HERO: Human-profile Enhanced Retrieval Optimization Framework for Long-term Agent Memory

Yuanhua Lin and Yile Li and Zhiyuan Zhao and Jing Shang and Jian Sun
China Mobile Information Technology Co., Ltd., Beijing, China
{linyuanhua, liyile, zhaozhiyuan, shangjing, sunjian01}@chinamobile.com

Abstract

Long-term memory is crucial for personalized responses and long-horizon agent interactions. Existing methods often rely on LLMs to compress or rewrite dialogue histories and use the transformed memories as retrieval evidence. Despite the progress in organizing fragmented contexts, two major drawbacks persist: (1) information loss from compression, which discards fine-grained but later useful details, and (2) semantic drift from rewriting, which erodes the original tone and situated context. In this work, we propose a novel Human-profile Enhanced Retrieval Optimization framework for long-term agent memory (HERO). Specifically, HERO converts the dialogue history into a traceable heterogeneous memory graph that preserves raw dialogue text as evidence for reasoning, thereby mitigating information loss. For retrieval, HERO extracts initial anchors from the current query and incorporates human profiles via an iterative graph traversal; these anchors and profiles provide guidance signals that adaptively activate the most informative regions of the graph. Experiments on two benchmark datasets show that HERO outperforms strong baselines on both factual and personalized reasoning, while providing more faithful access to raw dialogue evidence.

1 Introduction

The improvement of Large Language Models (LLMs) has prompted the evolution of agents (Schlegel et al., 2025; Yang et al., 2024; Pantiukhin et al., 2025). Agent memory is a core capability for such systems, enabling agents to reuse experiences and sustain personalized long-term interactions (Zhang et al., 2025c; Fang et al., 2025a). A common implementation is Retrieval-Augmented Generation (RAG) (Lewis et al., 2020; Borgeaud et al., 2022), which retrieves historical context or external knowledge as model input. While RAG helps mitigate hallucination (Huang et al., 2025), it remains a memory mechanism that is decoupled from the reasoning process. The retrieved isolated text fragments lack inherent structured connections, leading to substantial redundancy and partially relevant content. Consequently, the standard RAG paradigm is akin to static knowledge access from an external knowledge base rather than a human-like memory system, and thus lacks the capacity for multi-turn, dynamic interaction and reasoning during the cognitive process.

To overcome the aforementioned bottlenecks, episodic long-term memory (Svoboda et al., 2006; Kasap and Magnenat-Thalmann, 2010; Pink et al., 2025) has emerged as a promising research direction. It focuses on the dynamic integration of memory and reasoning, aiming to achieve flexible activation, adaptive reorganization, and continuous evolution of memory. From the perspective of memory organization, existing methods can be broadly divided into summary-based and graph-based methods. Summary-based methods refine refine dialogue histories to generate summaries or human profiles (Team, 2023; Yu et al., 2024; Lee et al., 2024); graph-based methods extract memory elements (such as entities and keywords) or leverage LLMs to summarize text blocks, and represent temporal, causal, or semantic associations via relational edges (Sarthi et al., 2024; Gutierrez et al., 2024; Edge et al., 2024; Rezazadeh et al., 2025). These methods address the fragmented retrieval and redundancy inherent in RAG. Nevertheless, both approaches diverge from human cognitive science (Dickerson and Eichenbaum, 2009; Tanguay et al., 2023): the efficacy of agent memory is fundamentally constrained by the narrative and essential memory extraction capabilities of LLMs. For instance, existing graph-based methods are prone to introducing noise and ambiguity when constructing the topology due to inaccurate relation extraction (Han et al., 2025; Guo et al., 2024), hierarchical clustering may be compromised by erroneous

LLM-generated summaries, causing hallucinations to propagate from lower-level representations to higher-level abstractions(Li et al., 2026).

This reliance on abstraction causes two closely related problems. First, compression may omit details that appear unimportant at write time but later become crucial, such as temporal references, causal conditions, numerical facts, or subtle preference changes. Second, rewriting dialogue into summaries, profiles, or standardized memory units may introduce semantic drift: even when the main fact is retained, the user's original wording, affect, speaker attribution, or situated context can be weakened.

Based on the above analysis, a natural idea is to construct a long-term memory that inspired by human cognition for agents, which can swiftly locate relevant concepts (including basic attributes, preferences, etc.), activate related episodic memory slices, perform factually faithful reasoning and generate personalized responses. In this work, we propose **HERO**, a Human-profile Enhanced Retrieval Optimization framework for long-term agent memory. HERO follows a human-cognition-inspired principle: raw dialogue traces are preserved as episodic evidence, while higher-level profile information is used only to guide memory activation. Specifically, HERO constructs a heterogeneous memory graph with Episodic Traces, Episodic Units, and Episodic Cues, connected through structural inclusion, semantic indexing, and temporal chaining. This graph preserves local dialogue contexts while supporting traversal across fine-grained textual evidence and adjacent chronological contexts. In parallel, HERO maintains profile insights derived from episodic traces, but these profile nodes act as navigational bridges rather than replacement memories.

For retrieval, we introduce an iterative cue-activation mechanism to accommodate distinct memory tendencies for factual and personalized reasoning. Specifically, we operate across two dimensions—query relevance and implicit reasoning based on human-profiles, generating cues across three levels: queries, episodic units, and profiles. This extends the naive memory retrieval process, which relies solely on query-embedding similarity, into a two-stage pipeline. First, we leverage human profiles to guide cue reconstruction and expansion of the original query; then, we perform path-based retrieval on the memory graph using the generated cues. In this way, our model can retrieve explicit episodic memories and trace implicit, cross-episodic long-range dependencies through the graph structure.

The contributions of this paper are as follows:

- We analyze the limitations of existing long-term memory approaches and propose a novel human-profile-enhanced retrieval optimization framework, which incorporates human-cognition-inspired design principles.
- We design a heterogeneous memory graph that preserves complete dialogue episodes as final evidence and uses human-profile information as retrieval guidance rather than memory replacement.
- We design a profile-aware cue activation and graph retrieval mechanism that retrieves explicit facts and implicit cross-episodic dependencies while constraining topic drift through query-conditioned filtering.
- We conduct experiments on two benchmark datasets. HERO achieves 56.06% F1 on LoCoMo and 70.63% accuracy on PERSONAMEM, demonstrating that HERO outperforms strong baselines on complex factual question answering and personalized reasoning tasks.

2 Related Work

To address the challenge of processing long-horizon interaction histories in agents, recent research on long-term memory has broadly diverged into two strands. One line is the summary-based approaches (Team, 2023; Kang et al., 2025a; Lu et al., 2023; Yu et al., 2024; Lee et al., 2024; Packer et al., 2023; Fang et al., 2025b; Li et al., 2025b; Pan et al., 2025; Li et al., 2025a). The main idea is to refine original texts into hierarchical summaries, or structured facts to bypass the context window limitation. Representative systems like MemoryBank (Zhong et al., 2024), LightMem (Fang et al., 2025b), EverMemOS (Hu et al., 2026), Mem0 (Chhikara et al., 2025), and CAM (Li et al., 2026) implement various cognitive-inspired write-time compression or memory lifecycle management schemes. However, these methods inevitably suffer from information loss and semantic drift due to their heavy reliance on LLMs to rewrite or summarize raw interactions.

Another line is the graph-based memory approaches (Sarthi et al., 2024; Xu et al., 2026; Guo

et al., 2024; Han et al., 2025; Anokhin et al., 2025; Zhang et al., 2025a; Gutierrez et al., 2024; Edge et al., 2024; Zhuang et al., 2025; Rezazadeh et al., 2025; Rasmussen et al., 2025), which extracts structured knowledge from text to construct graphs and performs path-based retrieval. This structure naturally supports multi-hop reasoning and alleviates knowledge conflicts through temporal modeling.

These methods significantly advance long-term agent memory. However, they primarily rely on compressing historical information and the re-narration power of LLMs, which inevitably leads to irreversible information loss or semantic distortion. In this paper, we construct a novel heterogeneous memory graph, which is faithful to the original text. Moreover, we leverage the human-profile and the current query through a multi-round iterative to generate initial cues, which in turn guide agents' personalized response.

3 Methodology

We define the user's dialogue sessions as $\mathcal{H} = \{s_1, s_2, \ldots, s_T\}$, where each session s_i consists of a sequence of events or multi-turn dialogues. We assume the existence of a latent human profile $\mathcal{P}$, representing the user's accumulated facts, evolving traits, and preferences derived from $\mathcal{H}$.

Given a current user query q, the goal of HERO is to retrieve an optimal memory subset $\mathcal{M}^* \subset \mathcal{H}$ that maximizes the generation probability of the target response r. This optimization problem can be formulated as:

$$\mathcal{M}^* = \arg\max_{\mathcal{M} \subset \mathcal{H}} P(r \mid q, \mathcal{M}, \mathcal{P}; \Phi_{\text{LLM}}), \quad (1)$$

where Φ_{LLM} denotes the parameters of the backbone Large Language Model.

We propose HERO, a long-term agent memory framework that enables personalized reasoning without sacrificing the fidelity of raw dialogue records. As shown in Figure 1, HERO has two components: (1) Memory Construction, which organizes multi-granular conversational experiences into a coarse-to-fine heterogeneous graph; and (2) Memory Retrieval, which uses multi-level cues to retrieve profile-consistent evidence for explicit fact retrieval and implicit cross-session reasoning.

3.1 Memory Construction

To support factually faithful reasoning over episodic memory, HERO constructs an episodic-centered heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The graph has two layers: an *Episodic Layer* that stores raw historical events and a *Profile Layer* that distills human profiles.

The Episodic Layer focuses on faithfully recording what happened in the original dialogues. It hierarchically segments the conversation flow into granular units to ensure memory traceability, and contains three types of nodes: Episodic Traces ($\mathcal{V}_t$), Episodic Units ($\mathcal{V}_u$), and Episodic Cues ($\mathcal{V}_{c\_epi} \subset \mathcal{V}_c$):

(1) Episodic Traces are partitioned into semantically coherent sessions or time windows. Each trace node $v_t \in \mathcal{V}_t$ represents a local conversation scenario memory, encompassing the full original text, timestamp, and sequential index.

(2) Episodic Units are the further segmented atomic sentence-level units v_u. Each unit node $v_u \in \mathcal{V}_u$ acts as the fundamental carriers of evidence, retaining detailed contextual constraints.

(3) Episodic Cues are entities and keywords that extracted from v_u by LLMs, and serve as associative bridges. Besides, entities extracted from different units are normalized and merged into shared nodes in the global cue set $\mathcal{V}_c$.

The Profile layer is designed for understanding users. This layer performs high-level semantic abstraction to model users' implicit attributes, and contains two types of node: Profile Insights ($\mathcal{V}_p$) and Profile Cues ($\mathcal{V}_{c\_pro} \subset \mathcal{V}_c$):

(1) Profile Insights ($\mathcal{V}_p$): For each episodic trace, the LLM distills human-centric information: specifically *Facts*, *Insights*, and *Traits* to generate profile nodes v_p, where $v_p \in \mathcal{V}_p$.

(2) Profile Cues ($\mathcal{V}_{c\_pro} \subset \mathcal{V}_c$): Fine-grained entities are further extracted from v_p to link abstract user portraits back to the global semantic space. Note that $\mathcal{V}_c = \mathcal{V}_{c\_epi} \cup \mathcal{V}_{c\_pro}$.

The final node set $\mathcal{V}$ is composed of multi-granular semantic abstractions:

$$\mathcal{V} = \mathcal{V}_t \cup \mathcal{V}_u \cup \mathcal{V}_p \cup \mathcal{V}_c. \quad (2)$$

The edge set $\mathcal{E}$ for topological construction contains three types:

(1) Structural Inclusion: Directed edges $\{(v_u, v_t), (v_p, v_t)\}$ link fine-grained units and profile nodes to their source episodic traces. Directed edges $\{(v_c, v_t), (v_c, v_p)\}$ connect cue nodes to their originating episodic traces and profile insights.

(2) Semantic Indexing: Weighted edges $\{(v_c, v_t), (v_c, v_p)\}$ connect retrieval cues with episodic traces and profile insights; weights $w_{c,t}$ and $w_{c,p}$ are based on occurrence frequency.

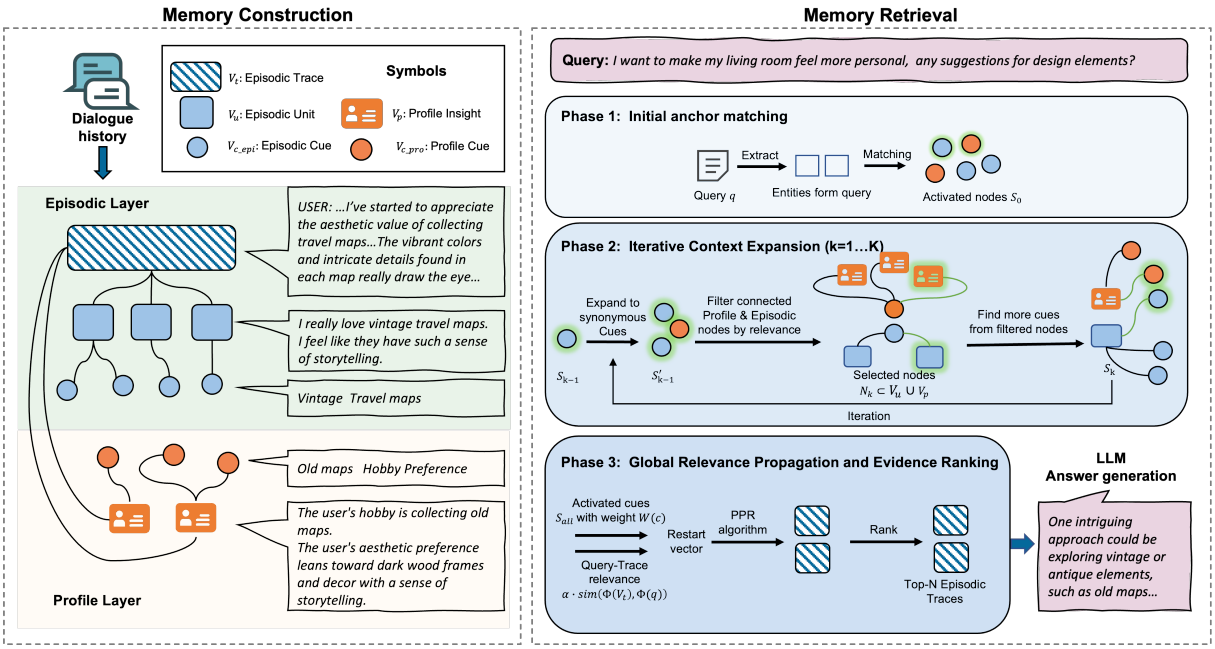

Figure 1: Overview of the proposed HERO framework. Left: Memory Construction. User dialogues are processed to extract Episodic Traces, which are further organized into Episodic Units and Episodic Cues to form the Episodic Layer. In parallel, Profile Insights and corresponding Profile Cues are derived from Episodic Traces to construct the Profile Layer. Episodic Cues and Profile Cues jointly constitute the cue nodes in a heterogeneous memory graph. Right: Memory Retrieval. Given a user query, entities are first extracted and matched against the cue nodes in the graph to obtain an initial candidate set. Through iterative expansion via Cue–Episodic Unit and Cue–Profile Insight bridges, additional relevant cue nodes are activated and incorporated into the Personalized PageRank initialization. In this process, Episodic Units support explicit semantic linking, while Profile Insights provide implicit, profile-consistent semantic connections.

(3) Temporal Chaining: Adjacent episodic traces are connected by $\{(v_{t_i}, v_{t_{i+1}})\}$ to preserve temporal order.

This construction turns streaming conversations into a traversable cognitive association graph with three benefits. First, the Episodic Layer supports faithful retrieval of raw evidence and multi-hop associations across sessions, enabling variable-resolution retrieval from coarse traces to fine-grained facts without irreversible compression. Second, the Profile Layer captures user preferences and provides implicit links for retrieving profile-consistent evidence even without direct keyword overlap. Third, HERO supports incremental updates: new dialogues create local nodes and edges (linking to the latest predecessor and merging new cues), without full graph reconstruction.

3.2 Memory Retrieval

During retrieval, HERO employs a profile-enhanced retrieval optimization strategy to handle two query types. Factual questions activate query-related cues to locate complete raw evidence and avoid losses from extraction or summarization. Personalized reasoning questions activate profile-related cues, because their target attributes may not appear as explicit entities in the dialogue. For example, answering "Describe the user's dietary preferences" requires profile insights to bridge abstract preferences with multiple raw fragments. HERO retrieves evidence in three phases: (1) Initial Anchor Identification, (2) Iterative Context Expansion, and (3) Global Relevance Propagation and Evidence Ranking.

Phase 1: Initial Anchor Identification We extract the explicit entity set $\mathcal{E}_q$ from the query q and map them to the cue nodes $\mathcal{V}_C$ to generate the initial active set $\mathcal{S}_0$, which is defined as:

$$\mathcal{S}_0 = \{v \in \mathcal{V}_C \mid \exists e \in \mathcal{E}_q, \text{sim}(\Phi(e), \Phi(v)) \geq \tau_{\text{init}}\},$$

where $\Phi(\cdot)$ denotes the embedding function, $\text{sim}(\cdot,\cdot)$ is the cosine similarity, and τ_{init} is the matching threshold.

Phase 2: Iterative Context Expansion In this phase, HERO emulates the human "following-the-clues" recall process: for K iterations, the system traverses a cue–context–cue pathway. Firstly, we

expand $\mathcal{S}_{k-1}$ to include synonymous nodes, handle the lexical diversity situation:

$$\mathcal{S}'_{k-1} = \{v \in \mathcal{V}_C \mid \exists u \in \mathcal{S}_{k-1}, \\ \text{sim}(\Phi(v), \Phi(u)) \geq \tau_{\text{alias}}\},$$

where τ_{alias} is proposed to constrain semantic drift. We then collect context nodes connected to the current cues $\mathcal{S}'_{k-1}$:

$$\mathcal{N}_k = \bigcup_{v \in \mathcal{S}'_{k-1}} \text{Adj}(v) \cap (\mathcal{V}_u \cup \mathcal{V}_p),$$

Next, we filter them by query relevance:

$$\mathcal{N}^*_k = \{n \in \mathcal{N}_k \mid \text{sim}(\Phi(n), \Phi(q)) \geq \tau_{\text{iter}}\},$$

we activate new cue nodes in $\mathcal{V}_C$ connected to the selected contexts $\mathcal{N}^*_k$ to generate cues:

$$\mathcal{S}_k = \{c \in \mathcal{V}_C \mid \exists n \in \mathcal{N}^*_k, (n, c) \in \mathcal{E}\} \setminus \bigcup_{i=0}^{k-1} \mathcal{S}_i,$$

for each new cue $c \in \mathcal{S}_k$, set its activation weight by transferring from a previous cue $c' \in \mathcal{S}'_{k-1}$ through the linking context $n \in \mathcal{N}^*_k$:

$$W(c)_k = W(c')_{k-1} \times \text{sim}(\Phi(n), \Phi(q)).$$

Phase 3: Global Relevance Propagation and Evidence Ranking After K iterations, the system aggregates the activation signals to identify the most salient episodic traces (v_t). The activation intensity of an episodic trace is calculated by combining dense retrieval relevance and the cue-matching bonus:

$$W(v_t) = \alpha \, \text{sim}(\Phi(v_t), \Phi(q)) \\ + \log\left(1 + \sum_{c \in \mathcal{S}_{\text{all}}} \frac{W(c) \ln(1 + \text{f}(c, v_t))}{L_c}\right),$$

where $\mathcal{S}_{\text{all}} = \bigcup_{k=0}^{K} \mathcal{S}_k$ denotes the global activated cue set. $W(c)$ is the activation weight of cue node c, $\text{f}(c, v_t)$ is its occurrence frequency in trace v_t, and L_c is the hierarchical depth of cue node c.

Then, node-level scores are integrated into a restart probability vector $\mathbf{r}$ constructed from $\{W(v) \mid v \in \mathcal{V}_C \cup \mathcal{V}_t\}$; v_p is masked during propagation and thus excluded from $\mathbf{r}$. Following (Gutierrez et al., 2024; Zhuang et al., 2025), we employ Personalized PageRank (PPR) (Haveliwala, 2002) on G with $\mathbf{r}$ as the restart distribution and the Profile Insights nodes v_p masked, which can be viewed as energy propagation on the memory graph:

$$\mathbf{a} = (1 - d)\, \mathbf{r} + d\, \mathbf{a}\, \mathbf{P}_{\text{mask}},$$

where $\mathbf{a}$ denotes the stationary PPR score vector, $d \in (0, 1)$ is the damping factor, and $\mathbf{P}_{\text{mask}}$ is the row-normalized transition matrix after masking profile nodes v_p (and their incident edges), ensuring that the propagation only flows through evidence-bearing nodes. Finally, we rank Episodic Traces v_t by their PPR scores, and apply a re-ranking model to return the top-k Episodic Traces as reasoning evidence for the LLM.

4 Experiments

In this section, we evaluate our proposed HERO framework with respect to the following research questions:

- **RQ1.** Compared with methods based on compression or rewriting, can HERO provide a more accurate and faithful retrieval of fine-grained facts from long-term memory?
- **RQ2.** Can HERO better support reasoning across multiple sessions and events?
- **RQ3.** Can HERO capture users' historical events, traits, and preferences to enable personalized responses?
- **RQ4.** How efficient is HERO during retrieval?

4.1 Experimental Setup

Datasets and Metrics. We evaluate HERO on two benchmarks. LoCoMo (Maharana et al., 2024) contains 10 long-term conversations with about 600 turns, 16K tokens, and up to 32 sessions each, covering single-hop, multi-hop, temporal, and open-domain QA; we report F1, BLEU-1, and LLM-judged accuracy following the standard protocol (Chhikara et al., 2025; Hu et al., 2026). PERSONAMEM-32k (Jiang et al.) evaluates personalized memory over 180 simulated user–agent histories across real-world scenarios, covering user facts, evolving preferences, update reasons, and preference-aligned recommendations; we report multiple-choice accuracy.

Baselines. We compare HERO with state-of-the-art baselines. Mem0(Chhikara et al., 2025), MemoryOS(Kang et al., 2025b) and EverMemOS(Hu

Method	Multi-hop			Temporal			Open-domain			Single-hop			Overall Average		
	ACC	F1	BLEU-1	ACC	F1	BLEU-1	ACC	F1	BLEU-1	ACC	F1	BLEU-1	ACC	F1	BLEU-1
HERO	**85.11**	**43.52**	**32.57**	**86.60**	**63.63**	**57.82**	**72.92**	**34.67**	**30.13**	**91.20**	**59.81**	**52.15**	**87.99**	**56.06**	**48.38**
CAM	76.60	29.89	14.72	76.32	40.34	35.51	66.67	19.37	15.57	87.40	30.70	24.55	81.82	31.85	24.48
EverMemOS	81.91	41.55	28.31	83.80	59.08	52.09	69.79	25.76	19.33	86.44	43.78	36.06	84.03	45.48	36.94
LinearRAG	52.13	28.83	21.11	53.27	41.50	36.82	55.21	22.44	16.73	66.35	44.70	39.24	60.32	39.74	34.01
MemoryOS	57.45	35.11	26.62	44.86	32.05	26.68	53.13	23.57	12.77	74.67	52.10	45.47	63.96	43.03	36.06
Mem0	67.02	34.99	24.07	40.19	32.88	29.02	63.54	21.83	15.98	73.48	39.79	33.30	64.74	36.35	29.64

Table 1: Performance Comparison on the LoCoMo benchmark. All values represent the performance metric for each task (e.g., accuracy %). Best results are highlighted in bold.

Category	**R@5**	**R@10**
Single-hop	0.935	0.960
Multi-hop	0.915	0.947
Temporal	0.935	0.953
Open-domain	0.604	0.667
Overall	**0.910**	**0.938**

Table 2: Retrieval quality on the LoCoMo benchmark.

et al., 2026) rely on LLM-based rewriting or memory updates to transform raw dialogue into compact memory representations. CAM(Li et al., 2026) builds hierarchical summary nodes through topic clustering. LinearRAG(Zhuang et al., 2025) is a lightweight graph retrieval method.

Implementation Details. For the main experiments, methods are evaluated using Qwen2.5-72B-Instruct (Qwen et al., 2025) as the backbone LLM and Qwen3-Embedding-8B (Zhang et al., 2025b) as the embedding model. Mem0 is evaluated using its official hosted version in the main comparison. All methods share the same answer-generation prompts, as detailed in the Appendix.

4.2 Overall Performance

RQ1. HERO achieves leading response correctness and fidelity by preserving raw episodic context. We compare HERO with baselines on the LoCoMo dataset. The four LoCoMo task types focus on fact-based question answering, where answers must be retrieved from the dialogue history or inferred by combining multiple clues. As shown in Table 1, HERO achieves the best overall performance across all metrics, reaching 87.99% accuracy, 56.06% F1, and 48.38% BLEU-1. These results indicate that HERO not only produces correct answers, but also stays closer to the source dialogue text in wording, because its generation is directly grounded in raw text.

To further verify whether the gain comes from better evidence retrieval, we evaluate retrieval quality on LoCoMo using Recall@5 and Recall@10 against annotated supporting spans. As shown in Table 2, HERO achieves an overall Recall@10 of 0.938, indicating that the graph-based retriever can locate the relevant raw evidence in most cases. The lower recall on open-domain questions is expected, since such questions often require external commonsense or world knowledge not fully expressed in the dialogue history.

These results highlight the importance of memory storage strategy. Methods based on LLM rewriting or summarization perform irreversible compression before the downstream query is known, which may discard details or introduce semantic drift. This is consistent with prior analyses of LLM-generated event summaries, which report missing temporal or causal links, hallucinated details, and incorrect speaker attributions (Maharana et al., 2024). The case study in Table 3 further illustrates this issue: generated memories may preserve surface keywords while distorting causal links or event contexts.

RQ2. HERO demonstrates cross-session reasoning capabilities. As shown in Table 1, HERO achieves an accuracy of 85.11% on the Multi-hop tasks, which require linking scattered clues across multiple sessions. This result indicates that HERO can recover long-range dependencies rather than relying only on locally similar dialogue snippets. For example, in a LoCoMo instance where the query asks "`Who did John go to yoga with?`", the relevant evidence is split across two temporally distant sessions. In an earlier dialogue on `2023-02-25`, John stated that "`my colleague Rob invited me to a beginner's yoga class`"; in a later session on `2023-04-07`, he mentioned that he "`started a weekend yoga class with a colleague`" without naming the person. Neither utterance alone is sufficient to answer the query. HERO bridges them by following cue-based links from `yoga class` to

Question & Gold Answer	Original Context	Generated Memory
Q: What did Evan start painting due to a friend's gift? **A: forest scene**	*Evan shared a photo of a painting of a forest scene on an easel, and said, "Thanks, Sam! It all started when a friend of mine gave me this painting one day, it inspired me a lot and that's when I started painting..."*	*Memory Content: Evan thanked Sam and mentioned starting painting classes a few days ago. Evan shared a forest painting, inspired by a gift from a friend.*
Q: How often does Audrey take her pups to the park for practice? **A: Twice a week**	*Audrey shared a photo of a group of dogs sitting on a dirt road and said, "...it's a great physical and mental workout. I take them to the park twice a week for practice—it's been a great bonding experience..."*	*Memory Content: Audrey highlighted her twice-weekly agility classes and the positive impact on her dogs' social and physical development.*

Table 3: Case study comparing original context and generated memory on the LoCoMo benchmark.

Method	Generalize to New Scenarios	Provide Preference-Aligned Recommendations	Recall User Shared Facts	Acknowledge Latest User Preferences	Revisit Reasons behind Preference Updates	Suggest New Ideas	Track Full Preference Evolution	Overall Avg
HERO	**82.46**	**80.00**	68.99	**76.47**	**87.88**	**37.63**	**72.66**	**70.63**
CAM	70.18	54.55	65.89	**76.47**	75.76	36.56	61.87	61.63
Mem0	73.68	72.73	61.24	70.59	79.80	35.48	68.34	64.52
EverMemOS	80.70	72.73	68.99	**76.47**	83.84	22.58	71.22	66.38
LinearRAG	68.42	74.55	**72.09**	70.59	81.82	21.51	71.94	65.35
MemoryOS	**82.46**	67.27	65.12	70.59	86.87	22.58	71.94	65.70

Table 4: Performance comparison on the PERSONAMEM benchmark. All values represent the performance metric for each task. Best results are highlighted in bold.

`colleague` and then to `Rob`, thereby retrieving the cross-session evidence needed to infer the answer. This example illustrates how HERO's episodic graph supports multi-hop retrieval: HERO connects episodic traces through entity cues, allowing retrieval to follow the graph topology and perform multi-hop walks, which better captures long-range dependencies across sessions.

RQ3. HERO is able to capture dynamic human profile and evolving preferences. We evaluate HERO on PERSONAMEM, which focuses on tracking user facts, evolving preferences, and preference-aligned responses across long-term interactions. As shown in Table 4, HERO achieves the best overall accuracy of 70.63%, outperforming all baselines.

HERO shows clear advantages on preference-oriented categories, including Preference-Aligned Recommendations (80.00%), Revisit Reasons behind Preference Updates (87.88%), and Track Full Preference Evolution (72.66%). These results suggest that the profile layer helps identify user-specific cues that are not always explicit in the current query. Importantly, profiles in HERO are used as retrieval guidance rather than final evidence: they steer the retriever toward relevant raw dialogue traces, while the answer is generated from the retrieved original snippets. The ablation results in Figure 2 further support this design: removing the profile layer decreases PERSONAMEM accuracy from 65.70% to 63.67%, indicating that profile-enhanced retrieval provides complementary signals beyond query-only matching. This is consistent with the finding in (Jiang et al.) that relevant raw messages are important for personalized reasoning.

4.3 Trade-off Analysis

RQ4. HERO achieves an optimal trade-off between system efficiency and response performance. We further analyze the LLM token cost and time efficiency during retrieval on the LoCoMo benchmark. All efficiency metrics (latency and token usage) are reported as the average value per query. Figure 3 plots different methods along three dimensions: retrieval latency (x-axis, in log scale), F1 score (y-axis), and token consumption (bubble size). The bubble size represents the per-query LLM token consumption, including both input and output tokens.

HERO achieves high effectiveness with low retrieval overhead. It obtains the highest F1 while using only 248 tokens on average, and its latency remains below one second. This efficiency comes from HERO's design: heavy semantic linking and structure building are moved to the memory construction and profile maintenance stages. During retrieval, HERO only performs lightweight cue extraction followed by graph traversal, avoiding long chains of online LLM reasoning.

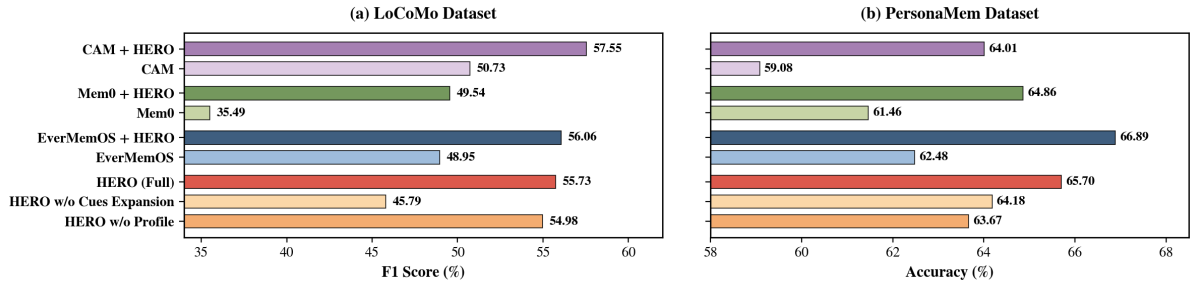

Figure 2: Ablation Studies of HERO.

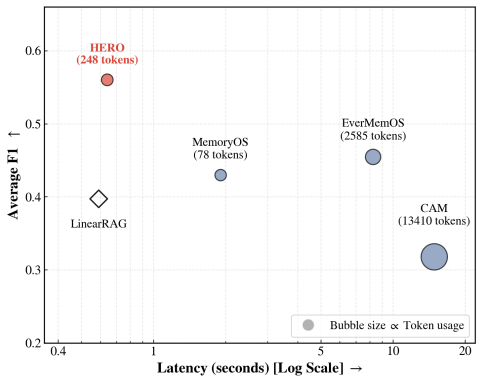

Figure 3: Performance–efficiency trade-off during retrieval on the LoCoMo benchmark.

Hyperparameter	Value	Acc	F1
τ_{alias}	0.6	0.818	0.552
	0.7	0.819	0.548
	0.8*	0.816	0.549
	0.9	0.816	0.546
$\tau_{\text{init}} = \tau_{\text{iter}}$	0.3	0.801	0.548
	0.4*	0.829	0.558
	0.5	0.814	0.557
	0.6	0.823	0.548
Top-k	5	0.814	0.556
	10*	0.828	0.560
	20	0.823	0.550
	40	0.818	0.555

Table 5: Hyperparameter sensitivity on LoCoMo. Default values are marked with *.

4.4 Ablation Study

To validate the effectiveness of HERO's components, we conduct an ablation study with two variants: (1) **HERO w/o Profile**, where the profile layer is deactivated to assess the impact of profile-enhanced retrieval; and (2) **HERO w/o Cues Expansion**, where iterative context expansion is disabled to evaluate multi-hop evidence retrieval. The results in Figure 2 demonstrate the importance of both components, with different effects across tasks. We also use HERO-retrieved raw dialogue segments as auxiliary evidence for existing baselines, which consistently improves their performance, further indicating the value of faithful raw-memory retrieval.

4.5 Hyperparameter Analysis

We examine whether HERO requires careful hyperparameter tuning. As shown in Table 5, LoCoMo accuracy remains within 0.801–0.829 and F1 within 0.546–0.560 when varying the alias matching threshold, retrieval filtering threshold, and retrieval top-k. These results show that HERO is stable under moderate hyperparameter changes and does not rely on delicate threshold tuning.

5 Conclusion

Long-term memory is a fundamental challenge for building personalized and coherent conversational agents. To empower episodic long-term memory for this task, we propose HERO, a novel human-profile enhanced retrieval optimization framework. HERO enables factually faithful reasoning through an episodic-centered heterogeneous memory graph and an iterative dual-cue activation mechanism, significantly improving personalized responses. Experiments on two benchmark datasets demonstrate encouraging performance gains, accurate retrieval capabilities, and robust cross-session reasoning. These results validate the effectiveness of our framework and suggest that preserving raw episodic context with structured retrieval is a promising direction for long-term agent memory.

Limitations

While HERO achieves promising results on long-term factual and personalized memory tasks, we acknowledge several areas for future exploration. First, although HERO organizes raw episodic traces through temporal chaining and semantic cue links,

the current implementation does not include an explicit memory lifecycle mechanism, such as time-aware decay, stale-trace pruning, or user-controlled deletion. Future work could explore incorporating these strategies as lightweight local graph updates for longer-running deployments. Second, HERO relies on LLMs to extract cues and profile insights. These profiles guide retrieval, but are not directly used as evidence for generation, mitigating the risk of grounding responses in noisy abstractions; however, inaccurate profile cues may still affect the retrieval trajectory. Future work could introduce confidence calibration, and timestamp-aware profile updates to further improve robustness.

Ethical Considerations

This work studies long-term memory for conversational agents, which involves storing and retrieving user dialogue history and personal profiles. All datasets used in this work are publicly released for research purposes and do not require additional ethical approval. For real-world deployment, we recommend that systems built on HERO adopt explicit data governance policies, including user-controlled memory deletion and access transparency, consistent with applicable data protection regulations.

We also note that inaccuracies in stored or retrieved memory could lead to misleading responses. HERO mitigates this risk by grounding generation in raw, unmodified dialogue text rather than LLM-rewritten summaries, reducing the chance of introducing hallucinated or distorted information.

References

Petr Anokhin, Nikita Semenov, Artyom Y. Sorokin, Dmitry Evseev, Andrey Kravchenko, Mikhail Burtsev, and Evgeny Burnaev. 2025. Arigraph: Learning knowledge graph world models with episodic memory for LLM agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025, Montreal, Canada, August 16-22, 2025*, pages 12–20. ijcai.org.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR.

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.

Bradford Dickerson and Howard Eichenbaum. 2009. The episodic memory system: Neurocircuitry and disorders. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology*, 35:86–104.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A graph RAG approach to query-focused summarization. *CoRR*, abs/2404.16130.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, and Zaiqiao Meng. 2025a. A comprehensive survey of self-evolving AI agents: A new paradigm bridging foundation models and lifelong agentic systems. *CoRR*, abs/2508.07407.

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. 2025b. Lightmem: Lightweight and efficient memory-augmented generation. *CoRR*, abs/2510.18866.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *CoRR*, abs/2410.05779.

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Haoyu Han, Yu Wang, Harry Shomer, Kai Guo, Jiayuan Ding, Yongjia Lei, Mahantesh Halappanavar, Ryan A. Rossi, Subhabrata Mukherjee, Xianfeng Tang, Qi He, Zhigang Hua, Bo Long, Tong Zhao, Neil Shah, Amin Javari, Yinglong Xia, and Jiliang Tang. 2025. Retrieval-augmented generation with graphs (graphrag). *CoRR*, abs/2501.00309.

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

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. 2026. EverMemOS: A Self-Organizing Memory Operating System for Structured Long-Horizon Reasoning. *arXiv e-prints*, arXiv:2601.02163.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen,

Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2):42:1–42:55.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose Taylor, and Dan Roth. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. In *Second Conference on Language Modeling*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025a. Memory OS of AI agent. *CoRR*, abs/2506.06326.

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

Zerrin Kasap and Nadia Magnenat-Thalmann. 2010. Towards episodic memory-based long-term affective interaction with a human-like robot. In *19th IEEE International Conference on Robot and Human Interactive Communication, Viareggio, Italy, RO-MAN, 2010, September 13-15, 2010*, pages 452–457. IEEE.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John F. Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025a. Hello again! llm-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 5259–5276. Association for Computational Linguistics.

Rui Li, Zeyu Zhang, Xiaohe Bo, Zihang Tian, Xu Chen, Quanyu Dai, Zhenhua Dong, and Ruiming Tang. 2026. Cam: A constructivist view of agentic memory for llm-based reading comprehension. *Advances in Neural Information Processing Systems*, 38:113381–113406.

Zhiyu Li, Shichao Song, Hanyu Wang, Simin Niu, Ding Chen, Jiawei Yang, Chenyang Xi, Huayi Lai, Jihao Zhao, Yezhaohui Wang, Junpeng Ren, Zehao Lin, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, Zhiqiang Yin, Qingchen Yu, Bo Tang, and 3 others. 2025b. Memos: An operating system for memory-augmented generation (MAG) in large language models. *CoRR*, abs/2505.22101.

Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *CoRR*, abs/2308.08239.

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

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. Memgpt: Towards llms as operating systems. *CoRR*, abs/2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Dmitrii Pantiukhin, Boris Shapkin, Ivan Kuznetsov, Antonia Anna Jost, and Nikolay Koldunov. 2025. Accelerating earth science discovery via multi-agent LLM systems. *Frontiers Artif. Intell.*, 8.

Mathis Pink, Qinyuan Wu, Vy Ai Vo, Javier Turek, Jianing Mu, Alexander Huth, and Mariya Toneva. 2025. Position: Episodic memory is the missing piece for long-term LLM agents. *CoRR*, abs/2502.06975.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

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 *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D. Manning. 2024. RAPTOR: recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Katja Schlegel, Nils Sommer, and Marcello Mortillaro. 2025. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*, 3.

Eva Svoboda, Margaret Mckinnon, and Brian Levine. 2006. The functional neuroanatomy of autobiographical memory: A meta-analysis. *Neuropsychologia*, 44:2189–208.

Annick Tanguay, Daniela Palombo, Brittany Love, Rafael Glikstein, Patrick Davidson, and Louis Renoult. 2023. The shared and unique neural correlates of personal semantic, general semantic, and episodic memory. *eLife*, 12.

LangChain Team. 2023. Conversational rag.

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

John Yang, Carlos E. Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Wenhao Yu, Hongming Zhang, Xiaoman Pan, Peixin Cao, Kaixin Ma, Jian Li, Hongwei Wang, and Dong Yu. 2024. Chain-of-note: Enhancing robustness in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 14672–14685. Association for Computational Linguistics.

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, Fei Huang, and Jingren Zhou. 2025b. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv: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 *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2025. Linearrag: Linear graph retrieval augmented generation on large-scale corpora. *CoRR*, abs/2510.10114.

A Appendix

A.1 Experimental Settings

Unless otherwise specified, all methods use Qwen2.5-72B-Instruct as the default LLM for memory construction and answer generation, and Qwen3-Embedding-8B as the embedding model. All LLMs in our experiments are accessed via commercial API services.

For ablation and hyperparameter analysis experiments, we use Qwen3-30B-A3B to reduce computational cost. Since all HERO variants are evaluated under the same backbone and pipeline, these results reflect the relative contribution of each component rather than absolute scores comparable to the main results.

For all methods, answer generation and evaluation are conducted under identical prompts and LLMs. Other components follow the configurations described in their original papers. In particular, EverMemOS and Mem0 are evaluated using the EverMemOS evaluation pipeline, which provides a standardized and publicly available implementation for memory-based systems and ensures consistent preprocessing and metric computation.

For Mem0, we report the official hosted version in the main comparison, as it represents the strongest available implementation.

On both the LoCoMo and PersonaMem benchmarks, the chunk size is set to 128. We set retrieval top-k to 10 on LoCoMo and 40 on PersonaMem.

A.2 Effect of Backbone LLMs

Table 6 reports LoCoMo overall accuracy under different backbone LLMs. Under the GPT-4o-mini setting, HERO achieves 0.870 accuracy, slightly higher than the strongest reported baseline EverMemOS 0.868. HERO also remains stable when

Backbone LLM	Method	ACC
GPT-4o-mini	MemoryOS†	0.547
	Mem0†	0.610
	MemU†	0.612
	MemOS†	0.759
	Zep†	0.811
	EverMemOS†	0.868
	HERO	0.870
Qwen2.5-72B-Instruct	**HERO**	0.880
Qwen3-30B-A3B	**HERO**	0.866

Table 6: LoCoMo overall accuracy under different backbone LLMs. Results marked with † are reported by EverMemOS (Hu et al., 2026); other results are from our evaluation.

Dataset	Top-k	ACC
LoCoMo	5	0.814
	10*	0.828
	20	0.823
	40	0.818
PERSONAMEM	20	0.642
	30	0.659
	40*	0.657
	50	0.667
	60	0.662

Table 7: Sensitivity to retrieval top-k. Both datasets report accuracy. Default values are marked with *.

using different Qwen backbones, achieving 0.880 with Qwen2.5-72B-Instruct and 0.866 with Qwen3-30B-A3B. These results suggest that HERO's gains are not tied to a single generation model.

A.3 Selection of Retrieval Top-k

We conduct a sensitivity analysis to examine whether HERO depends on a specific retrieval top-k. As shown in Table 7, HERO remains stable across different top-k values on both LoCoMo and PERSONAMEM. On LoCoMo, top-$k = 10$ achieves the best accuracy among the tested values. As for PERSONAMEM, it requires tracking dynamic profiles across multiple sessions, meaning evidence is highly scattered. The performance remains stable from top-$k = 30$ to top-$k = 60$; we use top-$k = 40$ as the default because it provides a practical coverage–cost trade-off while retrieving fewer traces than larger settings.

A.4 Scalability and Construction Cost

We report the graph size and construction cost of HERO on LoCoMo and PERSONAMEM.

Table 8 shows the average and maximum graph size per user. "Tr", "Cu", and "Pr" denote trace, cue, and profile nodes. The memory footprint stays below 3.4 MB per user in all cases.

Dataset	Memory (MB)	Tr/Cu/Pr Nodes	Edges
LoCo avg	2.39	209/1163/1234	9181
LoCo max	2.83	244/1398/1444	10941
PM avg	2.73	271/1645/1241	9810
PM max	3.35	338/2030/1528	12132

Table 8: Per-user graph statistics of HERO.

Dataset	Total Time (s)	Cue Tok.	Profile Tok.
LoCo	294.5	73,278	108,991
PM	345.5	66,237	162,107

Table 9: Graph construction overhead of HERO.

Table 9 reports the graph construction cost. On average, construction takes less than 6 minutes and fewer than 230K tokens per user. The graph is updated incrementally and does not require global recomputation.

A.5 Prompt Templates

A.5.1 Profile Extraction

```
### Role
You are a Knowledge Extraction System for a
↪  memory graph. Your goal is to extract key
↪  information from conversations to update the
↪  user's long-term memory.
Return JSON only: {"facts": ["...", "..."]}
### What to Extract
Extract items into a flat list, covering these
↪  three categories:
1. **Facts**
One-time, specific actions or events.
Example: "User visited Tokyo"
2. **User Insights**
Situational states or emotions
Example: "User is currently injured", "User is
↪  currently feeling anxious about exams"
3. **human profiles**
Stable or recurring traits, habits, or
↪  preferences.
Example: "User prefers spicy food", "Tom is a
↪  software engineer"
### Rules
1. **Third-Person Perspective**: Convert all
↪  first-person dialogue into third-person
↪  objective facts.
2. **Stand-Alone Sentences**: The output must
↪  make sense **without context**.
- BAD: "He liked *it*."
- GOOD: "User liked *the sci-fi movie 'Dune'*."
- Replace pronouns (he, she, it, that) with
↪  specific entities.
3. **Atomic**: Each string must contain only ONE
↪  key piece of information.
4. **No Redundancy**: Do not extract "User said
↪  that..." or "User mentioned...". Just state
↪  the fact directly.
```

```
5. **Resolve Time**: If relative time is
↪  mentioned (yesterday, last week), try to make
↪  it specific contextually; otherwise, keep
↪  the original phrasing but ensure clarity.

### Example
Input:
Tom: "I'm exhausted. I stayed up until 3 AM
↪  coding that Python script."
AI: "Do you do this often?"
Tom: "Yeah, I'm a night owl. I usually work best
↪  after midnight."
Output:
{"facts": [
  "Tom stayed up until 3:00 AM coding a Python
  ↪  script",
  "Tom is currently exhausted",
  "Tom usually works best after midnight"]}
```

A.5.2 Answer Generation (LoCoMo)

```
You are an intelligent memory assistant tasked
↪  with retrieving accurate information from
↪  episodic memories.
# DATA EXPLANATION:
The "Information" section contains original
↪  conversation fragments, each prefixed with a
↪  timestamp.
Note that the information is NOT necessarily
↪  sorted chronologically. You must carefully
↪  distinguish the timestamps and speakers
↪  within each fragment.

# INSTRUCTIONS:
Your goal is to synthesize information from all
↪  relevant memories to provide a comprehensive
↪  and accurate answer.
You MUST follow a structured Chain-of-Thought
↪  process to ensure no details are missed.
Actively look for connections between people,
↪  places, and events to build a complete
↪  picture. Synthesize information from
↪  different memories to answer the user's
↪  question.

# CRITICAL REQUIREMENTS:
1. NEVER omit specific names - use "Amy's
↪  colleague Rob" not "a colleague"
2. ALWAYS include exact numbers, amounts, prices,
↪  percentages, dates, times
3. PRESERVE frequencies exactly - "every Tuesday
↪  and Thursday" not "twice a week"
4. MAINTAIN all proper nouns and entities as they
↪  appear
5. For time-related answers, use this format: 20
↪  January 2026
6. Answer with exact words from the information
↪  context whenever possible
7. Double-check that your answer directly
↪  addresses the question asked
8. Ensure your final answer is specific and
↪  avoids vague time references

# Example:
Information: [2023-03-15]: Manry said, "I went to
↪  the vet yesterday."
Question: What day did Manry go to the vet?
Correct Answer: March 15, 2023
Explanation: Even though the phrase says
↪  "yesterday," the timestamp shows the event
↪  was recorded as happening on March 15th.
↪  Therefore, the actual vet visit happened on
↪  that date, regardless of the word "yesterday"
↪  in the text.

Information: [2023-03-15]: Manry said, "I went to
↪  the week before March 15, 2023."
Question: When did Manry go to the vet?
Correct Answer: the week before March 15, 2023.

## Output Format:
Please output the result in JSON format
↪  containing the following keys:
- **thought**: A step-by-step logical breakdown
↪  in the language of the input. Locate the
↪  specific sentence and identify the real
↪  names/dates.
- **answer**: The final, concise result (e.g., a
↪  specific name, date, or location) in the form
↪  of a short phrase.

Example:
```json
{
"thought": "Scanning the memories...",
"answer": "20 January 2026"
}
```
```

A.5.3 Answer Generation (PERSONAMEM)

```
# INSTRUCTIONS:
1. Carefully read all the conversation history
↪  provided as context
2. The question is a multiple-choice question
↪  with options (a), (b), (c), (d)
3. Select the option that **best demonstrates
↪  accurate memory of the user's history and
↪  preferences**.
4. **Prioritize specific, personalized responses
↪  over generic ones, but only if the historical
↪  details mentioned are factually correct
↪  based on the context.**
5. Pay attention to the reasons behind preference
↪  changes mentioned in conversations

# CRITICAL REQUIREMENTS:
1. Base your answer ONLY on the information
↪  provided in the context
2. If preferences have evolved, use the MOST
↪  RECENT preference
3. Verify that any "memories" mentioned in the
↪  options (e.g., "I remember you said...")
↪  strictly align with the provided text
4. Do NOT make assumptions beyond what is stated
↪  in the conversations

## Output Format:
- **Thought**: Analysis of relevant information
↪  from the conversation history
- **Answer**: Your final answer as a single
↪  letter: (a), (b), (c), or (d)
```

A.5.4 EVALUATION

```
You are an expert grader that determines if
↪  answers to questions match a gold standard
↪  answer.
Your task is to label an answer to a question as
↪  'CORRECT' or 'WRONG'. You will be given the
↪  following data:
```

```
(1) a question (posed by one user to another
↪  user),
(2) a 'gold' (ground truth) answer,
(3) a generated answer
which you will score as CORRECT/WRONG.

The generated answer may be more verbose.
↪  However, as long as the core information and
↪  factual claims are semantically consistent
↪  with the gold answer, it should be marked as
↪  CORRECT. Do not penalize for length or extra
↪  explanations.

Now it's time for the real question:
Question: {question}
Gold answer: {gold_answer}
Generated answer: {generated_answer}

First, provide a short (one sentence) explanation
↪  of your reasoning, then finish with CORRECT
↪  or WRONG.
Do NOT include both CORRECT and WRONG in your
↪  response. Just return the label CORRECT or
↪  WRONG in a json format with the key as
↪  "label".
```

A.6 Case study

As shown in Table 10, this case illustrates how HERO's dual-path retrieval integrates episodic evidence and profile guidance to support personalized reasoning. Starting from the initial cue LIVING SPACE, HERO activates two complementary paths. The episodic path surfaces concrete dialogue segments about vintage travel maps as home decor, while the profile path injects the user's long-term preference for visual storytelling and hobby-driven decoration. These two paths converge on a shared semantic theme—personal collections as meaningful decor—which allows HERO to retrieve multiple supporting memories and correctly ground its decision on option (c). Option (a) suggests natural elements and (d) suggests bold colors or geometric patterns, neither of which is supported by the retrieved evidence. Option (b) refers to personal collections but remains too generic to capture the user's specific interests. In contrast, option (c) directly aligns with the user's core preference for vintage and historically grounded artifacts, and appropriately generalizes this preference to related design elements, making it the most well-supported choice.

Stage	Content
Query	"I've been thinking about ways to make my living space more unique and personal. Any suggestions for incorporating interesting design elements?"
Options	**(a)** You might consider incorporating natural elements... **(b)** For a personalized space, consider showcasing personal collections or heirlooms. These items not only add uniqueness but also reflect your personal history and tastes... **(c)** One intriguing approach could be exploring vintage or antique elements, such as old maps, blueprints, or historical artwork... **(d)** A dramatic approach could be using bold colors or geometric patterns. This could add a modern flair and energize your living area. Such elements often speak to personal style ...
Cue Activation	**Initial match:** `[LIVING_SPACE, DESIGN_ELEMENTS]` **Path A (Episodic Association):** Context: "Assistant said, Collecting vintage travel maps enriches not only your living space with..." → Activated cues: `[VINTAGE_TRAVEL_MAPS]` → Associated cues: `[VINTAGE_TRAVEL_MAPS, VISUAL_STORYTELLING]` **Path B (Profile Guidance):** Profile insight: "User's hobby enriches their living space with visual stories" → Activated cues: `[HOBBY, VISUAL_STORIES]` → Associated cues: `[HOBBY, VISUAL_STORIES, VINTAGE_MAPS]` ...
Retrieved Episodic Traces	**T1 (User):** "...I've started to appreciate the aesthetic value of collecting travel maps! They're becoming beautiful decor in my home. The vibrant colors and intricate details found in each map really draw the eye, making every piece unique in its own right..." **T2 (User):** "I also started a collection of vintage travel maps to display in my home..." **T3 (Assistant):** "...Collecting vintage travel maps enriches not only your living space with visual stories but also your mind with historical context..." ...
HERO Answer	**Correct option:** "One intriguing approach could be exploring vintage or antique elements, such as old maps, blueprints, or historical artwork. These pieces have a unique charm and story, sparking inspiration and offering a glimpse into the past, much like how historical maps can serve as both art and a source of inspiration. If you're interested, I can share more about how to source or incorporate such elements into your decor."

Table 10: Case study illustrating HERO's retrieval path on the PERSONAMEM benchmark, where episodic associations and profile guidance jointly activate relevant cues and retrieve supporting memories.