Bi-Mem: Bidirectional Construction of Hierarchical Memory for Personalized LLMs via Inductive-Reflective Agents

Anonymous ACL submission

Abstract

Constructing memory from users' long-term conversations overcomes LLMs' contextual limitations and enables personalized interactions. Recent studies focus on hierarchical memory to model users' multi-granular behavioral patterns via clustering and aggregating historical conversations. However, conversational noise and memory hallucinations can be amplified during clustering, causing locally aggregated memories to misalign with the user's global persona. To mitigate this issue, we propose Bi-Mem, an agentic framework ensuring hierarchical memory fidelity through bidirectional construction. Specifically, we deploy an inductive agent to form the hierarchical memory: it extracts factual information from raw conversations to form fact-level memory, aggregates them into thematic scenes (*i.e.,* local scene-level memory) using graph clustering, and infers users' profiles as global persona-level memory. Simultaneously, a reflective agent is designed to calibrate local scene-level memories using global constraints derived from the persona-level memory, thereby enforcing global-local alignment. For coherent memory recall, we propose an associative retrieval mechanism: beyond initial hierarchical search, a spreading activation process allows facts to evoke contextual scenes, while scene-level matches retrieve salient supporting factual information. Empirical evaluations demonstrate that Bi-Mem achieves significant improvements in question answering performance on long-term personalized conversational tasks.

1 Introduction

Personalization of Large Language Models (LLMs) (Chen et al., 2024; Zhang et al., 2025b) aims to address user-specific requirements, such as recalling shared experiences and generating preference-aligned suggestions based on historical conversations. To facilitate personalized interactions in long-term conversational tasks, constructing memory (Chhikara et al., 2025; Li et al., 2025; Xu et al., 2025) has emerged as a critical paradigm due to the limitation of LLMs' context windows. Early work focused on extracting key entities or dialogue summaries from historical conversations as factual memories (Packer et al., 2023; Chhikara et al., 2025). However, such fragmented factual memories struggle to capture inter-fact relationships or users' high-level behavioral patterns (Rezazadeh et al., 2025; Pan et al., 2025). Consequently, recent research has explored clustering and aggregating these isolated facts with different granularities to form a hierarchical memory structure (*e.g.,* multi-level graph (Li et al., 2025; Gutierrez et al., 2024) or tree structure (Rezazadeh et al., 2025)). By facilitating coarse-to-fine retrieval (Zhao et al., 2025), the hierarchical systems provide more comprehensive contextual knowledge, enabling more coherent and personalized interactions.

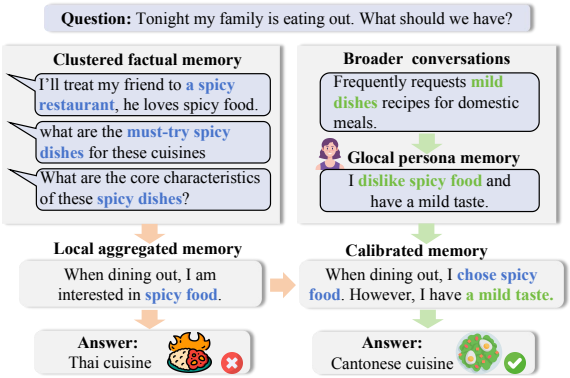

Figure 1: Illustration of local aggregated memory misaligning with the user's global persona in naive hierarchical memory systems, leading LLMs to generate persona-violating answers.

Despite the success of hierarchical memory, such clustering and aggregation face two critical limitations: noise amplification and cumulative hallucinations (Huang et al., 2025; Zhang and Zhang,

2025). First, conversational noise — such as irrelevant chitchat and incidental interaction — is often amplified within local clusters, causing aggregated local memory to skew toward trivialities. Second, fact-level hallucinations (Chen et al., 2025) and extraction errors propagate and accumulate during hierarchical consolidation. Together, these "cascading errors" cause locally aggregated memories to misalign with the user's global persona (Wang et al., 2025b). As illustrated in Figure 1, consider a user who frequently requests mild recipes for home-cooked meals but seeks spicy dishes for a one-off social dinner with spicy-loving friends. Naive clustering tends to amplify this transient social context, leading to the aggregation of these contexts into a local behavioral rule: "prefers spicy cuisine when dining out." In contrast, the global persona derived from broader conversations is "dislikes spicy food and has a mild taste." Consequently, the aggregated local memory diverges from the user's global persona, leading LLMs to generate persona-violating suggestions for family dining-out scenarios.

To tackle these problems, we propose Bi-Mem, an agentic framework ensuring hierarchical memory fidelity via bidirectional construction, rather than unidirectional aggregation (Li et al., 2025). Specifically, we employ an inductive agent to manage bottom-up hierarchical structure formulation: extracting atomic fact-level memory from raw conversations, aggregating them into thematic scene-level memory via graph clustering (Pattnaik et al., 2024), and further distilling them into a five-dimensional persona-level memory (Zhong et al., 2024) as a global constraint. Meanwhile, the reflective agent is designed to perform top-down calibration, imposing global constraints on local scene memories to enforce global-local alignment. As shown in Figure 1, calibrating the local scene memory by injecting the global constraint of "I have a mild taste" resolves the global-local contradictions, thereby enabling LLMs to generate preference-aligned suggestions. To connect memories of different granularities, we introduce an associative retrieval mechanism. Beyond the initial hierarchical search, it adopts a spreading activation process (Zhang et al., 2025a): retrieved facts evoke contextual thematic scenes, while scene-level matches retrieve salient supporting facts. Empirical results demonstrate that Bi-Mem significantly improves question answering performance in personalized long-term conversational tasks.

The main contributions of this work are summarized as follows:

- We identify global persona-local aggregated memory misalignment as a general challenge for hierarchical memory, which is induced by noise and hallucination.
- We design Bi-Mem, an agentic framework that ensures hierarchical memory fidelity through bidirectional construction (*i.e.,* an inductive and a reflective process).
- We propose an associative retrieval mechanism to facilitate coherent memory recall, which enhances the integration of granular factual information and contextual scenes.
- Extensive empirical evaluations on complex personalized long-term conversational tasks demonstrate the effectiveness of Bi-Mem.

2 Related Work

Personalized LLMs. Personalized LLMs aim to tailor model responses to align with specific user preferences and personalities based on user-LLM interaction histories (Chen et al., 2024; Zhang et al., 2025b). Instead of just delivering general knowledge (Kim et al., 2023), they meet user-specific needs, like recalling shared experiences or generating preference-aligned suggestions—boosting engagement and satisfaction. To handle long-term interaction histories, existing approaches either fine-tune user-specific modules (Tan et al., 2024b,a) on conversational data or adopt memory architectures (Wang et al., 2025b; Zhong et al., 2024) to store and manage user-specific conversations without modifying model parameters. Considering the computational resource of fine-tuning and LLMs' context window limitation, we focus on building memory for users' long-term conversational history to enable such personalization.

Memory for Personalized LLMs. Personalized LLMs' memory (Zhang et al., 2024; Hu et al., 2025) is a flexible component that stores and leverages users' historical conversation information to supplement the LLMs' context window limitations. It mainly consists of two parts: memory construction (Xu et al., 2025)(extracting key information from user-LLM interactions and managing it into structured forms) and memory retrieval (Wang et al., 2025a)(retrieving relevant memory through semantic similarity to support LLMs' response generation). Mainstream implementations adopt token-based storage. For example, Mem0 (Chhikara et al., 2025) constructs mem-

ory from conversational information by supporting ADD/UPDATE/DELETE/NOOP operations and retrieves relevant memory through vector similarity. MemoryBank (Zhong et al., 2024) empowers personalized LLMs by integrating human-like memory storage (dialogues, event summaries, user portraits) and dense retrieval.

Hierarchical Memory. Token-based memory for personalized LLMs has two core structures: Flat Memory (Packer et al., 2023) stores information as independent units (*e.g.,* raw interaction and summarized sessions) without explicit connections. Hierarchical Memory (Hu et al., 2025) uses multi-level architectures with different granularities and associations, supporting complex reasoning via cross-layer links. For example, CAM (Li et al., 2025) constructs hierarchical memory by adopting an incremental overlapping clustering algorithm and retrieves by a Prune-and-Grow strategy. HippoRAG (Gutierrez et al., 2024) employs a hierarchical schemaless knowledge graph (KG) for memory construction and leverages Personalized PageRank for cross-memory reasoning.

3 Preliminary

3.1 Task Formulation

Let $\mathcal{C} = \{c_1, ..., c_N\}$ denote the long-term conversational history between a user and an LLM agent, where each $c_i = (q_i, r_i)$ denotes the i-th interaction (user request q_i, LLM response r_i). The task of the memory agent is composed of a memory construction and retrieval task (Pan et al., 2025).

- **Memory construction:** Construct a structured memory bank $\mathcal{M}$ by encoding conversational history $\mathcal{C}$ via a memory construction function:

$$\mathcal{M} = f_{\text{cons}}(\mathcal{C}) \tag{1}$$

  where f_{cons} constructs $\mathcal{C}$ into discrete memory units $\boldsymbol{m} \in \mathcal{M}$.
- **Memory Retrieval:** Given a target user query q^* and a memory bank $\mathcal{M}$, extract K memory units relevant to q via a retrieval function f_R:

$$\{\boldsymbol{m}_k \in \mathcal{M}\}_{k=1}^{K} \leftarrow f_R(q^*, \mathcal{M}, K) \tag{2}$$

- **Response Generation:** The goal of LLM personalization (Wang et al., 2025b) is to generate an optimal personalized response r^* for the current query q^* based on retrieved N memory units:

$$r^* = f_{\text{LLM}}\left(q^*, \{\boldsymbol{m}_k\}_{k=1}^{K}\right) \tag{3}$$

where r^* aligns with the user's persona or accurately recalls relevant contextual details.

4 Method

In this section, we elaborate on our proposed BiMem, an agentic framework with bidirectional construction for hierarchical memory and associative retrieval, as shown in Figure 2. We first define the three levels of our designed hierarchical memory in 4.1. Then, we present bidirectional memory construction in Section 4.2, including the inductive-reflective process, followed by the details of the associative retrieval mechanism in Section 4.3. The detailed algorithm is presented in Appendix A.

4.1 Definition for Hierarchical Memory

The hierarchical memory $\mathcal{M} = (\mathcal{F}, \mathcal{S}, \mathcal{P})$ in BiMem is structured into three granularities:

- **Fact-level memory** ($\mathcal{F}$) is the atomic unit of memory, where each conversational interaction maps to a single fact unit. Formally, $\mathcal{F} = \{f_1, f_2, ..., f_N\}$, where each unit f_i is defined as the tuple $(i, \text{con}_i, t_i, e_i)$: i is the index of fact, con_i is the summarized content of the interaction, t_i is the timestamp, and e_i is the connection edges linking to related fact or interaction indices.
- **Scene-level memory** ($\mathcal{S}$) clusters related facts to form a thematic scene (*e.g.,* "dining-out with spicy-loving friends"). Formally, $\mathcal{S} = \{s_1, s_2, ..., s_J\}$, where each scene s_j is the aggregation of a subset of fact-level memory $\bar{\mathcal{F}}_j \subseteq \mathcal{F}$: $s_j = \frac{1}{|\bar{\mathcal{F}}_j|} \sum_{f \in \bar{\mathcal{F}}_j} f$.
- **Persona-level memory** ($\mathcal{P}$) represents the user's core profile, distilled from all aggregated scenes $\mathcal{S}$. It serves as a global constraint to ensure memory consistency. Formally, $\mathcal{P} = [p_1, p_2, p_3, p_4, p_5]$ is a five-dimensional persona, where p_1 (basic-info) covers name, age, occupation, and location, p_2 (interests) covers hobbies, likes, and dislikes, p_3 (personality) covers personality traits, emotional patterns, p_4 (values) covers core values, beliefs, and motivations, and p_5 (relationships) covers key social relationships.

4.2 Bidirectional Memory Construction

Bidirectional construction integrates a bottom-up inductive process (for memory formulation) and a top-down reflective process (for top-down calibration), realizing the global-local alignment.

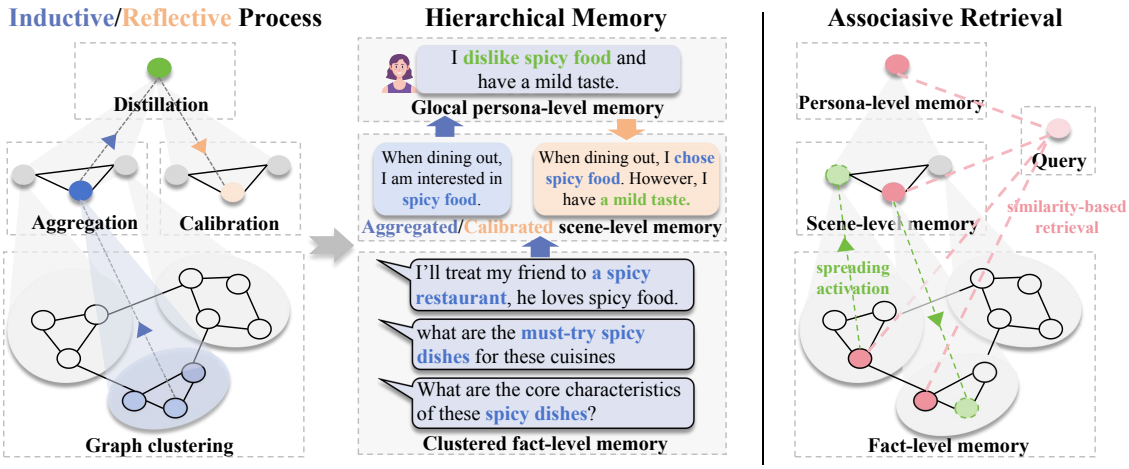

Figure 2: The framework of our proposed Bi-Mem, including the bidirectional construction (inductive process and reflective process) of hierarchical memory and the associative retrieval.

4.2.1 Inductive Process (Bottom-Up)

The inductive agent constructs the hierarchical memory from raw conversations through three sequential steps:

Fact-level Memory Extraction: To construct the fact unit $f_i = (i, \text{con}_i, t_i, e_i)$ for each interaction c_i in the conversation, we first apply a summarization operator $\mathcal{E}$ via an LLM to extract the summarized interaction content con_i, defined as:

$$\text{con}_i = \arg\max_{\text{con}\in\text{Con}} P_{\mathcal{E}}(\text{con}|c_i), \tag{4}$$

where $i \in [1, 2, \ldots, N]$, and Con is the candidate set of generated summaries by LLMs. The index i and timestamp t_i are directly assigned from the interaction metadata. The connection edges e_i are identified by computing the semantic similarity between the current content con_i and con_l from other fact units. Formally, $l \in e_i$ if:

$$\text{sim}(\phi(\text{con}_i), \phi(\text{con}_l)) > \tau, \tag{5}$$

where $\text{sim}(\cdot)$ is the cosine similarity of embeddings obtained by a pretrained embedding model ϕ, and τ is a predefined threshold.

Scene-Level Memory Aggregation: To provide more contextual knowledge and capture users' behavioral patterns, we transform fragmented atomic facts into high-level thematic scenes. We model the factual memory as an undirected graph $G = (\mathcal{F}, E)$, where an edge $(f_i, f_l) \in E$ is established if $l \in e_i$ or $i \in e_l$. We then perform graph clustering by applying the Label Propagation Algorithm (LPA) (Ugander and Backstrom, 2013) to G. This process utilizes local structural consensus (Lancichinetti and Fortunato, 2012) to partition the fact set $\mathcal{F}$ into J thematic clusters $\{\bar{\mathcal{F}}_1, \bar{\mathcal{F}}_2, \ldots, \bar{\mathcal{F}}_J\}$. Formally, for each fact cluster $\bar{\mathcal{F}}_j \subseteq \mathcal{F}$, we apply an aggregation operator $\mathcal{A}$ with LLMs to derive the corresponding scene s_j by maximizing the generation probability:

$$s_j = \arg\max_{s\in\mathcal{S}_{can}} P_{\mathcal{A}}(s \mid f_k \in \bar{\mathcal{F}}_j), \tag{6}$$

where $\mathcal{S}_{can}$ is the set of candidate generated scenes. Thus, we can construct thematic scene-level memory $\mathcal{S} = \{s_1, s_2, \ldots, s_J\}$.

Persona-Level Memory Distillation: To obtain the global constraint of memory, we distill the five-dimensional persona $\mathcal{P}$ from the entire scene-level memories, as detailed in 4.1. Specifically, we define a distillation operator $\mathcal{D}$ that leverages an LLM to project the set $\mathcal{S}$ onto each persona dimension:

$$p_d = \arg\max_{p\in\mathcal{P}_{cand}} P_{\mathcal{D}}(p \mid \mathcal{S}, \text{per}_d), \tag{7}$$

where $d \in 1, \ldots, 5$, per_d is a dimension-specific instruction that guides the LLM to focus on relevant scenes within $\mathcal{S}$.

4.2.2 Reflective Process (Top-Down)

To enforce global-local alignment and mitigate cascading errors from the inductive process, we implement a reflective agent that performs top-down calibration, ensuring that each local scene $s_j \in \mathcal{S}$ is grounded in the stable persona constraints $\mathcal{P}$.

Specifically, the reflective agent first assesses whether the local scene s_j contradicts or fails to

reflect the related global persona within $\mathcal{P}$. It then applies a Calibration Operator $\mathcal{R}$ to generate a compensatory condition Δs_j that explicitly bridges the local scene s_j with the global persona:

$$\Delta s_j = \arg \max_{\delta \in \Delta_{cand}} P_{\mathcal{R}}(\delta \mid s_j, \mathcal{P}), \quad (8)$$

where Δ_{cand} denotes the space of candidate compensatory conditions. The calibrated scene is formally defined as: $s'_j = s_j \oplus \Delta s_j$, where Δs_j is empty if the local scene is already consistent with the global persona. This reflective process injects global constraints on local scenes, resulting in a self-consistent hierarchical memory with scenes $\mathcal{S}' = \{s'_1, s'_2, \ldots, s'_J\}$. Since fact-level memory consists of detailed conversation facts, we refrain from calibrating it to align with reality.

4.3 Associative Memory Retrieval

To enable coherent recall across memory granularities, we propose an associative retrieval mechanism. This mechanism leverages Spreading Activation after initial search in hierarchical memory $\mathcal{M} = (\mathcal{F}, \mathcal{S}', \mathcal{P})$, balancing granular detail with high-level context.

Initial Hierarchical Search: Given a query q^*, we first project both the query and all memories into a latent space using a pre-trained embedding model $\phi(\cdot)$. Initial relevance scores a_0 are computed as the cosine similarity between the query and each memory unit:

$$a_0(x) = \text{sim}(\phi(q^*), \phi(x)), \quad (9)$$

where $x \in \mathcal{F} \cup \mathcal{S}' \cup \mathcal{P}$ is the memory units in hierarchical memory, $\text{sim}(\cdot)$ is the cosine similairty. The top-k x with the highest a_0 scores are selected as the retrieved memory set M_{ret}.

Associative Spreading Activation: To capture structural dependencies between granular facts and thematic scenes, we implement a bidirectional associative mechanism: the retrieved fact invokes its parent scene, while the retrieved scene triggers the recall of its most representative facts. Specifically, starting from the initial retrieved memory set $\mathcal{M}_{ret}$, the retrieval set is expanded through two associative paths. First, for each fact $f \in \mathcal{M}_{ret}$, its unique parent scene s is automatically associated and added to the expanded set. Second, for each scene $s_j \in \mathcal{M}_{ret}$, the mechanism spreads activation to its corresponding fact cluster $\bar{\mathcal{F}}_j$. We select the top-m facts with the highest relevance score a_1 to the scene:

$$a_1(x) = \text{sim}(\phi(s_j), \phi(x)), \quad \forall x \in \bar{\mathcal{F}}_j. \quad (10)$$

And for each persona $p \in \mathcal{M}_{ret}$, we treat it as a global anchor and do not associate it with other memories. The final retrieved memory set $\mathcal{M}_{ret}$ is updated as the union of the initial retrieved ones and those associated ones.

Response Generation: The final retrieved memory set $\mathcal{M}_{ret}$, which encompasses multi-granular evidence from atomic facts to global persona traits, is integrated into a unified prompt context. Formally, given the current query q^* and the retrieved information in $\mathcal{M}_{ret}$, the LLMs generates a personalized response r^* by maximizing the conditional probability:

$$r^* = \arg \max P_{\text{LLM}}(r \mid q^*, \mathcal{M}_{ret}). \quad (11)$$

5 Experiments

In this section, we conduct extensive experiments to evaluate the effectiveness of our proposed framework in long-term personalized conversational tasks by answering the following questions:

- RQ1: How does our framework Bi-Mem perform compared with leading baselines of memory?
- RQ2: What contribution does each of the Fact-Scene-Persona levels and reflective calibration make to constructing the Hierarchical Memory?
- RQ3: To what extent does the Associative Retrieval mechanism improve the QA accuracy?
- RQ4: How sensitive is the framework to the hyperparameters of initial retrieval size k?
- RQ5: How efficient is the Bi-Mem compared to existing memory frameworks?

5.1 Expermental Settings

Datasets. We adopt LoCoMo (Maharana et al., 2024) to conduct our experiments, which is a widely-used dataset in personalized long-term conversational tasks. Consisting of 50 dialogues, each averaging 305 turns, 20 sessions, and 9,000 tokens, it offers an advantage over existing conversational datasets (Jang et al., 2023; Xu et al., 2022) for constructing LLMs' memory. LoCoMo is built through a rigorous human-LLM co-creation pipeline: LLM-based agents with unique personas generate initial dialogues, and human annotators refine them to ensure long-term consistency. For

Model	Method	LoCoMo									
		Single Hop		Multi Hop		Temporal		Open Domain		Average	
		$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
GPT-4o-mini	LongContext	32.69	20.85	22.87	16.84	12.47	8.86	7.87	5.89	25.08	16.65
	Rag	52.45	47.94	27.50	20.13	46.07	40.35	23.23	17.94	44.67	39.33
	SeCom	16.67	13.96	15.06	11.93	14.58	12.20	17.29	14.76	15.97	13.25
	A-Mem	44.65	37.06	27.02	20.09	45.85	36.67	12.14	12.00	39.61	32.27
	Mem0	47.65	38.72	38.72	27.13	48.93	40.51	28.64	21.58	45.08	35.88
	LightMem	41.79	37.83	29.78	24.80	43.71	39.72	16.89	13.92	38.41	34.32
	CAM	49.94	46.68	36.63	25.37	18.85	12.90	22.71	18.93	39.25	33.92
	Bi-Mem	**53.68**	**47.99**	**39.17**	**31.81**	**54.44**	**41.56**	**30.99**	**26.90**	**49.74**	**42.33**
Qwen2.5-14b-Instruct	LongContext	21.00	13.17	18.09	13.21	10.54	5.82	7.31	4.71	17.41	11.10
	Rag	47.87	42.79	26.38	19.54	30.78	25.97	14.16	10.52	38.20	32.95
	SeCom	22.26	18.71	19.15	14.26	19.58	16.30	15.75	13.01	20.71	17.02
	A-Mem	33.75	30.04	22.09	15.28	27.19	22.05	13.49	10.74	28.95	24.43
	Mem0	42.58	35.15	31.73	24.82	28.96	26.24	15.03	11.28	35.99	29.88
	LightMem	34.92	31.22	25.45	19.61	32.03	27.70	15.81	11.81	31.37	27.12
	CAM	32.69	23.15	24.78	18.98	5.21	2.71	15.25	12.45	24.37	17.42
	Bi-Mem	**48.07**	**42.95**	**32.42**	**25.53**	**44.21**	**29.29**	**18.38**	**16.11**	**42.51**	**35.19**

Table 1: Overall performance of different methods on LoCoMo benchmarks. The highest score is typed in bold to indicate statistically significant improvements ($p < 0.05$), while the second-best score is underlined.

further question answering based on the historical conversations (constructed memory), LoCoMo includes 7,512 question-answer pairs spanning different types: single-hop questions derived from a single session, multi-hop questions requiring cross-session information synthesis, temporal reasoning questions assessing the grasp of time-related cues, open-domain knowledge questions integrating conversational context with external facts, and adversarial questions designed to test models' ability to recognize unanswerable queries. Note that the adversarial question category is not adopted in our work, as it does not align with the scope of memory construction and retrieval.

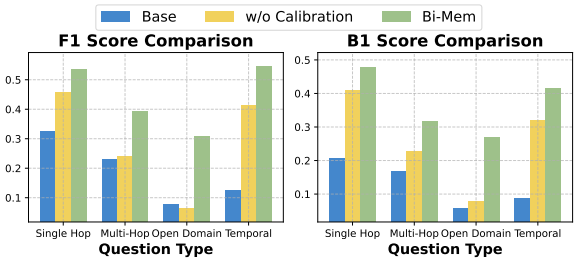

Figure 3: Ablation Study on Reflective Calibration in Hierarchical Memory Construction. "Base" denotes answer generation via long-context LLM backbones without memory, while "w/o Calibration" refers to unidirectional hierarchical memory construction without calibration in the reflective process.

Baselines We evaluate the performance of Bi-Mem against multiple leading baselines thoroughly, including: LongContext (LLM backbone without memory), RAG (Lewis et al., 2020), Mem0 (Chhikara et al., 2025), LightMem (Fang et al., 2025), A-MEM (Xu et al., 2025), SeCom (Pan et al., 2025), and CAM (Li et al., 2025). The detailed introduction for these baselines is presented in the Appendix B.1.

Implementation Details For a fair comparison, we apply the same LLM backbones and embedding models for both our Bi-Mem and all baselines. Specifically, we adopt GPT-4o-mini (Hurst et al., 2024) and Qwen2.5-14B-Instruct (Bai et al., 2025) (128K-token context window) to assess Bi-Mem's effectiveness and generalizability. For embedding models, we use all-MiniLM-L6-v2 (denoted $\phi(\cdot)$, which supports both graph edge construction (for fact-level memory) and retrieval tasks. The predefined threshold τ for edge formation between fact units (Eq. (5)) is set to 0.2. Regarding the retriever implementation, a hybrid retrieval approach is adopted, which fuses a cosine similarity-based dense retriever and a BM25 retriever with the weighting factor of 0.5. The k for the initial

Method	Single Hop		Multi-Hop		Open Domain		Temporal		Average	
	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
Fact only	49.99	44.66	27.11	20.46	9.43	7.80	51.28	36.60	43.49	36.19
Fact+Scene	50.12	44.66	32.01	23.56	25.75	22.08	52.35	37.30	45.72	37.81
Fact+Persona	49.26	43.99	27.91	20.60	9.65	9.21	51.84	37.09	43.38	36.04
Bi-Mem	**53.68**	**47.99**	**39.17**	**31.81**	**30.99**	**26.90**	**54.44**	**41.56**	**49.74**	**42.33**

Table 2: Ablation study to evaluate the respective contributions of the Fact-Scene-Persona levels in the hierarchical memory. "Fact only" denotes memory constructed solely at the fact level; "Fact+Scene" and "Fact+Persona" represent hierarchical memory combining fact-level with scene-level or persona-level information.

Method	Single Hop		Multi-Hop		Open Domain		Temporal		Average	
	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
Top-Down	13.87	11.85	12.92	9.51	14.47	13.43	16.63	11.58	14.32	11.46
Botom-Up	42.59	37.50	29.01	20.95	21.26	17.03	47.05	33.26	39.68	32.27
Hierarchical	51.26	45.76	32.28	24.58	20.17	17.46	51.42	36.16	45.84	38.06
+Scene_to_Fact	52.97	47.31	34.02	26.81	22.72	19.38	53.48	38.05	47.68	39.83
+Fact_to_Scene	52.40	46.75	34.16	26.96	25.14	20.51	51.62	36.44	47.16	39.29
Bi-Mem	**53.68**	**47.99**	**39.17**	**31.81**	**30.99**	**26.90**	**54.44**	**41.56**	**49.74**	**42.33**

Table 3: Ablation study on different retrieval strategies. "Top-down" and "Bottom-up" denote retrieving relevant memories from the top or bottom level, followed by level-by-level association. "Hierarchical" refers to retrieving the top-k relevant memories from all three levels simultaneously. "+ Scene_to_Fact" and "+ Fact_to_Scene" extend hierarchical retrieval by spreading from scenes to their child facts or from facts to their parent scenes.

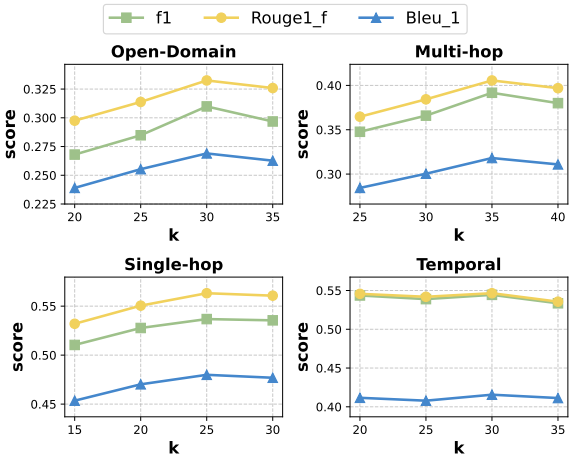

Figure 4: Sensitivity analysis of Bi-Mem to the hyperparameter k, which is the number of initially retrieved memory units in the hierarchical search stage.

search in Eq (9) is selected in $[15, 20, 30, 35, 40]$, while m for associative spreading in Eq (10) ranges from $[1, 2, 3]$. The optimal hyperparameter settings are presented in the Appendix B.2. To quantify the memory's effectiveness, we evaluate all methods by comparing their QA accuracy using two metrics F_1 (balancing answer precision and recall) and BLEU-1 (denoted B_1, which measures word overlap between generated and ground-truth answers).

5.2 Main Results (RQ1)

To answer Q1, we compare Bi-Mem with leading baselines on four types of questions from the LoCoMo benchmark. We conducted each experiment at least three times on different random seeds, with the averaged results reported in Table 1. We can observe that Bi-Mem consistently outperforms all baselines on both GPT-4o-mini and Qwen-2.5-14B-Instruct backbones, validating the effectiveness and generalizability of our bidirectional memory construction and associative retrieval for LLMs' personalization. Notably, RAG (Lewis et al., 2020) serves as flat factual memory in our setup, only outperforming on single-hop questions. This demonstrates that purely factual memory fails to handle complex questions requiring inter-fact relationship capture or high-level user behavioral pattern recognition, highlighting the necessity of a multi-granular memory structure.

5.3 Ablation for Memory Construction (RQ2)

To address Q2, we conduct experiments to verify the necessity of each granular memory in the

hierarchical structure. As shown in Table 2, fact-only memory construction exhibits the worst performance across all question types except single-hop, highlighting the importance of high-level memory. Benefiting from persona-level calibration that provides global constraints, "Fact+Scene" achieves comparable performance. Furthermore, Bi-Mem (integrating all three levels) attains the best performance, confirming the necessity of each granularity. To further evaluate the effectiveness of bidirectional construction, we conduct an ablation study focused on reflective process calibration. As illustrated in Figure 3, removing the reflective process (w/o Calibration) results in significant performance degradation across all question types. Additionally, we present a case in the Appendix C.2 to validate the significance of reflective calibration vividly.

5.4 Ablation for Associative Retrieval (RQ3)

To evaluate the effectiveness of our Associative Retrieval mechanism, we compare it with various retrieval strategies as presented in Table 3. The results indicate that unidirectional strategies exhibit clear limitations in bridging multi-granular memories. "Top-down" strategy may fail to retrieve the precise details grounded in the questions, and "Bottom-Up" strategy may frequently suffer from semantic fragmentation. The "Hierarchical" baseline, which retrieves the top-k candidates from all levels simultaneously, improves performance but remains constrained by the independence of each retrieved memory unit. The significant performance gains observed with "+ Fact_to_Scene" and "+ Scene_to_Fact" demonstrate the power of spreading activation after initial hierarchical search. Ultimately, the full Bi-Mem retrieval configuration achieves the best results, proving that bidirectional association effectively bridges the gap between abstract persona and concrete conversational facts.

5.5 Sensitivity Analysis (RQ4)

To address RQ4, we analyze Bi-Mem's sensitivity to the hyperparameter k (number of initially retrieved memory units in the hierarchical search stage). k is varied over $[15, 20, 25, 30, 35, 40]$. As shown in Figure 4, answer accuracy across all tasks initially increases with k — indicating that a larger initial retrieval budget enables our associative retrieval to capture a more comprehensive memory set. Performance then degrades after peaking, as excessive retrieved memory introduces additional noise that negatively impacts relevant information extraction. This underscores that an optimal k effectively bridges gaps between fragmented facts and thematic scenes. More sensitivity analysis on other hyperparameters (*e.g.,* m for associative spreading in Eq. 10) is listed in the Appendix C.1.

Metric	A-mem	CAM	Bi-Mem
Memory Construction (s)	3327	443	3448
Retrieving-Answering (s)	7.30	8.86	4.54
TOTAL (s)	4437	1790	4138
ANSWER QUALITY (F1)	39.65	39.25	49.73

Table 4: Cost-Efficiency Analysis: We measured the memory construction time and average retrieving-answering time per question for a single user, with total time encompassing memory construction and answering the subsequent 152 questions.

5.6 Efficiency (RQ5)

To evaluate efficiency, we compare Bi-Mem's memory construction and QA time costs with memory-based baselines (A-Mem (Xu et al., 2025), CAM (Li et al., 2025)). As shown in Table 4, Bi-Mem has superior QA efficiency (4.54s/question), significantly faster than A-Mem (7.30s) and CAM (8.86s). This speedup stems from our Associative Retrieval mechanism—pre-established hierarchical links enable faster multi-granular evidence identification than flat retrieval or complex re-ranking. Though its bidirectional Inductive-Reflective process increases construction time (3,448s) vs. CAM (443s), it matches A-Mem's efficiency (3,327s). Notably, this trade-off is justified by large accuracy gains, confirming the value of bidirectional hierarchical memory construction.

6 Conclusion

In this paper, we present Bi-Mem, a framework designed to ensure hierarchical memory fidelity via bidirectional construction and an associative retrieval mechanism. Instead of constructing hierarchical memory through undirectional aggregation, Bi-Mem employs an inductive agent for bottom-up formulation and a reflective agent for top-down calibration, mitigating the misalignment between local memory and the user's global persona. To ensure coherent recall, the associative retrieval mechanism connects memories of different granularities via spreading activation. Experimental results demonstrate that Bi-Mem significantly enhances QA accuracy in long-term personalized conversational tasks.

Limitations

While Bi-Mem demonstrates significant improvements in the memory system for personalized LLMs, there are several limitations:

1. Sensitivity to Model Reasoning: The fidelity of bidirectional memory construction is inherently tied to the instruction-following and reasoning capabilities of the underlying LLM. Variations in LLMs' instructions may impact the effectiveness of the inductive and reflective processes, suggesting a need for more robust prompting strategies.

2. Static vs. Dynamic Persona Modeling: While our framework ensures consistency with a stable persona anchor, it is primarily designed for users with relatively stable personas. Future work could extend the calibration mechanism to better capture the temporal evolution of user preferences and dynamic persona shifts.

3. Potential for Optimization via RL: Our current framework focuses on the structural design of memory construction and retrieval. While effective, there is an opportunity to further refine agent-specific behaviors, such as the reflective agent's calibration policy, through Reinforcement Learning (RL) or preference optimization to better align with diverse user interaction styles.

Ethics Statement.

This work is designed to enhance the fidelity of hierarchical memory systems, formulating the memory construction process as a bidirectional alignment between local scenes and global personas. The generative AI is used for coding and writing assistance. We do not foresee any direct, immediate, or negative societal impacts of our research.

Reproducibility Statement.

All the results in this work are reproducible. We have discussed the optimal hyperparameters and the details on devices and software environments in Appendix B.2.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

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

Jin Chen, Zheng Liu, Xu Huang, Chenwang Wu, Qi Liu, Gangwei Jiang, Yuanhao Pu, Yuxuan Lei, Xiaolong Chen, Xingmei Wang, Kai Zheng, Defu Lian, and Enhong Chen. 2024. When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web (WWW)*, 27(4):42.

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.

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

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. In *NeurIPS*.

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

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.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *EMNLP*, pages 13584–13606. Association for Computational Linguistics.

Jiho Kim, Yeonsu Kwon, Yohan Jo, and Edward Choi. 2023. KG-GPT: A general framework for reasoning on knowledge graphs using large language models. In *EMNLP (Findings)*, pages 9410–9421. Association for Computational Linguistics.

Andrea Lancichinetti and Santo Fortunato. 2012. Consensus clustering in complex networks. *Scientific reports*, 2(1):336.

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 *NeurIPS*.

Rui Li, Zeyu Zhang, Xiaohe Bo, Zihang Tian, Xu Chen, Quanyu Dai, Zhenhua Dong, and Ruiming Tang. 2025. CAM: A constructivist view of agentic memory for LLM-based reading comprehension. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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 *ACL (1)*, pages 13851–13870. Association for Computational Linguistics.

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*.

Anup Pattnaik, Cijo George, Rishabh Kumar Tripathi, Sasanka Vutla, and Jithendra Vepa. 2024. Improving hierarchical text clustering with llm-guided multi-view cluster representation. In *EMNLP (Industry Track)*, pages 719–727. Association for Computational Linguistics.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. 2025. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. In *ICLR*. OpenReview.net.

Zhaoxuan Tan, Zheyuan Liu, and Meng Jiang. 2024a. Personalized pieces: Efficient personalized large language models through collaborative efforts. In *EMNLP*, pages 6459–6475. Association for Computational Linguistics.

Zhaoxuan Tan, Qingkai Zeng, Yijun Tian, Zheyuan Liu, Bing Yin, and Meng Jiang. 2024b. Democratizing large language models via personalized parameter-efficient fine-tuning. In *EMNLP*, pages 6476–6491. Association for Computational Linguistics.

Johan Ugander and Lars Backstrom. 2013. Balanced label propagation for partitioning massive graphs. In *WSDM*, pages 507–516. ACM.

Chengbing Wang, Yang Zhang, Fengbin Zhu, Jizhi Zhang, Tianhao Shi, and Fuli Feng. 2025a. Leveraging memory retrieval to enhance llm-based generative recommendation. In *WWW (Companion Volume)*, pages 1346–1350. ACM.

Zehong Wang, Junlin Wu, Zhaoxuan Tan, Bolian Li, Xianrui Zhong, Zheli Liu, and Qingkai Zeng. 2025b. From personal to collective: On the role of local and global memory in LLM personalization. *CoRR*, abs/2509.23767.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *ACL (1)*, pages 5180–5197. Association for Computational Linguistics.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for LLM agents. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

BY Yan, Chaofan Li, Hongjin Qian, Shuqi Lu, and Zheng Liu. 2025. General agentic memory via deep research. *arXiv preprint arXiv:2511.18423*.

Wan Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics*, 13(5):856.

Yujie Zhang, Weikang Yuan, and Zhuoren Jiang. 2025a. Bridging intuitive associations and deliberate recall: Empowering LLM personal assistant with graph-structured long-term memory. In *ACL (Findings)*, pages 17533–17547. Association for Computational Linguistics.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2025b. Personalization of large language models: A survey. *Trans. Mach. Learn. Res.*, 2025.

Xinping Zhao, Yan Zhong, Zetian Sun, Xinshuo Hu, Zhenyu Liu, Dongfang Li, Baotian Hu, and Min Zhang. 2025. Funnelrag: A coarse-to-fine progressive retrieval paradigm for RAG. In *NAACL (Findings)*, pages 3029–3046. Association for Computational Linguistics.

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

Algorithm 1 Bidirectional Hierarchical Memory Construction

Input: Conversation history $\mathcal{C}$, similarity threshold τ
Output: Calibrated hierarchical memory $\mathcal{M} = (\mathcal{F}, \mathcal{S}', \mathcal{P})$

// Phase 1: Inductive Process (Bottom-Up)
$\mathcal{F} \leftarrow$ ExtractFacts($\mathcal{C}$) via operator $\mathcal{E}$
$G \leftarrow$ ConstructFactGraph($\mathcal{F}, \tau$)
$\{\bar{\mathcal{F}}_j\}_{j=1}^{J} \leftarrow$ LPA(G) ▷ Graph-based clustering
$\mathcal{S} \leftarrow$ AggregateScenes($\{\bar{\mathcal{F}}_j\}$) via operator $\mathcal{T}$
$\mathcal{P} \leftarrow$ DistillPersona($\mathcal{S}$) via operator $\mathcal{D}$ ▷ 5-dim persona profile
// Phase 2: Reflective Process (Top-Down)
$\mathcal{S}' \leftarrow \emptyset$
for each scene $s_j \in \mathcal{S}$ **do**
  if s_j misalign with $\mathcal{P}$ **then**
    $\Delta s_j \leftarrow$ Calibrate($s_j, \mathcal{P}$) via operator $\mathcal{R}$
    $s'_j \leftarrow (s_j, \Delta s_j)$ ▷ Condition-based calibration
  else
    $s'_j \leftarrow (s_j, \emptyset)$
  end if
  $\mathcal{S}' \leftarrow \mathcal{S}' \cup \{s'_j\}$
end for
return $\mathcal{M} = \langle \mathcal{F}, \mathcal{S}', \mathcal{P} \rangle$

Algorithm 2 Associative Retrieval Mechanism

Input: Hierarchical memory $\mathcal{M} = \langle \mathcal{F}, \mathcal{S}, \mathcal{P} \rangle$, Query q^*, parameters k, m
Output: Final retrieved memory set $\mathcal{M}_{ret}$

// Phase 1: Initial Hierarchical Search
Compute $a_0(x) = \text{sim}(\phi(q^*), \phi(x))$ for all $x \in \mathcal{F} \cup \mathcal{S} \cup \mathcal{P}$
$\mathcal{M}_{ret} \leftarrow$ Top-$k(\{x \mid a_0(x)\})$ ▷ Select initial seeds across layers
// Phase 2: Associative Spreading
for each element $u \in \mathcal{M}_{ret}$ **do**
  if $u \in \mathcal{F}$ **then**
    $s \leftarrow$ FindParent(u) ▷ Bottom-up: Associate parent scene
    $\mathcal{M}_{ret} \leftarrow \mathcal{M}_{ret} \cup \{s\}$
  else if $u \in \mathcal{S}$ **then**
    $\bar{\mathcal{F}}_u \leftarrow$ FactCluster(u) ▷ Top-down: Associate constituent facts
    $\mathcal{F}_{top\_m} \leftarrow$ Top-$m(\{f \in \bar{\mathcal{F}}_u \mid \text{sim}(u, f)\})$
    $\mathcal{M}_{ret} \leftarrow \mathcal{M}_{ret} \cup \mathcal{F}_{top\_m}$
  end if
end for
return $\mathcal{M}_{ret}$

A Algorithm

Here we list the algorithm of Bi-Mem's bidirectional hierarchical memory construction process in Algorithm 1, and associative retrieval mechanism in Algorithm 2.

B Detailed Experimental Settings

B.1 Details for Baselines

Here, we introduce the baseline methods in detail.

- LongContext: it directly uses LLMs (no memory mechanisms) for QA tasks, integrating the historical conversations and queries into the prompt.
- RAG (Lewis et al., 2020), the conversations are split into 2,048-token chunks, with the top-5 retrieved chunks used for QA tasks, following the settings of GAM (Yan et al., 2025).
- A-MEM (Xu et al., 2025) is an agentic framework with structured memory, dynamic links, and semantic matching-driven evolution.
- LightMem (Fang et al., 2025) is a lightweight system with a three-stage memory (sensory/short-term/long-term).
- Mem0 (Chhikara et al., 2025) extracts memory with ADD/UPDATE/DELETE/NOOP operations and retrieves via vector similarity.
- SeCom (Pan et al., 2025) builds denoised memory from topically coherent chunks and retrieves at the segment level.
- CAM (Li et al., 2025) is a hierarchical memory framework that adopts an incremental overlapping clustering algorithm for construction and employs a Prune-and-Grow strategy for retrieval.

B.2 More Implementation Details

To implement the ablation variant "Top-Down" and "Bottom-Up" in Section 5.4, we set the number of selected relevant memory units to 1, 15, and 25 for the fact, scene, and persona levels, respectively, in line with mainstream configurations. Total tokens consumed to construct the hierarchical memory per user are approximately 200,000–300,000, while those for answering subsequent questions per user range from 100,000 to 20,000. The embedding model "all-MiniLM-L6-v2" is deployed on an A800, supporting both edge construction in Eq. (5) and the retrieval process in Eqs. (9), (10). Additionally, the detailed settings for our hyperparameters are listed in Table 5. For reproduction, the

	Single Hop	Multi Hop	Temporal	Open Domain
k	35	25	30	25
m	3	3	3	3
τ	0.2	0.2	0.2	0.2
α	0.5	0.5	0.5	0.5
L_w	128k	128k	128k	128k

Table 5: Hyperparameter settings on different question types. α denotes the weighting factor between the dense retriever and the BM25 retriever when retrieving memories. L_w represents the context window for LLM backbones.

data and code is available at https://anonymous.4open.science/r/Bi-Mem-622D/.

C More Experimental Results

C.1 More Sensitivity Analysis

To validate Bi-Mem's sensitivity to the hyperparameter m (representing the number of activated facts from retrieved scenes during associative spreading activation), we present the experimental results in Table 6. We observe that performance peaks at $m = 3$, achieving an average F_1 of 49.74%. while both $m = 1$ and $m = 5$ lead to marginal degradation. Specifically, $m = 1$ (a small value) fails to provide sufficient factual memory for complex tasks, whereas increasing m to 5 may introduce redundant noise to distract the LLM from core user intent. This demonstrates that a balanced association scale can effectively retrieve optimal relevant memories for accurate question answering.

C.2 Case Study

To vividly illustrate the significance of bidirectional construction for hierarchical memory, we present a representative case study focusing on inductive memory construction and Reflective Calibration. As depicted in Figure 5 (original conversations) and Figure 6 (fact-level memory content), the factual information extracted from raw dialogues may contain noise (*e.g.,* "Taking care of ourselves is not always easy") and hallucinations (*e.g.,* "I am seeking help"). Such inaccuracies can distort the aggregated scene-level memory (Figure 7), leading to misalignment with the user's global persona-level memory—where Caroline is defined by empathy and courage to embrace her authentic self (Figure 8). Through Reflective Calibration, we inject the persona as a global constraint into the scene-level memory, ultimately generating a calibrated scene-level memory (Figure 9) that integrates inherent persona traits (*e.g.,* empathy and courage to embrace her authentic self) and ensures strict alignment between the scene-level memory and the core persona.

D Prompt Template

Below, we present the prompts that are designed for each corresponding step of the bidirectional hierarchical memory construction.

Fact-level memory extraction $\mathcal{E}$

Generate a structured fact con_i for the following interaction content c_i by:
1. Identifying the most salient keywords (focus on nouns, verbs, and key concepts)
2. Extracting core themes and contextual elements
3. Creating relevant categorical tags
Format the response as a JSON object:
{
- keywords: ["keyword1", "keyword2", ...],
- context: "one sentence summarizing the interaction content",
- tags: ["tag1", "tag2", ...]
}

Scene-level memory aggregation $\mathcal{A}$

You are a scene synthesizer specialized in factual comprehension..
Task: Summarize a cluster of related factual memories into a coherent 'Scene Memory'.
Input factual memories: facts_content
Instructions:
1. Identify the core theme connecting these facts.
2. Generate a descriptive summary capturing the progression of conversational facts.
3. Extract key entities and topics.
Format the response as a JSON object:
{
- scene_memory: "A comprehensive summarized scene",
- keywords: ["keyword1", "keyword2", ...],
- tags: ["tag1", "tag2", ...]
}

Method	Single Hop		Multi-Hop		Open Domain		Temporal		Average	
	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$	$F_1 \uparrow$	$B_1 \uparrow$
m=1	52.79	46.95	34.56	27.83	25.22	23.01	51.12	**44.77**	47.12	41.25
m=5	53.58	47.82	37.52	29.98	29.91	26.03	53.24	41.03	49.06	41.73
Bi-Mem$(m = 3)$	**53.68**	**47.99**	**39.17**	**31.81**	**30.99**	**26.90**	**54.44**	41.56	**49.74**	**42.33**

Table 6: Sensitivity experiment on the hyperparameter m, which represents the number of extended facts in Scene_to_Fact when conducting associative spreading activation.

Persona-Level Memory Distillation $\mathcal{D}$

You are a persona synthesizer specialized in psychological and behavioral analysis.
Task: Create a COMPREHENSIVE User persona based on the provided scene memories.
Input Scenes: all_scenes_content
Instructions:
1. Analyze these scenes deeply. Look for patterns in behavior, emotion, and choices.
2. For each dimension below, write a DETAILED paragraph (5-10 sentences). Do not be brief.
3. Use specific examples from the scenes to support your analysis.
Format the response as a JSON object:
{
- basic_info: "Detailed background...",
- interests: "Comprehensive list of hobbies and how they engage with them...",
- personality: "In-depth personality analysis...",
- values: "Core beliefs and motivations...",
- relationships: "Detailed social dynamics..."
}

Reflective Calibration.

You are a scene memory calibrator. Your goal is to align the given scene to the user's persona.
Persona-level memory: user_persona
Scene-level memory: current_scene
Instructions:
1. Read the persona-level memory to understand the user's key interests, values, and traits.
2. Check the current scene-level memory. Does it fail to mention any specific connection to the user persona that is likely present in the scene?
3. If yes, add a compensatory condition to append to the original scene. This addition should explicitly align the scene to the persona (e.g., "This aligns with her interest in ...").
4. CRITICAL: DO NOT REWRITE the existing summary. ONLY generate text to ADD.
5. If the current summary is already perfect, return an empty string for "added condition".
Format the response as a JSON object:
{
- needs_calibration: true/false,
- added condition: "Text to add (or empty string) as a condition",
- reason: The reason why you decide to calibrate."
}

Original converstaions:

1. I totally agree, Melanie. Taking care of ourselves is so important - even if it's not always easy. Great that you're prioritizing self-care.
2. I shared my own journey, the struggles I had and how much I've developed since coming out.
3. I struggled with mental health, and support I got was really helpful.
4. Lately, I've been looking into counseling and mental health as a career. I want to help people who have gone through the same things as me.

Figure 5: Case study in a cluster of original conversations.

Fact-level memory content:

1. Taking care of ourselves is not always easy. **(noise)**
2. I shared my own journey, the struggles I had and how hard to come out.
3. I struggled with mental health, and I am seeking help. **(hallucinations)**
4. I'm considering a counseling/mental health career to help those with similar experiences.

Figure 6: Case study on extracted fact-level memory content for the conversational cluster, where noise and hallucination exist.

Aggregated scene-level memory:

Due to the challenges of self-care and mental health struggles, she often shares her personal journey, seeks support during this process, and aspires to a career in the mental health/counseling field to assist others with similar experiences.

Figure 7: Case study on the aggregated scene-level memory from the fact cluster, which is misaligned with the global persona.

Related Persona-level memory:

Caroline is characterized by her empathy, creativity, and resilience, coupled with the courage to embrace her authentic self. She is distinguished by her warm and approachable demeanor, her steadfast loyalty to her family, and her exceptionally supportive nature.

Figure 8: Case study on the global persona-level memory related to the scene.

Calibrated scene-level memory:

Due to the challenges of self-care and mental health struggles, she often shares her personal journey, seeks support during this process, and aspires to a career in the mental health/counseling field to assist others with similar experiences. **Actually, she is inherently identified by empathy, supportive nature, and courage to embrace her authentic self.**

Figure 9: Case study on the calibrated scene-level memory, which is injected with the persona constraint.