HiMem: Hierarchical Long-Term Memory for LLM Long-Horizon Agents

Anonymous ACL submission

Abstract

Although long-term memory systems have made substantial progress in recent years, they still exhibit clear limitations in adaptability, scalability, and self-evolution under continuous interaction settings. Inspired by cognitive theories, we propose HiMem, a hierarchical long-term memory framework for long-horizon dialogues, designed to support memory construction, retrieval, and dynamic updating during sustained interactions. HiMem constructs cognitively consistent Episode Memory via a Topic-Aware Event-Surprise Dual-Channel Segmentation strategy, and builds Note Memory that captures stable knowledge through a multi-stage information extraction pipeline. These two memory types are semantically linked to form a hierarchical structure that bridges concrete interaction events and abstract knowledge, enabling efficient retrieval without sacrificing information fidelity. HiMem supports both hybrid and best-effort retrieval strategies to balance accuracy and efficiency, and incorporates conflict-aware Memory Reconsolidation to revise and supplement stored knowledge based on retrieval feedback. This design enables continual memory self-evolution over long-term use. Experimental results on long-horizon dialogue benchmarks demonstrate that HiMem consistently outperforms representative baselines in accuracy, consistency, and long-term reasoning, while maintaining favourable efficiency. Overall, HiMem provides a principled and scalable design paradigm for building adaptive and self-evolving LLM-based conversational agents.

1 Introduction

Large language models (LLMs) have demonstrated remarkable progress in language understanding and reasoning, enabling the development of LLM-based agents for complex, multi-turn tasks such as personalized assistance, planning, and long-term decision support (Yu et al., 2025; Su et al., 2025; Liu et al., 2025). In realistic interactive settings, however, these agents are required to operate over extended time horizons, where relevant information is scattered across long dialogues and multiple sessions. Despite strong short-term reasoning ability, existing LLM agents still struggle to reliably preserve, organize, and utilize information over long time spans. This limitation has emerged as a fundamental bottleneck for building adaptive and consistent long-horizon conversational agents (Xu et al., 2022; Wu et al., 2025a; Li et al., 2025).

Recent efforts to address this challenge can be broadly categorized into three directions. Retrieval-augmented generation (RAG) systems introduce external memory stores to fetch relevant information on demand, improving factual grounding (Alonso et al., 2024; Jiménez Gutiérrez et al., 2024; Sarthi et al., 2024; Edge et al., 2024). Long-context modeling approaches extend the context window to thousands or even millions of tokens, enabling direct reasoning over extended histories (Du et al., 2025; Fountas et al., 2025; Qian et al., 2025; Lee et al., 2024). More recently, structured long-term memory systems have been proposed to persistently store and retrieve dialogue information in compressed or structured forms (Hatalis et al., 2023; Nan et al., 2025). While these methods significantly improve efficiency and continuity, they still exhibit systematic limitations when applied to long-horizon interactions (Yue et al., 2024; Zhang et al., 2025c).

From both empirical observations and cognitive perspectives, we identify three recurring challenges that existing long-term memory systems struggle to address simultaneously. First, **semantic misalignment** arises when extracted memories are detached from their original dialogue context, leading to errors in resolving temporal references, coreference, and implicit semantics. Second, most systems rely on **monolithic or insufficiently hierarchical memory structures**, forcing a trade-off be-

tween information fidelity and retrieval efficiency. Fine-grained dialogue logs preserve rich context but incur high retrieval costs, whereas aggressively abstracted representations reduce cost at the expense of critical details needed for reasoning and personalization. Third, memory updates are typically **static or similarity-driven**, lacking principled mechanisms to revise or correct stored knowledge when new information partially overlaps with, extends, or contradicts existing memories. As a result, long-term consistency degrades over sustained interactions.

Inspired by cognitive theories of human memory (Gilboa and Marlatte, 2017; Ghosh and Gilboa, 2014; Fauconnier and Turner, 2003; Nader, 2015), we argue that effective long-term memory for LLM agents must satisfy three properties: (i) a *hierarchical structure* that bridges concrete interaction events and abstracted knowledge, (ii) a *unified semantic alignment mechanism* that preserves interpretability across memory representations, and (iii) a *conflict-aware update process* that supports continual self-evolution rather than static accumulation. Based on these principles, we propose **HiMem**, a hierarchical long-term memory framework designed for long-horizon conversational agents.

HiMem organizes memory into two semantically linked layers. *Episode Memory* preserves fine-grained, temporally grounded interaction segments constructed via a Topic-Aware Event–Surprise Dual-Channel Segmentation strategy, which aligns memory boundaries with both topical shifts and cognitively salient discontinuities. *Note Memory* abstracts stable knowledge such as facts, user preferences, and user profiles through a multi-stage information extraction pipeline. These two memory types form a hierarchical transition from concrete events to compact knowledge representations, enabling efficient retrieval without sacrificing information fidelity. During retrieval, HiMem supports both a hybrid retrieval strategy and a best-effort retrieval strategy that descends from abstract knowledge to concrete events only when necessary. Crucially, retrieval failures are treated as learning signals: HiMem performs conflict-aware Memory Reconsolidation to supplement missing knowledge and revise existing memories, enabling continuous self-evolution over long-term use.

We evaluate HiMem on long-horizon dialogue benchmarks and demonstrate that it consistently outperforms representative baselines in accuracy, consistency, and efficiency. Extensive ablation studies further validate the necessity of hierarchical memory organization, semantic alignment, and conflict-aware updating for robust long-term reasoning.

In summary, this paper makes the following contributions:

- We propose **HiMem**, a hierarchical long-term memory framework that integrates episodic and knowledge-oriented memories to support scalable and adaptive long-horizon conversational agents.
- We introduce a Topic-Aware Event–Surprise Dual-Channel Segmentation mechanism and a multi-stage information extraction pipeline to construct cognitively consistent and efficient memory representations.
- We design a conflict-aware Memory Reconsolidation mechanism that enables long-term memory systems to self-correct and evolve during sustained interactions.
- We provide extensive experimental evidence showing that principled hierarchical design and dynamic updating substantially improve long-horizon reasoning performance.

2 Methodology

HiMem is a modular long-term memory framework built upon a hierarchical architecture that integrates episodic interaction records with abstracted knowledge representations. It is designed to support efficient retrieval, semantic consistency, and continual memory evolution during long-horizon interactions.

2.1 Overall Framework

As shown in Figure 1, HiMem consists of three core modules: (i) a hierarchical memory construction module that builds Episode Memory and Note Memory from raw dialogues, (ii) a hierarchical memory retrieval module that supports both hybrid and best-effort retrieval strategies, and (iii) a conflict-aware memory updating module that enables continual self-evolution. Episode Memory preserves fine-grained interaction events, while Note Memory consolidates stable knowledge such as facts, user preferences, and user profiles. The two memory layers are semantically linked to form a hierarchy that bridges concrete experiences and abstract knowledge.

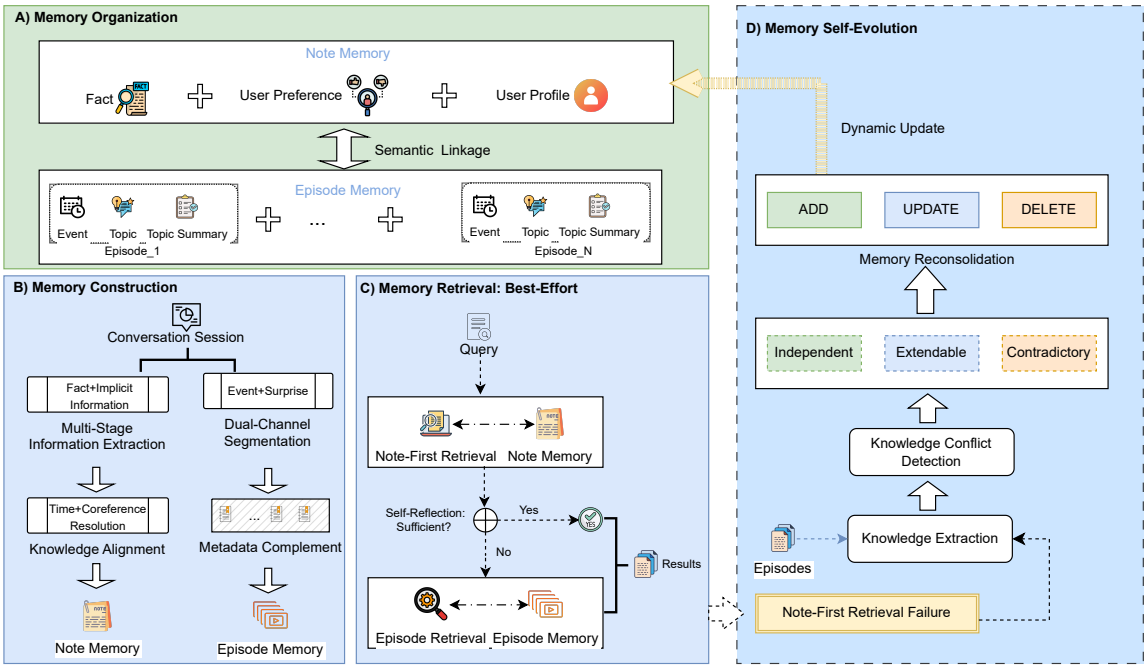

Figure 1: **Overview of HiMem.** (A) *Memory organization:* a hierarchical connection between Episode Memory and Note Memory. (B) *Memory construction:* pipelines that transform dialogue logs into Episode Memory and Note Memory. (C) *Best-effort retrieval:* hierarchical retrieval in the order of **Note Memory** → **Episode Memory**, with an LLM assessing evidence sufficiency. (D) *Memory self-evolution:* when evidence from Note Memory is insufficient, the system supplements potentially missing information from Episode Memory and triggers conflict detection and updating.

2.2 Memory Construction

Memory construction in HiMem follows a multi-stage pipeline that progressively transforms raw dialogue logs into structured long-term memory representations. This pipeline unifies event-level segmentation, knowledge extraction, and semantic alignment to ensure both fidelity and efficiency.

2.2.1 Episode Memory

Episode Memory records fine-grained interaction events aligned with topical and cognitive boundaries. Given a dialogue sequence, HiMem segments it into a sequence of non-overlapping episodes. Each episode is represented by a structured record containing an ID, timestamp, topic, topic summary, metadata, and the corresponding dialogue segment.

Dual-Channel Segmentation. To obtain cognitively coherent episodes, HiMem adopts a Topic-Aware Event–Surprise Dual-Channel Segmentation strategy. A segmentation boundary is introduced when either (i) a topical shift occurs in discourse goals or subtopics, or (ii) a salient discontinuity is detected, such as an abrupt change in intent or emotional state. These two criteria are fused using an OR rule, producing event units that align with both semantic continuity and cognitive salience.

Segmentation is performed in a single pass, where an LLM jointly evaluates topical and surprise signals and directly outputs the final segmentation. This design yields compact and self-contained episodes that reduce cross-segment interference while preserving critical contextual evidence for downstream reasoning.

2.2.2 Note Memory

Note Memory focuses on long-term storage of knowledge-oriented information that remains stable or reusable across interactions. From each dialogue, HiMem extracts three categories of knowledge:

$$K = \{K_{\text{fact}}, K_{\text{pref}}, K_{\text{profile}}\},$$

where K_{fact} denotes objective facts and events, K_{pref} captures user preferences, and K_{profile} represents relatively stable user traits.

Multi-Stage Knowledge Extraction. Knowledge extraction is decomposed into three stages to avoid semantic collapse. Stage 1 extracts in-

dependently interpretable factual and situational units. Stage 2 identifies high-confidence implicit information related to user preferences and profiles without introducing new facts. Stage 3 performs non-destructive normalization, including deduplication, coreference resolution, and temporal normalization, producing aligned knowledge representations suitable for long-term storage. Each aligned knowledge entry is stored as a note, represented as a structured record containing an identifier, the extracted content, a semantic category, and associated metadata.

2.2.3 Knowledge Alignment

To maintain semantic consistency across memory layers, HiMem applies a unified alignment process during memory construction. This process includes temporal alignment of relative time expressions, coreference resolution for entity mentions, and extraction of implicit semantic relations. Alignment is selectively applied: Episode Memory prioritizes preserving original dialogue context, while Note Memory emphasizes abstraction and normalization.

2.3 Memory Retrieval

HiMem supports two complementary retrieval strategies. In *hybrid retrieval*, the system retrieves information from both Episode Memory and Note Memory to maximize recall. In contrast, *best-effort retrieval* proceeds hierarchically by querying Note Memory first and falling back to Episode Memory only when evidence is deemed insufficient. Retrieved evidence is evaluated by an LLM to assess answerability, and unsupported queries are explicitly marked as unanswerable.

2.4 Memory Updating and Self-Evolution

During best-effort retrieval, HiMem employs a fixed LLM-based self-evaluation prompt to assess whether the retrieved evidence is sufficient to answer the query. This evaluation produces a binary judgment (*sufficient* or *insufficient*) under deterministic decoding (temperature = 0) and serves solely as a control signal, without introducing or revising memory content. While related to iterative self-refinement approaches (Madaan et al., 2023), HiMem confines the LLM to deterministic routing and decision control.

Memory reconsolidation is triggered only when two conditions are jointly satisfied: (i) retrieval from Note Memory alone is insufficient, and (ii) the subsequently retrieved Episode Memory provides adequate supporting evidence. This conjunctive trigger grounds updates in episodic context and prevents premature revisions. Although conceptually related to reflective agent frameworks such as (Shinn et al., 2023), HiMem performs structured, evidence-grounded memory operations rather than free-form verbal reflection.

When reconsolidation is triggered, HiMem conducts query-conditioned knowledge extraction over the supporting episodes and compares the extracted information with existing notes. Their relationship is classified as *independent*, *extendable*, or *contradictory*, based on which the system applies ADD, UPDATE, or DELETE operations to revise Note Memory. This typed design avoids indiscriminate overwriting and echoes classic belief revision perspectives (Gärdenfors, 1988), promoting long-term stability and semantic consistency.

In contrast, episodic memory is treated as immutable: newly constructed episodes are appended chronologically without modification, preserving the temporal integrity of interaction histories.

2.5 Adaptive Forgetting

To regulate memory growth under sustained interactions, HiMem optionally employs an adaptive forgetting mechanism based on usage frequency. In this work, forgetting primarily serves as a scalability-oriented mechanism to control memory size and maintain retrieval efficiency, and does not contribute directly to the performance gains reported in our experiments.

3 Experiments

3.1 Datasets

We evaluate HiMem on **LoCoMo** (Maharana et al., 2024), a benchmark designed to assess long-horizon conversational reasoning under sustained interactions. LoCoMo consists of multi-session dialogues with an average length of approximately 600 turns (around 16K tokens) and spans up to 32 interaction stages, posing significant challenges for long-range dependency modeling and memory management.

The benchmark covers diverse reasoning scenarios, including *Single-Hop* questions answerable within a single session, *Multi-Hop* questions requiring aggregation across distant dialogue turns, *Temporal Reasoning* questions involving implicit or explicit time relations, and *Open-Domain* ques-

tions that combine dialogue content with external or commonsense knowledge. Following prior work (Chhikara et al., 2025), we exclude the Adversarial category from quantitative evaluation, as it focuses on unanswerability detection rather than answer correctness.

3.2 Evaluation Metrics

Since different long-term memory systems may apply varying degrees of compression or abstraction over dialogue histories, we adopt a multi-dimensional evaluation protocol to assess answer quality comprehensively. Specifically, following prior work that systematically studies LLM-as-a-Judge and its biases (Zheng et al., 2023; Pan et al., 2025; Xu et al., 2025b), we use GPT-4o-mini as the LLM judge to compute evaluation score (denoted as GPT-Score) as the primary metric to approximate semantic correctness and consistency, together with F1 to measure lexical overlap.

In addition, for efficiency evaluation, we report **latency** (Lat.) and **token consumption** (Tok.). Latency is measured as the time required for memory retrieval only, excluding LLM inference and response generation, in order to isolate the efficiency of the memory system.

3.3 Baselines

We compare HiMem with representative long-term memory frameworks that cover different design paradigms. **Mem0** (Chhikara et al., 2025) represents structured memory systems based on atomic factual extraction and graph-based organization. **SeCom** (Pan et al., 2025) adopts event-level semantic segmentation and compression to improve contextual completeness. **A-MEM** (Xu et al., 2025b) augments event-level memory with entities, relations, and temporal features to support time-aware retrieval and reasoning. These baselines enable a systematic comparison across retrieval-based, compressed-context, and structured memory approaches.

These baselines are selected based on their compatibility with long-horizon conversational memory, availability of reproducible implementations, and suitability for evaluation under a unified agent interface with comparable inference budgets; methods that primarily target system-level context management or non-dialogue memory access are therefore not included.

3.4 Settings

To ensure fair comparison, all methods are evaluated using the same base language model and identical decoding configurations. We use **GPT-4o-mini** as the backbone LLM and a shared embedding model for vector representations. For each evaluation setting, we conduct three independent trials with fixed prompts.

For baseline-comparative main results, we report mean±std over multiple runs to reflect run-to-run variability. For auxiliary analyses and ablations (e.g., Table 3 and Table 4), we report mean values only for compact presentation, as these results are primarily intended to validate relative trends rather than to serve as headline comparisons.

Additional implementation details, including model configurations and hardware specifications, are provided in the Appendix.

4 Results and Analyses

4.1 Main Results

We evaluate HiMem on the LoCoMo benchmark to assess its ability to preserve, retrieve, and utilize information over long-horizon dialogues. Table 1 reports the performance of HiMem and representative baseline methods across diverse reasoning categories, including Single-Hop, Multi-Hop, Temporal Reasoning, and Open-Domain questions.

Overall, HiMem consistently outperforms all baselines across almost all categories. In particular, HiMem achieves substantial improvements on Multi-Hop and Temporal Reasoning tasks, which require aggregating scattered evidence across long interaction histories. These results indicate that hierarchical memory organization enables more effective modeling of long-range dependencies and semantic consistency than flat or monolithic memory structures. Moreover, the strong performance on Open-Domain questions suggests that HiMem can reliably integrate dialogue-derived knowledge with external or implicit information over extended time spans.

4.2 Ablation Study: Memory Components

To examine the contribution of different memory components, we conduct an ablation study by selectively removing Episode Memory or Note Memory from HiMem. The results are shown in Table 2.

Removing Episode Memory leads to a pronounced performance degradation across most categories, particularly on Multi-Hop and Temporal

Table 1: **Performance comparison of HiMem and baseline methods on LoCoMo.** Results are reported as mean (std) over three runs in percentages (%). Best results of GPT-Score are shown in **bold**, and second-best results are underlined.

Task	A-MEM		SeCom		Mem0		HiMem	
	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1
Single Hop	59.33(0.51)	34.45(0.46)	87.02(0.35)	23.70(0.06)	75.90(0.74)	53.05(0.65)	**89.22(0.06)**	43.93(0.24)
Multi Hop	40.78(0.77)	20.98(0.05)	59.10(1.17)	13.21(0.01)	56.62(2.86)	32.90(1.11)	**70.92(0.77)**	28.32(0.05)
Temporal	50.26(1.55)	35.84(0.26)	33.54(0.39)	4.28(0.06)	68.54(0.51)	56.37(0.74)	**74.77(0.25)**	22.05(0.22)
Open Domain	24.65(2.14)	9.30(0.50)	**60.07(0.49)**	8.57(0.10)	42.36(0.49)	22.70(0.20)	54.86(1.30)	18.92(0.45)
Overall	51.88(0.52)	30.71(0.29)	69.03(0.24)	16.77(0.02)	68.74(0.98)	48.16(0.73)	**80.71(0.21)**	34.95(0.11)

Table 2: **Ablation study on memory components in HiMem. HiMem** includes both *Note Memory* and *Episode Memory*; **w/o Note Memory** removes the Note Memory while retaining the Episode Memory; **w/o Episode Memory** removes the Episode Memory while retaining the Note Memory.

Task	HiMem		- w/o Episode		- w/o Note	
	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1
Single Hop	**89.22(0.06)**	43.93(0.24)	76.50(0.24)	41.09(0.09)	89.02(0.20)	45.14(0.03)
Multi Hop	**70.92(0.77)**	28.32(0.05)	56.26(0.33)	26.29(0.12)	70.33(0.44)	26.13(0.25)
Temporal	**74.77(0.25)**	22.05(0.22)	68.12(0.59)	23.65(0.31)	72.48(0.39)	29.35(0.16)
Open Domain	**54.86(1.30)**	18.92(0.45)	48.26(0.49)	22.58(0.55)	48.61(0.98)	16.81(0.15)
Overall	**80.71(0.21)**	34.95(0.11)	69.29(0.05)	33.59(0.06)	79.63(0.22)	36.60(0.02)

Reasoning tasks. This observation highlights the importance of preserving fine-grained contextual evidence aligned with the original interaction process. Without Episode Memory, the system struggles to recover detailed event-level information necessary for tracing complex reasoning chains over long dialogues.

In contrast, removing Note Memory results in a smaller but still consistent performance drop. This suggests that structured knowledge representations primarily serve to accelerate information localization and stabilize semantic anchors, while detailed contextual evidence remains indispensable for coverage and reasoning. Together, these findings demonstrate that Episode Memory and Note Memory play asymmetric yet complementary roles: effective long-term memory systems must balance information fidelity and abstraction rather than relying solely on either raw dialogue context or aggressively compressed representations.

4.3 Ablation Study: Knowledge Alignment

We further examine the role of the **Knowledge Alignment** module by comparing different memory types with and without a unified semantic alignment space. As shown in Table 3, disabling Knowledge Alignment causes a pronounced performance drop for *Note Memory*, indicating that a unified semantic space is crucial for extraction-based memories that do not retain raw dialogue context. Such alignment substantially improves intent understanding and memory localization. In contrast, removing Knowledge Alignment for *Episode Memory* slightly improves performance, suggesting that when segmentation is well-structured, additional semantic fusion may dilute information inherent in raw dialogue. Overall, although extracted knowledge representations are more compact and explicit, they are more sensitive to implicit semantics and coreference resolution, and therefore benefit more from unified semantic alignment.

4.4 Memory Self-Evolution

During best-effort retrieval, when *Note Memory* fails to return self-validated results while *Episode Memory* provides sufficient evidence, HiMem triggers the **Memory Reconsolidation** mechanism. Specifically, the system performs query-conditioned information extraction over retrieved Episode Memory results and supplements missing knowledge in Note Memory through conflict detection and dynamic updating. As shown in Fig-

Table 3: **Ablation study of the Knowledge Alignment module.** For clarity, we denote Knowledge Alignment as KA. (1) **HiMem**: alignment applied only to Note Memory; (2) **HiMem w/o KA**: alignment disabled for both Note Memory and Episode Memory; (3) **Note Memory**: alignment applied to extracted knowledge; (4) **Note Memory w/o KA**: extracted knowledge retained without alignment; (5) **Episode Memory**: alignment applied to segmented episodes; (6) **Episode Memory w/o KA**: segmented episodes retained without alignment.

Method	Single Hop		Multi Hop		Temporal		Open Domain		Average	
	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1
HiMem	**89.22**	43.93	**70.92**	28.32	**74.77**	22.05	**54.86**	18.92	**80.71**	34.95
- w/o KA	87.51	43.75	69.86	28.53	75.18	28.14	52.08	15.96	79.50	35.98
Note Memory	**66.51**	35.88	**54.26**	24.92	**67.39**	19.89	**50.35**	23.88	**63.44**	29.79
- w/o KA	61.79	34.16	46.81	23.57	60.12	19.77	42.71	18.52	57.51	28.25
Episode Memory	88.31	45.53	65.25	24.76	71.55	29.08	**48.61**	18.09	78.12	36.59
- w/o KA	**89.02**	45.14	**70.33**	26.13	**72.48**	29.35	**48.61**	16.81	**79.63**	36.60

Table 4: **Ablation study of the Note Memory.** For clarity, we denote Knowledge Alignment as KA, and refer to enabling Memory Self-Evolution as ME.

Method	Single Hop		Multi Hop		Temporal		Open Domain		Average	
	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1
Note Memory										
w/o KA	61.79	34.16	46.81	23.57	60.12	19.77	42.71	18.52	57.51	28.25
+KA	66.51	35.88	54.26	24.92	67.39	19.89	50.35	23.88	63.44	29.79
+KA & +ME	**76.50**	**41.09**	**56.26**	**26.29**	**68.12**	**23.65**	**48.26**	**22.58**	**69.29**	**33.59**

ure 2 and Table 4, enabling Memory Self-Evolution improves Note Memory performance by approximately **5.85%**, which further leads to a slight overall performance gain of about **0.28%**. These results demonstrate that Memory Reconsolidation is an effective mechanism for enabling long-term memory self-evolution.

4.5 Discussion of Extended Analyses

Additional analyses, including the retrieval strategies, hyperparameter sensitivity, and efficiency trade-offs, are provided in the Appendix.

5 Discussion

Long-horizon conversational agents require more than extended context windows or incremental memory accumulation; they critically depend on how information is structured, abstracted, and revised over time. The empirical results consistently demonstrate that hierarchical memory organization is a necessary condition for robust long-term reasoning rather than an optional architectural refinement. Episode Memory and Note Memory play asymmetric yet complementary roles: the former preserves fine-grained contextual evidence aligned with the original interaction process, while the latter consolidates stable, high-frequency knowledge into compact representations that substantially reduce retrieval cost. The performance degradation observed when either component is removed confirms that effective long-term memory systems must balance fidelity and abstraction, instead of relying solely on raw dialogue context or aggressive compression.

Beyond static organization, our findings highlight that memory updating cannot be treated as a purely similarity-driven or append-only process. In long-term interactions, newly observed information often partially overlaps with, extends, or contradicts existing knowledge. HiMem's conflict-aware Memory Reconsolidation explicitly distinguishes these cases and applies differentiated update strategies, which proves essential for maintaining semantic consistency over time. Importantly, the gains brought by memory self-evolution do not arise from heuristic rewriting, but from a conservative feedback loop between retrieval failure, episodic evidence inspection, and targeted knowledge supplementation.

Finally, the comparison between hybrid and best-

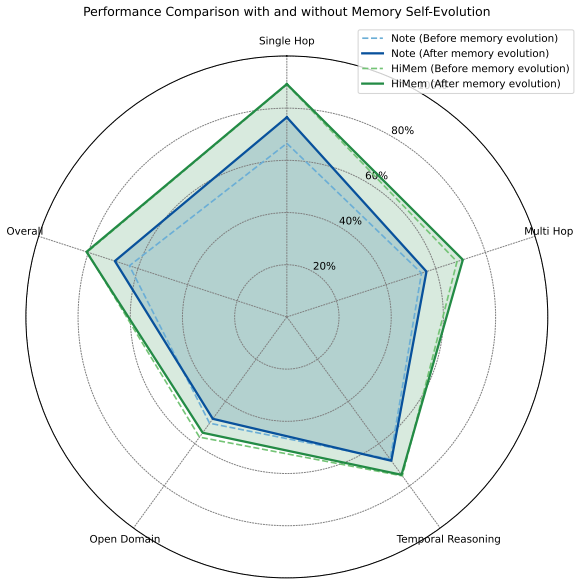

Figure 2: **Performance comparison (GPT-Score) before and after enabling Memory Self-Evolution for Note Memory and HiMem.** Memory Self-Evolution is triggered through conflict-aware *Memory Reconsolidation* during best-effort retrieval.

effort retrieval strategies indicates that hierarchical memory is not only a representational choice but also an efficiency mechanism. Retrieving abstract knowledge first and descending to concrete events only when necessary achieves a favorable trade-off between accuracy and computational cost, while simultaneously exposing latent information that can drive further memory evolution. Together, these observations suggest that long-horizon LLM agents should treat memory as a dynamic, multi-level system tightly coupled with retrieval and usage, rather than as a static external store.

6 Conclusion

This paper proposes **HiMem**, a hierarchical long-term memory framework for long-horizon dialogues, aimed at addressing several fundamental challenges faced by existing LLM agents in sustained interactions, including memory fragmentation, semantic drift, and the lack of self-evolution capability. Grounded in cognitive theories of human long-term memory, HiMem organically integrates event-level experiences with knowledge-level abstractions, and realizes efficient storage, retrieval, and dynamic updating of long-term information through a structured system design.

Methodologically, HiMem constructs cognitively consistent Episode Memory via **Topic-Aware Event–Surprise Dual-Channel Segmentation**, providing fine-grained and semantically stable contextual support for complex reasoning tasks. Meanwhile, through a multi-stage information extraction pipeline and selective **Knowledge Alignment**, high-frequency and stable facts as well as user-specific attributes are consolidated into dense Note Memory representations, significantly reducing retrieval costs while preserving semantic fidelity. Furthermore, HiMem introduces a conflict-aware **Memory Reconsolidation** mechanism that closes the loop between retrieval and memory updating, enabling continuous correction and evolution of knowledge through usage.

Extensive experiments across multiple long-horizon conversational scenarios systematically validate the effectiveness of these design choices. HiMem consistently outperforms existing methods in terms of accuracy, temporal reasoning, and open-domain understanding. Ablation and analysis studies further reveal that these gains arise from the *synergistic interaction* among hierarchical memory structures, cognitively aligned event segmentation, memory-type-aware semantic alignment, and self-evolution mechanisms, rather than from isolated component-level improvements. In addition, analyses of retrieval modes and hyperparameters demonstrate that HiMem achieves a robust balance between knowledge coverage and system efficiency.

Overall, HiMem's contributions extend beyond empirical performance improvements. More importantly, it offers a **practical paradigm for systematically integrating cognitive theories into the design of long-term memory for LLM agents**. By emphasizing memory-type distinctions, structured organization, and usage-driven feedback, this paradigm provides a methodological foundation for building scalable, interpretable, and self-evolving LLM agents. We hope that this work will inspire future research on long-term memory in more complex settings, including multi-agent, multimodal, and richly interactive environments.

Limitations

Although HiMem demonstrates stable and significant performance advantages on long-horizon conversational tasks, several limitations remain that warrant further investigation. These limitations do not stem from flaws in the design itself, but rather reflect broader challenges commonly faced by long-term memory systems in realistic interac-

tive settings.

Dependence on LLM Judgment Capabilities. First, HiMem relies extensively on the semantic and pragmatic judgment capabilities of the underlying LLM during memory construction and updating, including event segmentation, information extraction, conflict detection, and evidence sufficiency evaluation. While experimental results indicate that such one-shot, rule-constrained judgments are stable and effective in practice, their quality inevitably depends on the capability of the base model. In scenarios involving noisy inputs, metaphorical language, or cross-cultural pragmatic variations, the accuracy of segmentation and knowledge extraction may be affected. Future work could explore incorporating lightweight auxiliary classifiers or uncertainty estimation mechanisms at critical decision points to further enhance robustness under complex linguistic conditions.

Expressive Limits of One-Shot Segmentation. Second, HiMem currently adopts a one-shot segmentation strategy, which offers clear advantages in efficiency and controllability, but also imposes an upper bound on expressive capacity. This strategy assumes that the event structure of a conversation can be sufficiently identified through a single global pass. However, in extremely long or highly interleaved dialogues, event boundaries may exhibit hierarchical or recursive structures. Future extensions could investigate multi-granularity or iterative event restructuring strategies, while preserving the simplicity of the current design, to better accommodate non-linear conversational dynamics in Episode Memory.

Conservative Triggers for Knowledge Evolution. Regarding memory self-evolution, HiMem primarily relies on retrieval failure or insufficient evidence as triggers for **Memory Reconsolidation**. While this conservative design promotes stability and avoids unnecessary updates, it may allow certain latent inconsistencies or outdated knowledge to persist if they are not explicitly surfaced during retrieval. Designing more proactive yet noise-resistant evolution triggers remains an open challenge. For example, future work could incorporate user feedback, cross-task consistency checks, or long-term statistical signals to detect and resolve implicit conflicts more effectively.

Limited Evaluation Scope. Finally, although HiMem is evaluated on representative long-horizon dialogue benchmarks, the experiments are mainly confined to single-user, text-based interaction scenarios. Real-world long-term interactions often involve multiple users, multimodal inputs, and richer social contexts, which impose additional demands on memory organization and updating. Extending HiMem to multi-agent or multimodal settings, and studying how memories interact, conflict, and propagate across different agents, constitutes an important direction for future research.

Overall, these limitations highlight key research frontiers in advancing long-term memory systems from *usable* to truly *general-purpose*. HiMem provides a viable pathway for systematically integrating cognitive theories into the design of long-term memory for LLM agents. How to further enhance adaptability and generalization while maintaining structural clarity and interpretability remains a central focus for future work.

Ethical Considerations

We acknowledge that the development of hierarchical long-term memory systems for LLM agents carries significant ethical responsibilities, particularly concerning data privacy, knowledge integrity, and potential societal impacts.

Data Privacy and User Profiling HiMem is designed to extract and store structured information, including user preferences and profiles, to maintain long-term interaction coherence. In real-world applications, this involves the persistent storage of potentially sensitive personal information. We emphasize that any practical implementation of HiMem should adhere to privacy-by-design principles, such as the General Data Protection Regulation (GDPR). This includes implementing robust data encryption, ensuring transparency regarding what information is being "memorized," and providing users with the "right to be forgotten" by allowing them to inspect and delete specific entries in both Episode and Note Memory. In addition, the datasets used in this study are all publicly available and used in accordance with their respective licenses.

Knowledge Integrity and Hallucinations The "Memory Reconsolidation" mechanism introduces a dynamic self-evolution process where the system updates its internal knowledge based on new interactions. While this improves adaptability, it also poses a risk of "consolidating" hallucinations or incorrect information if the backbone LLM makes

erroneous judgments during the conflict-aware update phase. We have mitigated this through a conservative update strategy, but we caution that such systems should not be deployed in high-stakes domains (e.g., medical or legal advice) without human-in-the-loop verification.

Bias Amplification As HiMem relies on the semantic understanding and summarization capabilities of pre-trained LLMs, it may inadvertently inherit or amplify biases present in the foundation models during the memory abstraction process (Stage 1-3). We encourage future research to integrate bias-detection filters within the memory extraction pipeline to ensure that the "Notes" stored do not perpetuate harmful stereotypes or unfair social biases.

Intended Use and Transparency HiMem aims to foster more meaningful and efficient human-AI collaboration. However, the ability of an agent to form a "long-term bond" through persistent memory could potentially be misused for manipulative purposes. We advocate for full disclosure: users should be explicitly informed when they are interacting with an agent equipped with long-term memory capabilities to manage expectations and ensure informed consent.

References

Nick Alonso, Tomás Figliolia, Anthony Ndirango, and Beren Millidge. 2024. Toward conversational agents with context and time sensitive long-term memory. *Preprint*, arXiv:2406.00057.

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

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z. Pan. 2025. Rethinking memory in AI: Taxonomy, operations, topics, and future directions. *Preprint*, arXiv:2505.00675.

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

Gilles Fauconnier and Mark Turner. 2003. Conceptual blending, form and meaning. *Recherches en communication*, 19:57–86.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou-Ammar, and Jun Wang. 2025. Human-inspired episodic memory for infinite context LLMs. In *International Conference on Learning Representations (ICLR)*, Singapore. OpenReview.net.

Peter Gärdenfors. 1988. *Knowledge in Flux: Modeling the Dynamics of Epistemic States*. MIT Press.

Vanessa E. Ghosh and Asaf Gilboa. 2014. What is a memory schema? a historical perspective on current neuroscience literature. *Neuropsychologia*, 53:104–114.

Asaf Gilboa and Hannah Marlatte. 2017. Neurobiology of schemas and schema-mediated memory. *Trends in Cognitive Sciences*, 21(8):618–631.

Kostas Hatalis, Despina Christou, Joshua Myers, Steven Jones, Keith Lambert, Adam Amos-Binks, Zohreh Dannenhauer, and Dustin Dannenhauer. 2023. Memory matters: The need to improve long-term memory in LLM-agents. In *Proceedings of the AAAI Symposium Series*, volume 2, pages 277–280.

Zihong He, Weizhe Lin, Hao Zheng, Fan Zhang, Matt W. Jones, Laurence Aitchison, Xuhai Xu, Miao Liu, Per Ola Kristensson, and Junxiao Shen. 2024. Human-inspired perspectives: A survey on AI long-term memory. *Preprint*, arXiv:2411.00489.

Xun Jiang, Feng Li, Han Zhao, Jiahao Qiu, Jiaying Wang, Jun Shao, Shihao Xu, Shu Zhang, Weiling Chen, Xavier Tang, Yize Chen, Mengyue Wu, Weizhi Ma, Mengdi Wang, and Tianqiao Chen. 2024. Long-term memory: The foundation of AI self-evolution. *Preprint*, arXiv:2410.15665.

Bernal Jiménez Gutiérrez, 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*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. 2025. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning (ICML)*, Vancouver, Canada. OpenReview.net.

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 *International Conference on Machine Learning (ICML)*, Vienna, Austria. OpenReview.net.

Yubo Li, Xiaobin Shen, Xinyu Yao, Xueying Ding, Yidi Miao, Ramayya Krishnan, and Rema Padman. 2025. Beyond single-turn: A survey on multi-turn interactions with large language models. *Preprint*, arXiv:2504.04717.

Jiahao Liu, Shengkang Gu, Dongsheng Li, Guangping Zhang, Mingzhe Han, Hansu Gu, Peng Zhang, Tun Lu, Li Shang, and Ning Gu. 2025. AgentCF++: Memory-enhanced LLM-based agents for popularity-aware cross-domain recommendations. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025)*, pages 2566–2571, Padua, Italy. ACM.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in 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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

Karim Nader. 2015. Reconsolidation and the dynamic nature of memory. *Cold Spring Harbor Perspectives in Biology*, 7(10):a021782.

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

Kai Tzu-iunn Ong, Namyoung Kim, Minju Gwak, Hyungjoo Chae, Taeyoon Kwon, Yohan Jo, Seungwon Hwang, Dongha Lee, and Jinyoung Yeo. 2025. Towards lifelong dialogue agents via timeline-based memory management. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8631–8661, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

OpenSearch Project. 2025. OpenSearch. `https://opensearch.org`. Version 3.3.2.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *Preprint*, arXiv: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 *International Conference on Learning Representations (ICLR)*, Singapore. OpenReview.net.

Qdrant Team. 2025. Qdrant. `https://qdrant.tech`. Version 1.16.

Hongjin Qian, Zheng Liu, Peitian Zhang, Kelong Mao, Defu Lian, Zhicheng Dou, and Tiejun Huang. 2025. MemoRAG: Boosting long context processing with global memory-enhanced retrieval augmentation. In *Proceedings of The Web Conference (WWW 2025)*, pages 2366–2377, Sydney, Australia. ACM.

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. 2025. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for LLMs. In *International Conference on Learning Representations (ICLR)*, Singapore. OpenReview.net.

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, Yi Zhang, and Yassine Benajiba. 2025. MemInsight: Autonomous memory augmentation for LLM agents. *Preprint*, arXiv:2503.21760.

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

Sentence-Transformers. 2021. sentence-transformers/all-mpnet-base-v2. Hugging Face model card.

Lianlei Shan, Shixian Luo, Zezhou Zhu, Yu Yuan, and Yong Wu. 2025. Cognitive memory in large language models. *Preprint*, arXiv:2504.02441.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*.

Xiaorui Su, Yibo Wang, Shanghua Gao, Xiaolong Liu, Valentina Giunchiglia, Djork-Arné Clevert, and Marinka Zitnik. 2025. KGARevion: An AI agent for knowledge-intensive biomedical QA. In *International Conference on Learning Representations (ICLR)*, Singapore. OpenReview.net.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439, Vienna, Austria. Association for Computational Linguistics.

Yaxiong Wu, Sheng Liang, Chen Zhang, Yichao Wang, Yongyue Zhang, Huifeng Guo, Ruiming Tang, and Yong Liu. 2025a. From human memory to AI memory: A survey on memory mechanisms in the era of LLMs. *Preprint*, arXiv:2504.15965.

Yaxiong Wu, Yongyue Zhang, Sheng Liang, and Yong Liu. 2025b. SGMem: Sentence graph memory for long-term conversational agents. *Preprint*, arXiv:2509.21212.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Wenlin Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, and Tong Xu. 2025a. Towards multi-granularity memory association and selection for long-term conversational agents. *Preprint*, arXiv:2505.19549.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland. Association for Computational Linguistics.

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

Yangyang Yu, Haohang Li, Zhi Chen, Yuechen Jiang, Yang Li, Jordan W. Suchow, Denghui Zhang, and Khaldoun Khashanah. 2025. FinMem: A performance-enhanced LLM trading agent with layered memory and character design. *IEEE Transactions on Big Data*, 11(6):3443–3459.

Xihang Yue, Linchao Zhu, and Yi Yang. 2024. FragRel: Exploiting fragment-level relations in the external memory of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16348–16361, Bangkok, Thailand. 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. *Preprint*, arXiv:2506.07398.

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

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 Transactions on Information Systems*, 43(6):155:1–155:47.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems*.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. MemoryBank: Enhancing large language models with long-term memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 19724–19731.

A Related Work

In recent years, research on long-term memory for LLM agents has progressed along multiple technical directions in parallel, including retrieval-augmented generation, long-context modeling, and structured memory systems. These paradigms respectively emphasize external information access, context capacity expansion, and long-term knowledge organization. However, most existing approaches are developed from isolated design perspectives and lack a unified abstraction to systematically characterize the commonalities, differences, and inherent trade-offs among different memory mechanisms. To this end, we introduce a three-dimensional analytical framework, termed **Memory Form–Memory Organization–Memory Operation**, which revisits the design space of LLM long-term memory systems from the perspectives of memory unit representation, organizational structure, and dynamic operations, providing a unified basis for comparing different research directions.

A.1 Memory Form

Memory Form describes the fundamental representation and granularity of memory units, determining their content structure and serving as the foundational component of long-term memory systems. Early approaches predominantly rely on static segmentation strategies based on dialogue turns or sessions (Tan et al., 2025; Wu et al., 2025b; Salama et al., 2025), which often fail to align with semantic boundaries or the temporal evolution of events. Recent work such as SeCom (Pan et al., 2025) introduces semantic and event boundary detection mechanisms (Event Segmentation) to segment dialogues at the semantic level, leading to notable improvements in semantic coherence and contextual completeness. MemGAS (Xu et al., 2025a) further provides multiple segmentation modes, allowing systems to select different memory granularities according to task objectives.

Regarding memory content construction, SeCom builds structured memory representations through event segmentation and semantic summarization, while Mem0 constructs fragmented fact units via information extraction. A-MEM (Xu et al., 2025b) and THEANINE (Ong et al., 2025) further enrich memory representations by incorporating multi-dimensional features such as entities, relations, and temporal attributes. Collectively, these approaches

explore the balance between information completeness and noise suppression, while supporting temporal reasoning through timestamps or explicit temporal modeling. For instance, A-MEM preserves timestamps to track event order, whereas THEANINE explicitly models temporal dependencies among memory units to capture dynamic semantics.

Despite these advances, existing methods largely focus on explicit semantic segmentation and static feature encoding, without sufficiently modeling implicit semantic dependencies or hierarchical interactions among memory units. In contrast, long-context modeling approaches such as MemGPT (Packer et al., 2023) approach the problem from a system capacity management perspective, emphasizing contextual continuity and memory scheduling rather than semantic structuring of memory content. As a result, they exhibit limitations in fine-grained factual recall and long-term consistency modeling.

Unlike prior work, HiMem refines event granularity through **Dual-Channel Segmentation** and integrates semantic-level fusion with multi-dimensional feature encoding, unifying explicit structural representations with implicit semantic modeling. This design reflects a shift in memory form from structure-centric storage toward semantic-alignment-centric modeling.

A.2 Memory Organization

Memory Organization characterizes how memory units are connected and organized, directly affecting retrieval efficiency and scalability during reasoning. Existing approaches generally follow two main directions. On one hand, structured memory organizations are adopted by methods such as SeCom and A-MEM, which connect memory units linearly along temporal or topical dimensions to maintain semantic continuity. Some frameworks (Rezazadeh et al., 2025; Zhang et al., 2025b; Chhikara et al., 2025; Zhang et al., 2025a) further introduce tree or graph structures to capture cross-event and cross-topic semantic relations. On the other hand, some approaches draw inspiration from human cognition or operating systems (He et al., 2024; Shan et al., 2025). For example, HippoRAG2 (Jiménez Gutiérrez et al., 2025) combines graph structures with vector spaces to simulate hippocampal indexing mechanisms, enhancing semantic association and retrieval accuracy. MemGPT mimics page caching mechanisms in operating systems, organizing memory hierarchically based on access cost and treating the context window as a high-speed working memory.

However, these methods typically seek compromises within a single organizational structure, making it difficult to simultaneously accommodate diverse task requirements and storage efficiency. Moreover, they primarily focus on data structure design while paying limited attention to explicitly modeling hierarchical differences in memory content. To address this limitation, HiMem adopts a multi-level hybrid organization strategy that constructs a hierarchical long-term memory structure spanning from concrete events to abstract knowledge, based on the information density and abstraction level of memory units. During retrieval, semantic filtering progressively narrows down relevant information, significantly reducing computational cost and noise while preserving high precision.

Beyond structural optimization, HiMem's key contribution lies in achieving **Hierarchical Decoupling** in memory organization. By jointly optimizing semantic association, knowledge indexing, and retrieval efficiency within a unified hierarchical framework, HiMem transitions from structure-driven organization to semantic-driven organization.

A.3 Memory Operation

Memory Operation focuses on dynamic memory updates and operational mechanisms during usage, which are critical for long-term adaptability and self-evolution. Self-evolving memory is widely regarded as a core capability for long-horizon LLM agents, enabling systems to continuously learn, update, and refine their knowledge structures through sustained interactions (Jiang et al., 2024). Several works, including Mem0, A-MEM, and THEANINE, support dynamic updates of memory units to reflect new conversational content. MemoryBank (Zhong et al., 2024) introduces a forgetting-curve-based decay mechanism that periodically removes low-frequency or irrelevant information to reduce storage pressure and semantic interference.

MemGPT emphasizes self-management operations of memory: the LLM dynamically reads, writes, and schedules memory across different storage layers based on task requirements, triggering paging and summarization when the context window is constrained, and writing historical information to external storage to maintain contextual continuity. Notably, unlike systems that achieve

Method	Memory Form	Degeneracy	Memory Organization	Degeneracy	Memory Operation	Degeneracy
SeCom	Event-level semantic summaries	Yes	Linear/hybrid index or ganized by time	Yes	Append-only memory construction	Yes
A-MEM	Zettelkasten-style notes (note-level abstraction)	Partial	Linked note graph with local semantic connections	Partial	Incremental update and merge without explicit conflict typing	Yes
Mem0	Atomic factual triples	Yes	Graph-based memory with semantic edges	Partial	Update and deletion based on similarity and recency	Yes
MemGPT	Page-based memory blocks constrained by context window	Yes	External memory swapping driven by capacity management	Yes	Eviction and replacement without semantic consis- tency modeling	Yes
HiMem	**Hierarchical Episode → Knowledge**	**No**	**Semantic-driven hierarchical decoupling**	**No**	**Conflict-aware reconsolidation (assimilation/accommodation)**	**No**

Figure 3: **Mapping of representative long-term memory systems under the Memory Form–Memory Organization–Memory Operation framework.** The framework characterizes long-term memory systems along three dimensions: memory unit representation, organizational structure, and memory operations. When a dimension collapses into a single fixed design choice that restricts adaptive trade-offs among granularity, structure, or temporal evolution, it is considered to exhibit design degeneration. In contrast, HiMem maintains non-degenerate designs across all three dimensions, enabling more flexible and evolvable long-term memory modeling.

self-evolution through content-level updates or reconstructions, MemGPT does not directly modify the semantic structure of memory, but instead manages limited context resources via operating-system-style scheduling and compression.

Nevertheless, although these methods support dynamic updates to some extent, they largely remain at the level of content addition or deletion, lacking mechanisms for conflict awareness and semantic reintegration at the knowledge level. When semantic conflicts arise between new and existing information, systems often struggle to decide whether to retain, merge, or revise memories, leading to degraded knowledge consistency or increased computational overhead due to excessive filtering and rewriting.

To address this limitation, HiMem introduces a conflict-aware dynamic evolution mechanism based on memory type differentiation. For memory units that record objective events, whose semantics are relatively stable, conflict detection is unnecessary. In contrast, for knowledge-oriented memories representing user preferences or personal traits, semantic conflicts trigger assimilation or accommodation operations, enabling Memory Reconsolidation and supporting self-correction and continuous evolution. This design closely aligns with cognitive theories of Memory Reconsolidation (Nader, 2015), allowing LLM agents to maintain consistency during long-term interactions.

As illustrated in Figure 3, we map representative long-term memory systems into the proposed three-dimensional analytical framework. Most existing approaches avoid design degeneration in only one dimension, while implicitly simplifying the remaining dimensions into fixed design choices. For instance, some methods improve memory form through event segmentation or summarization, but retain static assumptions in memory organization and updating mechanisms. Others enhance organizational flexibility or operational strategies, yet remain constrained in memory form and semantic modeling.

In contrast, HiMem is among the few systems that maintain non-degenerate designs across all three dimensions. It introduces a hierarchical "event-to-knowledge" memory form, adopts semantic hierarchy rather than static topology as the organizing principle, and explicitly models conflict-aware reconsolidation in memory operations. This three-dimensional non-degeneracy enables HiMem to support more adaptive and evolvable memory management in long-term interaction scenarios.

B Implementation Details

To ensure fair and stable comparisons, all methods are evaluated under identical experimental settings. We use **GPT-4o-mini** as the base language model for all methods with fixed decoding parameters (`temperature`=0.0, `max_tokens`=8192), set `top-k`=10 for memory selection during retrieval, and adopt **all-mpnet-base-v2** (Sentence-Transformers, 2021) for vector representations. All remaining implementation details strictly follow the official implementations and recommended configurations of the corresponding baselines.

For each benchmark evaluation, we keep model parameters and prompts fixed, run three repeated

evaluations, and report the mean and standard deviation of all key metrics across runs. We apply the same evaluation protocol to HiMem and all baselines. In addition, we use GPT-4o-mini as the judge for GPT-Score.

All experiments are conducted on the same hardware environment—a MacBook Pro with Apple M4 Max and 128GB unified memory—to eliminate hardware differences as a confounding factor for efficiency metrics.

For HiMem, we use **OpenSearch** (OpenSearch Project, 2025) as the storage backend for Episode Memory, and adopt **Qdrant** (Qdrant Team, 2025) to manage knowledge-oriented memory representations in Note Memory.

C Extended Analyses

C.1 Memory Retrieval Modes

We compare two retrieval strategies in HiMem: **hybrid retrieval**, which simultaneously queries Note Memory and Episode Memory, and **best-effort retrieval**, which first queries Note Memory and descends to Episode Memory only when necessary. As shown in Table 5, hybrid retrieval achieves high accuracy while significantly reducing response latency, indicating that the hierarchical memory structure enables parallel localization of relevant information and thus shortens reasoning time. However, as shown in Table 6, hybrid retrieval incurs the highest token consumption due to aggregating information from both memory layers. In contrast, best-effort retrieval introduces additional latency due to self-evaluation and potential lower-layer queries, but consumes substantially fewer tokens since most queries are resolved at the Note Memory level. These results validate that hierarchical retrieval from abstract knowledge to concrete events achieves an effective trade-off between knowledge coverage and system efficiency.

C.2 Hyperparameter Analysis

We further analyze the effect of the top-k parameter ($k \in \{5, 10, 15, 20, 25\}$) on system performance. The results show that increasing k improves retrieval coverage and accuracy, but performance plateaus when $k \geq 10$. Meanwhile, retrieval latency and token consumption increase with larger k, reflecting higher retrieval costs. These findings indicate that HiMem, through fine-grained Episode segmentation and multi-stage knowledge extraction, can capture sufficient information within a small retrieval window, achieving optimal performance without expanding the search scope. In contrast, excessively large k introduces irrelevant information and unnecessary processing overhead. This analysis further confirms the efficiency and information-density advantages of HiMem in long-horizon retrieval.

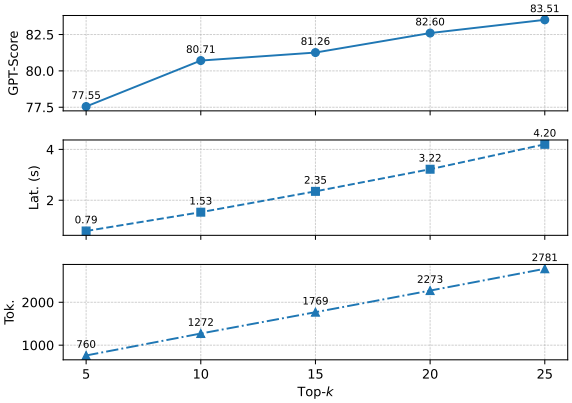

Figure 4: **Effect of top-k on performance and efficiency.** We report GPT-Score, latency, and token consumption as a function of k. Latency reports search latency only.

C.3 Efficiency Analysis

We evaluate efficiency by comparing response latency and token consumption across methods. Although HiMem introduces semantic fusion and hierarchical retrieval, its latency is not significantly worse than that of other baselines, indicating that the computational overhead of lightweight semantic alignment is minimal. In contrast, by leveraging accurate intent modeling and prior knowledge retrieval, HiMem can more rapidly localize required information. In terms of token consumption, HiMem substantially outperforms all baselines, benefiting from its hierarchical memory structure that prioritizes highly condensed knowledge units and thereby reduces context length and generation burden.

D Discussion

The experimental results systematically demonstrate that HiMem's advantages in long-horizon conversational tasks do not stem from isolated improvements in individual components. Instead, they arise from a set of *interdependent design decisions* spanning memory representation, organization, and evolution. Together, these decisions form a closed loop from *information acquisition* to

Table 5: **Performance comparison of HiMem under hybrid and best-effort retrieval strategies.**

Strategy	Single Hop		Multi Hop		Temporal		Open Domain		Average	
	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1	GPT-Score	F1
hybrid retrieval	**89.22**	43.93	**70.92**	28.32	**74.77**	22.05	**54.86**	18.92	**80.71**	34.95
best-effort retrieval	83.59	43.93	62.88	27.42	72.38	25.26	47.92	20.34	75.24	35.54

Table 6: **Efficiency comparison of HiMem under hybrid and best-effort retrieval strategies,** measured by latency and token consumption.

Strategy	Single Hop		Multi Hop		Temporal		Open Domain		Average	
	Lat.(s)	Tok.	Lat.(s)	Tok.	Lat.(s)	Tok.	Lat.(s)	Tok.	Lat.(s)	Tok.
hybrid retrieval	1.57	1292.53	1.55	1292.17	1.35	1177.69	1.67	1343.25	1.53	1271.69
best-effort retrieval	1.63	1016.45	2.03	1257.29	1.88	1200.50	2.66	1583.06	1.82	1134.24

knowledge consolidation and further to *continuous correction*, enabling stable performance in complex and dynamic conversational environments.

D.1 Hierarchical Memory as a Fundamental Constraint for Long-Term Dialogue Modeling

Both the main results and ablation studies consistently indicate that the **hierarchical memory structure** (Episode Memory + Note Memory) constitutes the core prerequisite for long-term conversational performance. Episode Memory preserves fine-grained contextual information aligned with the original interaction process, allowing the system to accurately trace back critical evidence in tasks such as multi-hop reasoning and temporal dependency modeling. In contrast, Note Memory compresses high-frequency and stable knowledge into dense semantic units through information extraction and structured representation, substantially reducing retrieval cost.

The functional asymmetry between these two memory types explains the patterns observed in the ablation results. Removing Episode Memory leads to a severe performance drop, highlighting that *raw contextual information remains indispensable* for complex reasoning tasks. Removing Note Memory also degrades performance, but to a lesser extent, indicating that structured knowledge primarily serves as a mechanism for accelerating localization and stabilizing semantic anchors. These findings underscore a key insight: **the effectiveness of long-term memory systems lies not in complete abstraction, but in maintaining a balance between abstraction and fidelity**.

D.2 The Decisive Role of Dual-Channel Segmentation in Modeling Long-Term Dependencies

The effectiveness of Episode Memory further depends on how it is constructed. Experimental results show that Episode units built via **Topic-Aware Event–Surprise Dual-Channel Segmentation** consistently outperform coarse-grained or purely topic-based segmentation strategies in long-term retrieval and reasoning. This suggests that semantic continuity alone is insufficient to capture event boundaries in real conversations; *cognitive-level discontinuities*, such as shifts in emotion, intent, or discourse function, are equally critical signals for long-term memory modeling.

By explicitly incorporating both "topic continuity" and "surprise-driven discontinuity" criteria at the segmentation stage and fusing them via an OR rule, HiMem generates memory units that better align with human event perception. Such cognitively consistent segmentation reduces cross-segment interference and improves the likelihood that retrieval targets genuinely relevant contexts, which directly manifests as performance gains in Multi-Hop and Temporal Reasoning tasks.

D.3 Selective Effects of Knowledge Alignment Reveal Memory-Type Differences

The ablation study on Knowledge Alignment uncovers a more nuanced and informative phenomenon: **a unified semantic alignment space does not benefit all memory types equally**. For Note Memory, removing Knowledge Alignment re-

Table 7: **Efficiency comparison of HiMem and baseline methods on the LoCoMo dataset.** The lowest latency and smallest token consumption are highlighted in bold, while the second-best results are underlined. Latency (Lat.) reports search latency only and is measured in seconds (s). Notably, SeCom achieves lower latency by preloading per-sample data into memory prior to inference; therefore, its latency is not directly comparable to that of other methods.

Task	A-MEM		SeCom		Mem0		HiMem	
	Lat.(s)	Tok.	Lat.(s)	Tok.	Lat.(s)	Tok.	Lat.(s)	Tok.
Single Hop	**0.93(0.08)**	2698.35(1.20)	-	2738.62(0.05)	4.65(0.13)	1586.37(207.51)	1.57(0.03)	**1292.53(0.03)**
Multi Hop	**0.91(0.14)**	2715.48(3.35)	-	2742.68(0.19)	3.96(0.19)	1588.96(209.64)	1.55(0.03)	**1292.17(0.15)**
Temporal	**0.95(0.12)**	2697.38(1.43)	-	2612.02(0.05)	4.64(0.22)	1591.95(209.63)	1.35(0.02)	**1177.69(0.11)**
Open Domain	**0.94(0.09)**	2675.39(6.14)	-	2732.16(0.73)	4.66(0.19)	1498.12(189.94)	1.67(0.02)	**1343.25(0.18)**
Overall	**0.93(0.10)**	2699.85(1.62)	-	2712.56(0.08)	4.53(0.16)	1582.51(207.25)	1.53(0.03)	**1271.69(0.05)**

sults in a substantial performance drop, indicating that extraction-based memories—once detached from raw dialogue context—rely heavily on semantic normalization processes such as coreference resolution and temporal alignment to maintain retrievability and consistency.

In contrast, enabling Knowledge Alignment for Episode Memory may even degrade performance. This observation suggests that when segmentation already achieves sufficient cognitive coherence, further semantic rewriting or fusion of raw dialogue can obscure implicit cues or fine-grained details. This result highlights a central design principle of HiMem: **semantic alignment should be memory-type aware rather than applied as a uniform preprocessing step**.

D.4 Memory Reconsolidation as a Key Mechanism for Long-Term Self-Evolution

Beyond static memory construction, experiments on Memory Self-Evolution further validate the importance of **Memory Reconsolidation** in sustained interactions. When Note Memory alone cannot support a query but Episode Memory provides sufficient evidence, the system performs targeted information extraction and conflict detection, writing missing knowledge back into Note Memory to correct and enrich its representation.

The effectiveness of this mechanism is evident at two levels. First, Note Memory exhibits a marked performance improvement after self-evolution is enabled, indicating substantive gains in knowledge coverage. Second, the corresponding improvement in overall performance demonstrates that **retrieval and memory updating should form a feedback loop rather than operate as independent processes**. This stands in contrast to memory systems that only support additive or replacement-based updates, and more closely mirrors human cognition, where memory is reshaped through use.

D.5 Global Consistency Across Retrieval Modes and Efficiency Analyses

Finally, comparisons across retrieval modes and the Top-K analysis provide system-level validation of the above design synergies. The complementary behaviors of hybrid and best-effort retrieval in terms of latency and token consumption reflect the flexibility afforded by hierarchical memory under an "abstraction-first, descend-when-necessary" strategy. Moreover, the observation that performance saturates at relatively small K values indicates that HiMem's memory units possess high information density, avoiding inefficient compensation through expanded retrieval scopes.

D.6 Summary

In summary, HiMem's advantages do not arise from larger models or longer context windows, but from a series of *design choices closely aligned with human long-term memory mechanisms*: cognitively consistent event segmentation, memory-type-aware semantic alignment, hierarchically complementary memory representations, and usage-driven Memory Reconsolidation. Together, these elements form a scalable, interpretable, and self-evolving long-term memory framework, providing robust support for sustained interaction and complex reasoning in long-horizon LLM agents.