 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. We observed that a semantic gap exists between the query and the memory representation, making direct query-memory embedding similarity insufficient for reliable recall.

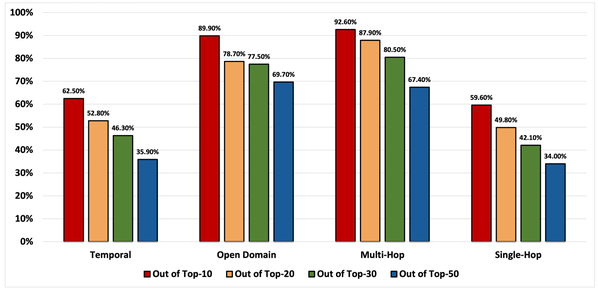


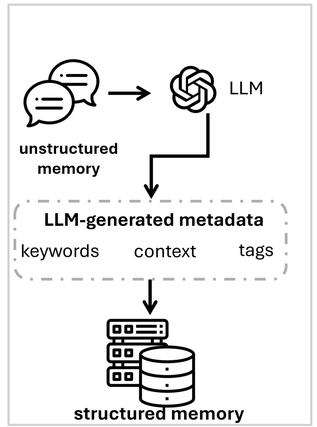
Figure 1: failure mode on LoCoMo dataset

In Figure.1 Across all four LoCoMo categories, many gold memories are absent from the top-10 results, and even top-50 retrieval fails to recover them reliably. This indicates that similarity-based retrieval may miss relevant evidence or retrieve it alongside substantial irrelevant and redundant content.

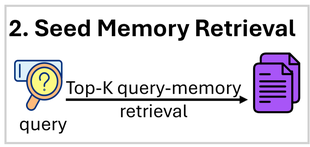
.

Methodology

Memory Construction

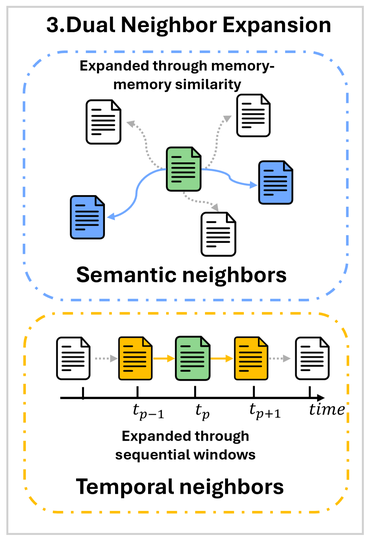


Step 1
- ✓ Treat each interaction as an atomic memory.
- ✓ Use an LLM to generate keywords, tags, and contextual descriptions for each memory.
- ✓ Construct a structured memory note with the original content and LLM-generated metadata.



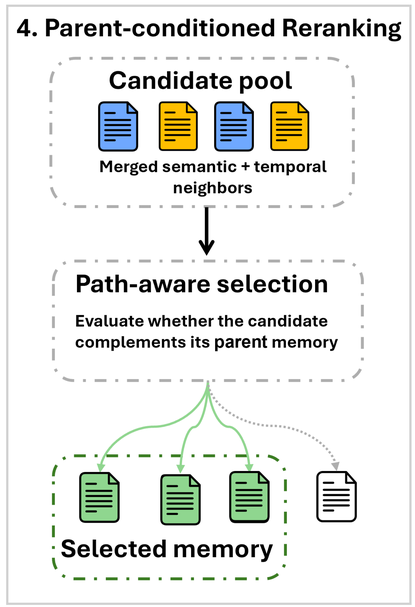
Step 2
- ✓ Using query–memory embedding similarity to select top-K memories as seeds.
- ✓ Use seed memories as anchors rather than treating them as the final retrieval results.

Memory Expansion



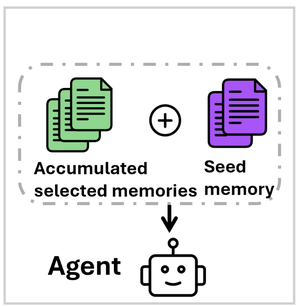
Step 3
- ✓ Semantic expansion: retrieve memories most similar to each anchor for global associative exploration.
- ✓ Temporal expansion: retrieve nearby memories within a local time window to preserve conversational context.
- ✓ Explore both global semantic links and local sequential context to recover evidence missed by direct query–memory similarity.

Memory Selection



Step 4
- ✓ Merge expanded neighbors from all anchors into a shared candidate pool.
- ✓ Score each candidate with its parent context instead of evaluating the candidate alone.
- ✓ Select top-B candidates that provide complementary evidence along the current retrieval path.
- ✓ Add selected memories to the evidence set for final answer generation.
- ✓ Reuse selected memories as next-hop anchors for continued multi-hop exploration.

Final Answer



Step 5
- ✓ Using seeds and accumulated selected memories to answer the question

Experiment results

Table 1: Results on GVD using Llama-3.1-8B-Instruct

Method	ACC	Corr.	Cohe.
A-MEM	0.78	0.795	0.81
Ours	0.88	0.92	0.92

Table 2: F1 score comparison on PerLTQA using Llama-3.1-8B-Instruct

Method	Profile	Social Relationship	Dialogues	Event	Overall
A-mem	28.3	18.2	11.7	10.4	12.4
Ours	48.3	47.2	29.5	46.0	40.1

As shown in the above tables, our method consistently outperform baseline method AMEM[4] on two different datasets, which validates the effectiveness of out method.

Case Study

Q: When did Maria adopt Shadow?

Step 1 — Seed retrieval:
- Seed1: "Her name is Shadow..."

→ This memory matches the query because it mentions "Shadow," but it does not state when the adoption occurred.
- The gold evidence is absent from the initial top-10 seeds.

Step 2 — Temporal-neighbor expansion:
- Temporal neighbor: "I just adopted this cute pup ... last week."

→ ✓ Gold evidence recovered through temporal neighbors of Seed1, two turns earlier in the same session

This example shows that temporal expansion can recover locally adjacent evidence that is difficult to retrieve through direct semantic similarity alone.

References

[1]Wang, Guanzhi, et al. "Voyager: An open-ended embodied agent with large language models." arXiv preprint arXiv:2305.16291 (2023).
[2] Packer, Charles, et al. "Memgpt: Towards llms as operating systems." arXiv preprint arXiv:2310.08560 (2023).
[3] Anderson, John R. "A spreading activation theory of memory." Journal of verbal learning and verbal behavior 22.3 (1983): 261-295.
[4] Xu, Wujiang, et al. "A-mem: Agentic memory for llm agents." Advances in Neural Information Processing Systems 38 (2026): 17577-17604.

Acknowledgement: The entire structure of the paper, including the initial and subsequent versions, was created by the authors. LLM tools were used for language editing, including formatting, grammar correction. All ideas, analyses, and conclusions are solely those of the authors.