StructMem: Structured Memory for Long-Horizon Behavior in LLMs

Buqiang Xu[1]*, Yijun Chen[1]*, Jizhan Fang[1], Ruobin Zhong[1], Yunzhi Yao[1], Yuqi Zhu[1,3], Lun Du[2,3], Shumin Deng[1]†

[1]Zhejiang University [2]Ant Group
[3]Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph

Abstract

Long-term conversational agents need memory systems that capture relationships between events, not merely isolated facts, to support temporal reasoning and multi-hop question answering. Current approaches face a fundamental trade-off: flat memory is efficient but fails to model relational structure, while graph-based memory enables structured reasoning at the cost of expensive and fragile construction. To address these issues, we propose **StructMem**, a structure-enriched hierarchical memory framework that preserves event-level bindings and induces cross-event connections. By temporally anchoring dual perspectives and performing periodic semantic consolidation, StructMem improves temporal reasoning and multi-hop performance on LoCoMo, while substantially reducing token usage, API calls, and runtime compared to prior memory systems[1].

1 Introduction

Persistent memory systems are essential for language model agents to maintain coherence in long-term interactions (Park et al., 2023). Beyond factual recall, long-horizon dialogue requires reasoning over temporal dependencies, causal chains, and multi-hop relationships across turns (Weller et al., 2025; Huang et al., 2025; Maharana et al., 2024; Wu et al., 2024; Yang et al., 2018). This necessitates memory representations that organize events into temporally grounded and relational structures (Kwiatkowski et al., 2019).

Existing memory systems largely fall into two paradigms, flat memory and graph memory, exhibit a trade-off between efficiency and structured reasoning, as illustrated in Figure 1. Specifically, flat memory systems (Fang et al., 2026; Zhong et al., 2024; Packer et al., 2023) store facts or summaries as independent units, but fail to preserve cross-event relations, causing retrieval over long histories to degrade into shallow similarity matching (Liu et al., 2023; Zhuang et al., 2026). Graph-based systems (Chhikara et al., 2025; Rasmussen et al., 2025) recover relational structure via entity–relation extraction, but incur high construction cost, require cascaded inference (Edge et al., 2024), and are vulnerable to error accumulation from noisy extractions (Zhuang et al., 2026). We argue that these limitations arise from an inappropriate memory unit. Rather than isolated facts or triplets, the fundamental unit of conversational memory should be a *temporally grounded relational event*, which preserves causal and interpersonal context without imposing rigid schemas.

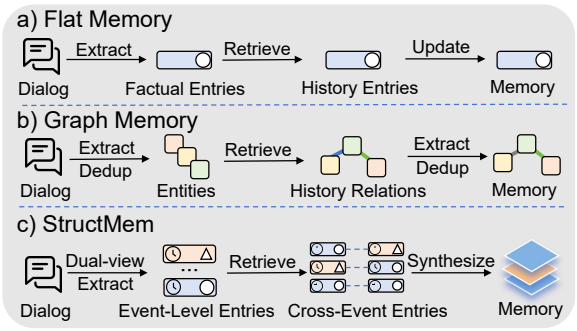


Figure 1: Three paradigms of Memory systems.

Based on this insight, we propose **StructMem**, a hierarchical memory framework built around event-centric representations. This abstraction preserves both what happened and how events relate across agents and time, while avoiding explicit schema design, entity resolution, and symbolic graph traversal. Specifically, at the event level, StructMem constructs structured episodes through dual-perspective extraction, capturing both event content and interactional relations within temporal context. At the cross-event level, it performs periodic consolidation over semantically related events, exploiting temporal locality to efficiently

* Equal contribution.
† Corresponding author.
[1]https://github.com/zjunlp/LightMem

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 122–146
July 2-7, 2026

induce higher-level relational structure. Experiments on LoCoMo show that StructMem improves long-horizon reasoning while significantly reducing computational overhead.

2 Related Work

Long-term memory serves as the cognitive foundation for agents to maintain persona consistency and perform reasoning across extended horizons (Maharana et al., 2024; Wu et al., 2024; Dong et al., 2025; Huang et al., 2025).

Early approaches addressed the context window limitation by externalizing history into flat vector databases (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024). While efficient for semantic matching, this paradigm fundamentally treats interaction history as an unordered bag of propositions, severing the temporal progression, causal dependencies, and relational substrate that bind events into coherent narratives (Gao et al., 2023; Liu et al., 2023). This flat representation leads to fragmented retrieval where isolated facts are returned without the contextual scaffolding necessary for complex reasoning (Weller et al., 2025; Li et al., 2025). Recent work has explored enhanced retrieval strategies through reflective reasoning and closed-loop control mechanisms (Du et al., 2025), yet these improvements still operate within the fundamental constraints of flat representations. Even with extended context windows, flat memory systems suffer from the Lost-in-the-Middle phenomenon (Liu et al., 2023), where attention mechanisms degrade in ultra-long sequences, ultimately reducing multi-hop reasoning to superficial similarity search over disconnected facts (Zhuang et al., 2026).

To bridge this reasoning gap, the field has increasingly pivoted towards structure-enriched architectures, particularly those leveraging Knowledge Graphs. Static graph approaches, such as Microsoft GraphRAG (Edge et al., 2024) and HippoRAG (Gutiérrez et al., 2025), employ hierarchical community detection and Personalized PageRank to facilitate global sense-making and multi-hop traversal. Concurrently, dynamic memory systems tailored for agents, such as Mem0^g (Chhikara et al., 2025) and Zep (Rasmussen et al., 2025), have introduced evolving schemas to capture the fluidity of user interactions. Recent advances further explore trainable graph representations (Xia et al., 2025) and lightweight hierarchical graphs with entity-relation indexing (Huang et al., 2025), demonstrating substantial improvements in multi-agent collaboration (Zhang et al., 2025) and procedural skill reuse (Fang et al., 2025). Despite these advances, imposing explicit graph structures on natural dialogue introduces inherent trade-offs. Compressing fluid narratives into rigid entity-relation triplets often incurs semantic loss (Chaudhri et al., 2022; Zhuang et al., 2026), while extraction instability allows hallucinated relations to propagate as persistent structural noise (Zhong and Chen, 2021; Kolluru et al., 2020). The computational overhead of continuous graph maintenance further poses latency challenges for real-time agentic applications (Edge et al., 2024; Fang et al., 2026).

A parallel line of research seeks a middle ground by enabling structured consolidation without rigid graph schemas. HiMem (Zhang et al., 2026) organizes memory into hierarchical text segments bounded by physical session boundaries, optimizing for compression and retrieval indexing. TiMem (Li et al., 2026) introduces per-turn reflective thinking chains to deepen single-turn understanding, though at the cost of continuous per-turn overhead. PREMem (Kim et al., 2025) shifts inference burden to the memory stage by pre-reasoning user preferences before storage, targeting long-term persona consistency. EMem (Zhou and Han, 2025) prioritizes retrieval faithfulness through raw episode preservation, relying on retrieval-driven passive consolidation rather than active synthesis. MemWeaver (Yu et al., 2025) introduces lightweight entity extraction to organize experiences at the session level.

3 Method

We propose **StructMem**, a framework that achieves structure-enriched organization through hierarchical design. The framework operates at two levels: event-level structure (§3.1) preserves relational bindings within individual utterances, while cross-event structure (§3.2) connects information across temporal boundaries.

3.1 Event-Level Binding

Event-level binding preserves the connection between factual content and relational context within individual utterances through dual-perspective extraction and temporal anchoring.

Dual-Perspective Extraction. For each utterance m_i in the dialogue stream, we extract entries from two complementary perspectives using lan-

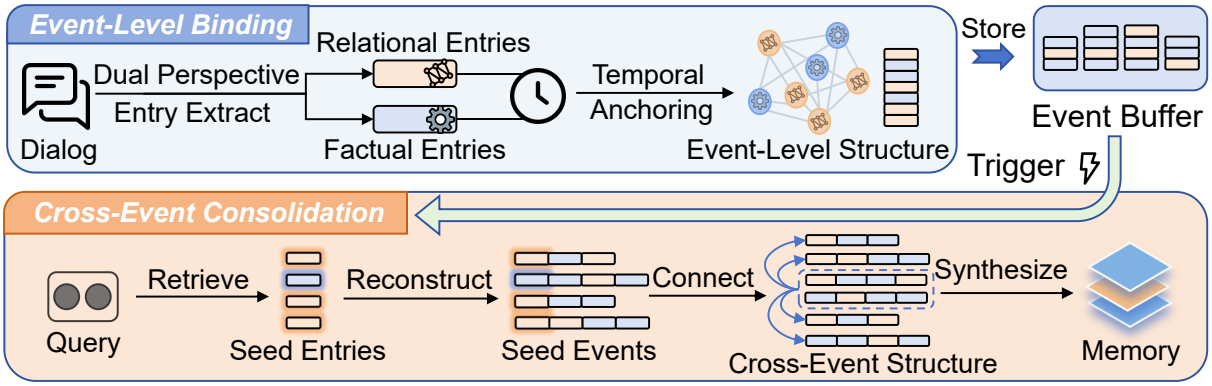


Figure 2: StructMem's hierarchical memory organization. **Event-Level Binding** constructs event-level structure by extracting dual perspectives and anchoring them temporally. **Cross-Event Consolidation** constructs cross-event structure through semantic retrieval, event reconstruction, and consolidation synthesis.

guage model $\mathcal{L}$ with prompts P_{fact} and P_{rel}:

$$\Phi_i \cup \Psi_i = \mathcal{L}(P_{fact} \| m_i) \cup \mathcal{L}(P_{rel} \| m_i), \quad (1)$$

where $\Phi_i = \{c_{i,1}, \ldots, c_{i,j}\}$ contains *factual entries* describing event content, and $\Psi_i = \{r_{i,1}, \ldots, r_{i,k}\}$ contains *relational entries* capturing interpersonal dynamics, causal influences, and temporal dependencies.

By representing both in natural language rather than rigid triplets, we preserve the contextual nuances required for episodic grounding while avoiding entity resolution overhead.

Temporal Anchoring. To preserve the binding between relational and factual information, all entries are anchored to their originating timestamp τ_i, forming an event-level unit:

$$\mathcal{M} \leftarrow \bigcup_{i=1}^{N} \{\langle x, \mathbf{e}_x, \tau_i \rangle \mid x \in \Phi_i \cup \Psi_i\}, \quad (2)$$

where $\mathbf{e}_x$ denotes the embedding of entry x. This temporal coupling enables reconstruction of complete factual-relational events during retrieval.

3.2 Cross-Event Consolidation

Cross-event consolidation connects information across temporal by periodically synthesizing semantically related events. We trigger synthesis when accumulated events exceed a time threshold.

Semantic Event Connections. We buffer unconsolidated entries since the last consolidation. The buffered entries are temporally ordered:

$$\mathcal{C}_{buf} = \text{Sort}_{\tau}\{x \in \mathcal{M}_{buffer}\}, \quad (3)$$

where $\mathcal{M}_{buffer}$ denotes the buffered entries. We encode the buffered context into an aggregated query by concatenating all buffered entry texts and encoding them with an embedding model. We then rank all historical entries by cosine similarity to this query and retrieve the top-K most semantically similar entries as seeds, denoted as $\mathcal{S}_k$.

For each seed entry $x^* \in \mathcal{S}_k$, we reconstruct its complete event context by retrieving all entries sharing the same timestamp:

$$E_{\tau}(x^*) = \{x' \in \mathcal{M} \mid \tau(x') = \tau(x^*)\}. \quad (4)$$

These reconstructed events, together with the buffered events, form the cross-event structure grounded in semantic relevance.

$$\mathcal{C}_{cross} = \mathcal{C}_{buf} \cup \bigcup_{x^* \in \mathcal{S}_k} E_{\tau}(x^*). \quad (5)$$

Memory Consolidation through Synthesis. Unlike conventional summarization that performs lossy compression on sequential text, our consolidation mechanism operates on semantically-reconstructed event clusters. It explicitly synthesizes cross-event relational hypotheses, forming a complementary abstraction layer that enables multi-hop reasoning while faithfully preserving the fidelity of raw episodic memory.

$$\mathcal{M} \leftarrow \mathcal{C}_{cons} = \mathcal{L}(P_{cons} \| \mathcal{C}_{cross}). \quad (6)$$

4 Experiments

4.1 Experimental Setup

We first describe the dataset and evaluation metrics, followed by the baseline systems used for comparison. To ensure reproducibility, complete set of prompt templates and implementation details used for memory construction, question answering, and evaluation is provided in Appendix A.7.

Method	Overall ↑	Performance by Type ↑				Build Tokens (M) ↓			Calls ↓	Time (s) ↓
		Multi	Open	Single	Temp	In	Out	Sum		
OpenAI	71.82	69.86	53.12	84.66	45.48	–	–	–	–	–
FullContext	73.83	68.79	56.25	**86.56**	50.16	–	–	–	–	–
MiniRAG	63.51	56.74	58.33	75.74	38.94	9.022	1.081	10.103	2508	**2566**
LightRAG	68.83	66.31	50.00	77.53	53.89	10.014	1.916	11.931	13576	60469
LangMem	58.10	62.23	47.92	71.12	23.43	9.873	1.192	11.066	5990	26281
A-Mem	64.16	56.03	31.25	72.06	60.44	9.126	2.368	11.494	11754	60607
Mem0	66.88	67.13	51.15	72.93	59.19	10.958	1.239	12.196	9181	30057
MemoryOS	58.25	56.74	45.83	67.06	40.19	1.889	0.939	2.868	5534	24220
Mem0^g	68.44	65.71	47.19	75.71	58.13	33.512	2.313	35.825	53514	115670
Zep	75.14	**74.11**	**66.04**	79.79	67.71	–	–	–	–	–
Memobase	75.78	70.92	46.88	77.17	**85.05**	–	–	–	–	–
StructMem	**76.82**	68.77	46.88	81.09	81.62	**1.501**	**0.436**	**1.937**	**1056**	22854

Table 1: Performance and resource consumption comparison of memory systems on LoCoMo dataset. ↑: larger is better; ↓: smaller is better. **The best results** are marked in bold, the second-best results are underlined. Row colors distinguish method categories: RAG methods , Flat Memory methods , and Structural Memory methods . OpenAI and FullContext have no construction cost; Zep and Memobase do not expose construction details.

Dataset and Metrics. We evaluate on the LoCoMo benchmark (Maharana et al., 2024) (see Appendix A.2 for detailed statistics). Effectiveness is measured using LLM-as-a-judge evaluation; efficiency is measured by token usage, API calls, and runtime during memory construction.

Baselines. We compare StructMem against RAG-based systems (OpenAI, FullContext, MiniRAG, LightRAG), flat memory methods (LangMem, A-Mem, Mem0), and structural memory methods (MemoryOS, Mem0^g, Zep, Memobase). All methods use gpt-4o-mini as the backbone and text-embedding-3-small for embeddings. Detailed retrieval and configuration parameters for all baselines are provided in Appendix A.4.

4.2 Overall Performance

Table 1 shows StructMem achieves state-of-the-art overall performance on LoCoMo, with substantial gains in multi-domain and temporal reasoning where cross-event connections are critical for understanding causal relationships across dialogue sessions. Beyond effectiveness, StructMem demonstrates exceptional efficiency: compared to existing memory systems, it reduces token consumption and requires significantly fewer API calls, as our progressive structural organization avoids the expensive post-hoc graph construction. These results hold consistently across multiple judge models, as verified in Appendix A.5.

4.3 Analysis

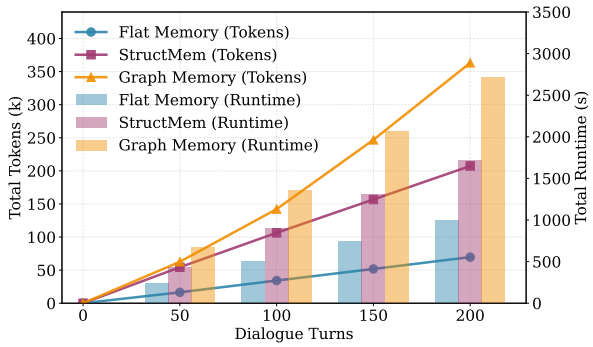
We analyze StructMem from two complementary perspectives: a paradigm-level comparison that evaluates effectiveness and efficiency across all three memory paradigms, and an internal analysis that examines the mechanism underlying StructMem's reasoning gains.

Method	Multi	Open	Single	Temp
Flat Memory	66.31	46.88	78.83	78.50
Graph Memory	66.67	48.96	80.50	76.64
w/o Cross-Event	66.31	46.88	80.86	79.44
StructMem	68.77	46.88	81.09	81.62

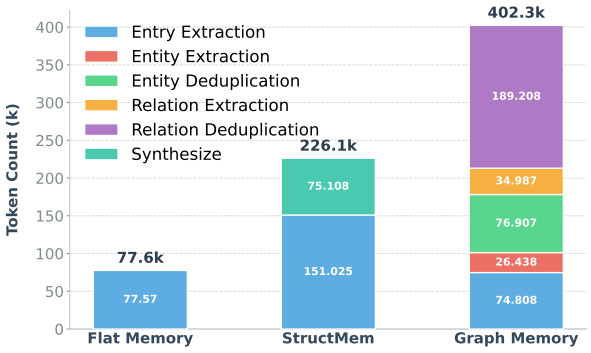
Table 2: Paradigm comparison and ablation study on LoCoMo dataset.

Paradigm Comparison. To validate the effectiveness of each paradigm, we conduct studies in Table 2. Starting from Flat Memory as the baseline, Graph Memory achieves improvements on single-session and open-domain tasks, though it decreases on temporal reasoning. In contrast, our approach demonstrates consistent improvements across all task types. Event-level structure improves performance in temporal reasoning and single-session. Cross-event structure yields further gains by capturing cross-temporal causal relationships.

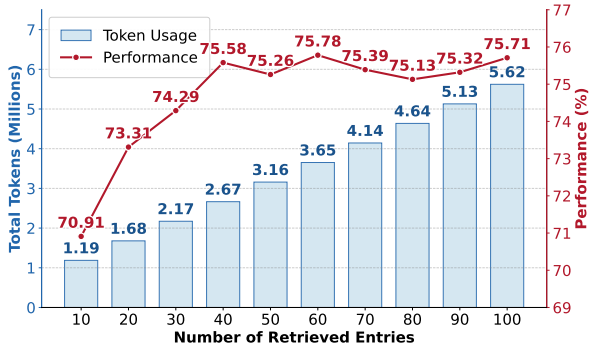
To examine computational efficiency, we analyze token usage and runtime on the first conversation of LoCoMo. Figure 3(a) shows that Graph



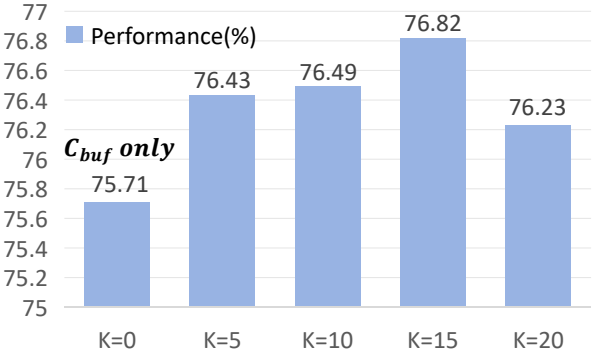
(a) Token consumption over dialogue turns



(b) Component-wise token consumption



(c) Effect of the number of retrieved entries



(d) Effect of the number of semantic retrieval seeds K

Figure 3: Analysis of efficiency across memory paradigms and internal mechanisms of StructMem.

Memory incurs significantly higher token usage and runtime as dialogue progresses. Figure 3(b) reveals the source: graph construction requires four cascading LLM operations per event, with deduplication overhead growing quadratically. In contrast, StructMem achieves efficiency through buffered consolidation: by exploiting temporal locality, in which semantically related events naturally cluster within short time windows, the system accumulates events and processes them in batch during periodic synthesis. This effectively reduces cross-event organization from per-event operations to periodic batch processing, substantially cutting both API calls and token consumption.

StructMem Internal Mechanisms. We analyze whether hierarchical organization provides genuine reasoning gains beyond retrieval scaling.

Figure 3(c) reveals that flat retrieval performance peaks at 60 entries and plateaus thereafter, indicating that simply retrieving more atomic entries cannot improve effectiveness, as the bottleneck is knowledge reasoning rather than coverage. Cross-event consolidation addresses this by synthesizing semantically related events into higher-level relational hypotheses, creating information that does not exist in any individual memory entry.

Figure 3(d) confirms this: without event connections ($K = 0$), performance matches the flat retrieval plateau, but introducing cross-event synthesis yields substantial gains, demonstrating that hierarchical consolidation reconstructs causal relationships across temporal boundaries and enables fundamentally new reasoning capabilities. Fidelity analyses in Appendix A.6 further confirm that these synthesized connections are well-grounded, with minimal spurious associations.

5 Conclusion

We propose StructMem, which achieves structure-enriched organization through hierarchical design: preserving event-level bindings and enabling cross-event consolidation, StructMem preserves temporal and relational structures without the computational overhead of continuous graph maintenance. Experiments on LoCoMo demonstrate that StructMem achieves better performance with strong results in multi-hop and temporal reasoning, while substantially reducing token consumption, API calls, and runtime compared to prior memory systems.

Limitations

Despite its strong performance, StructMem has several limitations. The quality of dual-perspective extraction is highly dependent on instruction prompts, where suboptimal design may result in incomplete or inaccurate relational information capture. Future research could investigate automated prompt optimization to improve robustness across various dialogue contexts. Additionally, the framework primarily addresses memory expansion and synthesis but currently lacks an explicit mechanism for conflict resolution and memory updating. As user facts or preferences may evolve over long horizons, the absence of a revision process could lead to inconsistencies between historical summaries and new information. Future iterations should incorporate memory decay or updating strategies to ensure the hierarchical organization accurately reflects the most current state of the interaction.

Acknowledgements

We would like to express sincere gratitude to the reviewers for their thoughtful and constructive feedback. This work was supported by the National Natural Science Foundation of China (No. 62576307), Yongjiang Talent Introduction Programme (2021A-156-G), and Information Technology Center and State Key Lab of CAD&CG, Zhejiang University. This work was supported by Ant Group and Zhejiang University - Ant Group Joint Laboratory of Knowledge Graph.

References

Vinay K. Chaudhri, Chaitanya Baru, Naren Chittar, Xin Luna Dong, Michael Genesereth, James Hendler, Aditya Kalyanpur, Douglas B. Lenat, Juan Sequeda, Denny Vrandečić, and Kuansan Wang. 2022. Knowledge graphs: introduction, history and, perspectives. *AI Magazine*, 43(1):17–29.

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

Cody V Dong, Qihong Lu, Kenneth A Norman, and Sebastian Michelmann. 2025. Towards large language models with human-like episodic memory. *Trends in Cognitive Sciences*.

Xingbo Du, Loka Li, Duzhen Zhang, and Le Song. 2025. Memr^3: Memory retrieval via reflective reasoning for llm agents. *Preprint*, arXiv:2512.20237.

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. *arXiv preprint arXiv:2404.16130*.

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. 2026. Lightmem: Lightweight and efficient memory-augmented generation. In *The Fourteenth International Conference on Learning Representations*.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2025. Memp: Exploring agent procedural memory. *arXiv preprint arXiv:2508.06433*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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 *Forty-second International Conference on Machine Learning*.

Zhengjun Huang, Zhoujin Tian, Qintian Guo, Fangyuan Zhang, Yingli Zhou, Di Jiang, and Xiaofang Zhou. 2025. Licomemory: Lightweight and cognitive agentic memory for efficient long-term reasoning. *arXiv preprint arXiv:2511.01448*.

Sangyeop Kim, Yohan Lee, Sanghwa Kim, Hyunjong Kim, and Sungzoon Cho. 2025. Pre-storage reasoning for episodic memory: Shifting inference burden to memory for personalized dialogue. *arXiv preprint arXiv:2509.10852*.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. 2020. OpenIE6: Iterative Grid Labeling and Coordination Analysis for Open Information Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3748–3761, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Kai Li, Xuanqing Yu, Ziyi Ni, Yi Zeng, Yao Xu, Zheqing Zhang, Xin Li, Jitao Sang, Xiaogang Duan, Xuelei Wang, Chengbao Liu, and Jie Tan. 2026.

Timem: Temporal-hierarchical memory consolidation for long-horizon conversational agents. *arXiv preprint arXiv:2601.02845*.

Mo Li, L. H. Xu, Qitai Tan, Long Ma, Ting Cao, and Yunxin Liu. 2025. Sculptor: Empowering llms with cognitive agency via active context management. *Preprint*, arXiv:2508.04664.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

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.

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

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee. 2025. On the theoretical limitations of embedding-based retrieval. *Preprint*, arXiv:2508.21038.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. Longmemeval: Benchmarking chat assistants on long-term interactive memory.

Siyu Xia, Zekun Xu, Jiajun Chai, Wentian Fan, Yan Song, Xiaohan Wang, Guojun Yin, Wei Lin, Haifeng Zhang, and Jun Wang. 2025. From experience to strategy: Empowering llm agents with trainable graph memory. *arXiv preprint arXiv:2511.07800*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 2369–2380.

Shuo Yu, Mingyue Cheng, Daoyu Wang, Qi Liu, Zirui Liu, Ze Guo, and Xiaoyu Tao. 2025. Memweaver: A hierarchical memory from textual interactive behaviors for personalized generation. *arXiv preprint arXiv:2510.07713*.

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025. G-memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*.

Ningning Zhang, Xingxing Yang, Zhizhong Tan, Weiping Deng, and Wenyong Wang. 2026. Himem: Hierarchical long-term memory for llm long-horizon agents. *arXiv preprint arXiv:2601.06377*.

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.

Zexuan Zhong and Danqi Chen. 2021. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, Online. Association for Computational Linguistics.

Sizhe Zhou and Jiawei Han. 2025. A simple yet strong baseline for long-term conversational memory of llm agents. *arXiv preprint arXiv:2511.17208*.

Luyao Zhuang, Shengyuan Chen, Yilin Xiao, Huachi Zhou, Yujing Zhang, Hao Chen, Qinggang Zhang, and Xiao Huang. 2026. LinearRAG: Linear graph retrieval augmented generation on large-scale corpora. In *The Fourteenth International Conference on Learning Representations*.

A Appendix

A.1 License

This work uses the LoCoMo benchmark dataset, which is publicly available for academic research purposes. We follow all usage terms specified by the dataset authors.

A.2 Dataset

We evaluate on the **LoCoMo** benchmark (Maharana et al., 2024), which contains 10 long-term conversations with an average of 588 turns and 16,618 tokens per conversation. We focus on the question answering task, utilizing four reasoning types from the benchmark. Table 3 shows the statistics of questions used in our evaluation. Model performance is evaluated using LLM-as-a-judge.

Reasoning Type	# Questions
Single-hop	841
Multi-hop	282
Temporal	321
Open-domain	96

Table 3: Statistics of LoCoMo questions used.

A.3 Implementation Details

We provide key implementation details for StructMem to facilitate reproducibility. **Memory construction:** We set the time window threshold to 1 hour for triggering consolidation. For cross-event consolidation, we retrieve top-15 semantically similar seed entries from historical memory. **Question answering:** During inference, we retrieve 60 entries and 5 synthesis from memory to provide context for answer generation.

A.4 Baseline Configurations

To ensure empirical rigor and reproducibility, we provide the detailed retrieval and architectural configurations for all evaluated systems:

FullContext utilizes the entire raw dialogue history fed into the prompt in reverse chronological order via a full-scan with $k = -1$. **OpenAI** processes all conversation turns concatenated as a flat, unordered text sequence directly without a retrieval step.

MiniRAG and **LightRAG** retrieve the top-20 relevant entries per question to provide factual context. Similarly, **A-MEM** and **LangMem** employ a global search mechanism to retrieve the top-40 most relevant memory entries for each query.

MemoryOS implements a three-tier hierarchical system featuring exhaustive recall of all Short-Term Memory (STM) pages, a two-stage selection for Mid-Term Memory (MTM) comprising the top-5 segments and top-10 dialogue pages, and the extraction of the top-10 relevant entries from Long-term Personal Memory (LPM).

For API-based systems including **Mem0**, **Mem0^g**, **Zep**, and **Memobase**, the top-10 relevant memories per speaker are retrieved for response generation.

A.5 Robustness of Evaluation

We validate the reliability of our LLM-as-a-judge protocol by conducting extensive cross-model evaluations and statistical analyses.

Table 4 summarizes the performance of memory systems across three distinct judge model families: gpt-4o-mini, Qwen2.5-32b-Instruct, and DeepSeek-V3.2. We further calculate the inter-judge agreement and correlation across all judge pairs, as detailed in Table 5. The Fleiss' κ among different models reaches **0.8341**, reflecting a near-perfect agreement that substantially exceeds the commonly accepted reliability threshold of 0.8. This high level of consensus, combined with significant Pearson correlation coefficients ($r > 0.81$, $p < 10^{-300}$), confirms that the automated evaluation protocol provides a stable and objective assessment of semantic response quality.

A.6 Fidelity and Hallucination Study

We conducted a systematic study to ensure that the induced structures are grounded in the source dialogue.

Event-Level Extraction Fidelity. We first evaluated whether the atomic memory entries accurately reflect the original utterances. We employed three independent judge models (gpt-4o-mini, Qwen2.5-32B-Instruct, and DeepSeek-V3.2) to identify hallucinated entries across conversations.

Specifically, for each extracted memory entry, the judges are provided with the corresponding source dialogue segment and tasked with determining if any factual or relational information is fabricated or unsubstantiated by the original text. The full prompt templates are provided in Figure 17.

As detailed in Table 6, the mean hallucination rate is only 2.36%, confirming that the **Dual-Perspective Extraction** of our hierarchical memory is highly faithful to the source context.

Cross-Event Consolidation Fidelity. The most critical verification involves the synthesis of cross-event links. To isolate and audit these links, we employed three independent judge models (gpt-4o-mini, Qwen2.5-32B-Instruct, and DeepSeek-V3.2) to identify hallucinated links across conversations.

For each consolidation step, the judge is provided with: (1) *Buffer Text* (current events), (2) *Supplementary Text* (retrieved history), and (3) two summaries, including **Summary A** (Baseline, $k = 0$, consolidates only buffer events) and **Summary B** (Test, $k = 15$, establishes cross-event links). The judge identifies cross-event links present in Summary B that are absent from Summary A, then classifies each cross-event link to judge if the link

Table 4: Robustness check of memory systems across different LLM judges on the LoCoMo dataset. The table is categorized by three judge models: gpt-4o-mini, Qwen2.5-32B-Instruct, and DeepSeek-V3.2. **Bold** and underline denote the best and second-best results within each judge block, respectively. ↑: larger is better.

Judge Model	Method	Overall ↑	Single Hop	Multi Hop	Temporal	Open Domain
gpt-4o-mini	FullContext	73.83	**86.56**	68.79	50.16	56.25
	A-MEM	64.16	72.06	56.03	60.44	31.25
	MemoryOS	58.25	67.06	56.74	40.19	45.83
	Memobase	75.78	77.17	70.92	**85.05**	46.88
	Zep	75.14	79.79	**74.11**	67.71	**66.04**
	StructMem	**76.82**	81.09	68.77	81.62	46.88
Qwen2.5-32B-Instruct	FullContext	71.17	**83.83**	67.02	48.29	48.96
	A-MEM	60.26	68.85	53.19	52.65	31.25
	MemoryOS	60.32	69.80	62.41	39.88	39.58
	Memobase	76.36	78.00	71.28	**85.05**	47.92
	Zep	75.52	79.19	**75.18**	70.40	**61.46**
	StructMem	**77.01**	82.16	69.86	78.19	48.96
DeepSeek-V3.2	FullContext	70.97	**85.49**	67.02	41.43	54.17
	A-MEM	63.90	74.55	56.38	47.35	47.92
	MemoryOS	61.56	72.53	62.06	35.83	50.00
	Memobase	79.29	80.86	**77.66**	**85.67**	48.96
	Zep	75.45	80.98	75.89	64.80	**61.46**
	StructMem	**79.35**	85.14	73.05	77.57	53.12

Table 5: Inter-judge agreement and correlation across different judge model pairs. GPT, Qwen, and DS denote gpt-4o-mini, qwen2.5-32b-instruct, and DeepSeek-V3.2, respectively.

Judge Pair	Cohen's κ	Pearson r	p-value
Qwen vs. DS	0.8395	0.8438	$< 10^{-300}$
Qwen vs. GPT	0.8326	0.8362	$< 10^{-300}$
DS vs. GPT	0.8184	0.8234	$< 10^{-300}$
Overall (Fleiss' κ)	**0.8341**	—	—

Table 6: Hallucination rates in the event-level extraction stage across 10 conversations.

Conversation	DS	Qwen	GPT	Mean
conv-26	2.07%	0.52%	0.78%	1.12%
conv-30	1.81%	0.60%	4.83%	2.41%
conv-41	2.04%	1.88%	2.35%	2.09%
conv-42	3.16%	1.97%	3.94%	3.02%
conv-43	2.94%	1.63%	3.10%	2.56%
conv-44	1.68%	0.92%	1.68%	1.43%
conv-47	5.28%	2.44%	3.05%	3.59%
conv-48	2.71%	1.45%	1.08%	1.75%
conv-49	3.25%	1.16%	1.62%	2.01%
conv-50	3.55%	2.84%	4.26%	3.55%
Overall	**2.84%**	**1.61%**	**2.63%**	**2.36%**

is spurious. The full prompt templates are provided in Figure 18 and Figure 19.

To evaluate the specific impact of our grounding anchors, we conduct a sensitivity analysis by comparing our default *Constrained* prompt against an *Unconstrained* variant. As illustrated in Figure 20, the *Unconstrained* version is created by removing explicit requirements for timestamp citations and concrete dependency focus (highlighted in gray).

The results in Table 7 demonstrate that removing these grounding constraints leads to a dramatic surge in hallucination rates across all judge models. This trend underscores that the high fidelity of StructMem's hierarchical organization is directly tied to our constrained synthesis mechanism, confirming that the **Memory Consolidation** of our hierarchical memory is highly faithful to the source context.

A.7 Prompt Templates

We present the prompt templates used for memory construction, question answering, and evaluation in StructMem.

For memory construction, we design prompts for different paradigms implemented in the LightMem framework. For Flat Memory, the factual entry extraction prompt (Figure 4 and Figure 5) guides the model to decompose utterances into objective event descriptions. For StructMem, the relational entry extraction prompt (Figure 6 and Figure 7) instructs the model to capture interaction dynamics, causal

Table 7: Detailed cross-event link quality comparison across conversations. **S**, **T**, and **R** denote the number of **S**purious links, **T**otal links, and the error **R**ate (%), respectively. `GPT`, `Qwen`, and `DS` represent `gpt-4o-mini`, `Qwen2.5-32B-Instruct`, and `DeepSeek-V3.2`. *Constrained* is the default setting for StructMem.

Conversation	Config	Judge: GPT			Judge: Qwen			Judge: DS		
		S	T	R (%)	S	T	R (%)	S	T	R (%)
conv-26	Constrained	0	83	0.00	3	70	4.29	1	12	8.33
	Unconstrained	4	72	5.56	23	101	22.77	10	58	17.24
conv-30	Constrained	0	73	0.00	0	59	0.00	1	23	4.35
	Unconstrained	4	84	4.76	6	79	7.59	2	44	4.55
conv-41	Constrained	4	108	3.70	1	131	0.76	1	31	3.23
	Unconstrained	13	104	12.50	20	115	17.39	6	74	8.11
conv-42	Constrained	0	95	0.00	5	93	5.38	1	47	2.13
	Unconstrained	8	100	8.00	20	106	18.87	6	65	9.23
conv-43	Constrained	0	104	0.00	5	130	3.85	6	40	15.00
	Unconstrained	4	109	3.67	11	114	9.65	9	74	12.16
conv-44	Constrained	0	110	0.00	9	91	9.89	2	39	5.13
	Unconstrained	6	104	5.77	30	135	22.22	16	88	18.18
conv-47	Constrained	1	107	0.93	4	90	4.44	0	33	0.00
	Unconstrained	11	103	10.68	32	108	29.63	16	69	23.19
conv-48	Constrained	1	102	0.98	2	101	1.98	0	36	0.00
	Unconstrained	5	100	5.00	27	107	25.23	11	62	17.74
conv-49	Constrained	0	94	0.00	2	90	2.22	0	52	0.00
	Unconstrained	7	81	8.64	14	86	16.28	10	61	16.39
conv-50	Constrained	0	106	0.00	2	113	1.77	1	45	2.22
	Unconstrained	10	109	9.17	30	114	26.32	18	92	19.57
Overall	**Constrained**	**6**	**982**	**0.61%**	**33**	**968**	**3.41%**	**13**	**358**	**3.63%**
	Unconstrained	**72**	**966**	**7.45%**	**213**	**1065**	**20.00%**	**104**	**687**	**15.14%**

influences, and temporal dependencies. The narrative synthesis prompt (Figure 8) consolidates local and retrieved contexts into coherent summaries during Macro Synthesis. For Graph Memory, the entity extraction prompt (Figure 9) identifies key entities from dialogue. The entity deduplication prompt (Figure 10) normalizes extracted entities to eliminate redundancy. The relation extraction prompt (Figure 11) constructs connections between entities. The relation deduplication prompt (Figure 12) resolves contradictions in the knowledge graph.

For question answering, we provide separate prompts tailored to different memory architectures. Figure 13 shows the prompt for StructMem with dual-circuit retrieval that leverages both atomic entries and consolidated summaries. Figure 14 and Figure 15 present prompts adapted for flat memory and graph-based memory baselines, respectively.

For evaluation, we use the LLM-as-a-judge prompt (Figure 16) to assess response correctness and coherence.

For fidelity and hallucination analysis, we provide the specialized templates used for memory auditing. Figure 17 presents the prompt for verifying event-level extraction. Figures 18 and 19 presents the prompt for verifying cross-event consolidation. We also include the *Unconstrained* synthesis template in Figure 20, where the grounding constraints are intentionally removed to evaluate the impact of explicit temporal anchors on reducing hallucinated associations.

A.8 Case Study

Table 8 presents a case study comparing how different memory paradigms handle temporal reasoning

Query	*When did Caroline and Melanie go to a pride festival together?*
Method	**Retrieved Content**
Flat Memory	**Factual Entries:** • Caroline attended pride parade on 2023-08-11 • Caroline had a blast at Pride fest last year (recorded 2023-08-17) • Melanie enjoyed time with the whole gang at Pride fest (recorded 2023-08-17)
Graph Memory	**Factual Entries:** • Caroline attended pride parade on 2023-08-11 • Caroline had a blast at Pride fest last year (recorded 2023-08-17) • Melanie enjoyed time with the whole gang at Pride fest (recorded 2023-08-17) **Entity-Relation Graph:** • caroline → attended → pride_parade • caroline → had_blast_at → pride_fest • melanie → enjoyed_time_at → pride_fest • melanie → expressed_excitement → caroline's_pride_involvement
StructMem	**Event Memory:** • Caroline attended pride parade on 2023-08-11 • Caroline had a blast at Pride fest last year (recorded 2023-08-17) • Melanie showed interest in Caroline's pride parade experience • Melanie enjoyed time with the whole gang at Pride fest (recorded 2023-08-17) • Melanie expressed excitement about Caroline's LGBTQ+ community involvement **Synthesis Memory:** "On August 17, 2023... **As they reminisced about their enjoyable time at Pride fest last year**, Melanie suggested planning a family outing, while Caroline proposed a special outing just for the two of them this summer..."
Prediction	**Flat Memory:** "They haven't gone together." **Graph Memory:** "Last month, June 2023." **StructMem:** "Last year, August 2022." **Reference:** 2022

Table 8: Case study comparing three memory paradigms on joint participation reasoning. Flat Memory and Graph Memory cannot establish co-participation from isolated entries, while StructMem's synthesis correctly identifies shared experiences.

over joint participation. The query asks when two speakers attended an event together, requiring inference over co-participation relationships that are not explicitly stated in individual conversational turns.

Flat Memory retrieves factual entries independently: Caroline attended Pride fest "last year" (temporally anchored to August 17, 2023, referring to 2022), while Melanie enjoyed time "with the whole gang" at Pride fest. Without any mechanism to connect these isolated facts, the system concludes "they haven't gone together," failing to recognize the implicit joint participation.

Graph Memory constructs entity-relation triples on top of the same factual entries. While the graph captures individual attendance, these remain isolated nodes without explicit co-participation edges. The post-hoc graph structure cannot infer that mentions of the same event by different speakers within the same conversation indicate joint attendance. Consequently, it produces an incorrect temporal inference: "Last month, June 2023."

StructMem addresses this limitation through two mechanisms. First, relational entries capture interpersonal dynamics during extraction: "Melanie showed interest in Caroline's pride parade experience" provides crucial context about their shared discussion. Second, synthesis consolidates temporally co-located entries. When Caroline's Pride fest mention appears adjacent to Melanie's in the chronologically sorted context, the relational entry's possessive pronoun "their" signals joint participation. The synthesis then makes this implicit connection explicit: "their enjoyable time at Pride fest last year," enabling the system to correctly answer "Last year, August 2022."

This case demonstrates why extraction-time structural capture outperforms post-hoc graph construction for temporal reasoning. By organizing information hierarchically during memory formation rather than overlaying structure afterward, StructMem preserves the temporal and relational context necessary for inferring implicit relationships across conversational turns.

Prompts for Factual Entry Extract

```
[SYSTEM]: You are a Personal Information Extractor.
Your task is to extract **all possible facts or information** about the speakers from a
conversation,
where the dialogue is organized into topic segments separated by markers like:
--- Topic X ---
[timestamp, weekday] <source_id>.<SpeakerName>: <message>
...
Note: Messages may include visual context marked as [visual_context: ...] which provides
additional scene information.
Important Instructions:
0. You MUST process messages **strictly in ascending source_id order** (lowest → highest).
   For each message, stop and **carefully** evaluate its content before moving to the next.
   Do NOT reorder, batch-skip, or skip ahead — treat messages one-by-one.
1. You MUST process every user message in order, one by one.
   For each message, decide whether it contains any factual information.
   - If yes → extract it and rephrase into a standalone sentence.
   - Do NOT skip just because the information looks minor, trivial, or unimportant.
     Extract ALL meaningful information including:
     * Past events and current states
     * Future plans and intentions
     * Thoughts, opinions, and attitudes
     * Wants, hopes, desires, and preferences
2. **CRITICAL - Preserve All Specific Details**:
   When extracting facts, you MUST include ALL specific entities and details mentioned:
   - **Full names with context**: "The Name of the Wind" by Patrick Rothfuss (not just "a
book")
   - **Complete location names**: Galway, Ireland; The Cliffs of Moher; Barcelona (not just
"a city")
   - **Specific event names**: benefit basketball game, study abroad program (not just "an
event")
   - **Product/item details**: vintage camera, brand new fire truck (not just "a camera")
   - **Numbers and quantities**: 4 years ago, next month, last week
   - **Company/organization names**: beverage company, fire-fighting brigade
   Additionally, **infer implied information** when clearly supported:
   - If multiple related items mentioned → may infer general pattern
   - Keep BOTH specific facts AND inferred insights as separate entries
3. Perform light contextual completion so that each fact is a clear standalone statement.
4. **Time Handling**:
   Note: Distinguish mention time (when said) vs event time (when happened).
   - For events with relative time (yesterday, last week, X ago, next month):
     Preserve the relative time and reference the message timestamp (YYYY-MM-DD).
     Format: "<fact with ALL details> <relative time> <timestamp>."
   - For ongoing/timeless facts: No time annotation needed.
```

Figure 4: Factual entry extraction prompt (Part 1).

Prompts for Factual Entry Extract

```
5. Output format:
   Always return a JSON object with key `"data"`, which is a list of items:
   {
     "source_id": "<source_id>",
     "fact": "<completed standalone fact with all specific details>"
   }
Examples:
--- Topic 1 ---
[2024-01-07T17:24:00.000, Sun] 0.Tim: Hey John! Next month I'm off to Ireland for a
semester in Galway
[2024-01-07T17:24:01.000, Sun] 1.John: That's awesome! Where will you stay?
[2024-01-07T17:24:02.000, Sun] 2.Tim: In Galway. I also want to visit The Cliffs of Moher
[2024-01-07T17:24:03.000, Sun] 3.John: Nice! By the way, I held a benefit basketball game
last week
[visual_context: a basketball court with players and audience]
[2024-01-07T17:24:04.000, Sun] 4.Tim: Cool! I'm currently reading "The Name of the Wind" by
Patrick Rothfuss
[2024-01-07T17:24:05.000, Sun] 5.John: That sounds interesting!
--- Topic 2 ---
[2024-01-12T13:41:00.000, Fri] 6.John: Got great news! I got an endorsement with a popular
beverage company last week
[2024-01-12T13:41:01.000, Fri] 7.Tim: Congrats! That's amazing
[2024-01-12T13:41:02.000, Fri] 8.John: Thanks! By the way, Barcelona is a must-visit city
[2024-01-12T13:41:03.000, Fri] 9.Tim: I'll add it to my list!
{"data": [
  {"source_id": 0, "fact": "Tim is going to Ireland for a semester in Galway the month
after 2024-01-07."},
  {"source_id": 0, "fact": "Tim will study in Galway, Ireland the month after 2024-01-07."},
  {"source_id": 2, "fact": "Tim will stay in Galway."},
  {"source_id": 2, "fact": "Tim wants to visit The Cliffs of Moher."},
  {"source_id": 3, "fact": "John held a benefit basketball game at a basketball court with
players and audience the week before 2024-01-07."},
  {"source_id": 4, "fact": "Tim is currently reading 'The Name of the Wind' by Patrick
Rothfuss."},
  {"source_id": 4, "fact": "Tim is reading a fantasy novel."},
  {"source_id": 6, "fact": "John got an endorsement with a beverage company the week before
2024-01-12."},
  {"source_id": 8, "fact": "John recommends Barcelona as a must-visit city."},
  {"source_id": 9, "fact": "Tim has a travel list and plans to add Barcelona to it."}
]}
Reminder: Be exhaustive and ALWAYS include specific names, titles, locations, and details
in every fact.
[USER]:
--- Topic {global_topic_id} ---
{topic_text}
```

Figure 5: Factual entry extraction prompt (Part 2).

Prompts for Relational Entry Extract

```
[SYSTEM]: You are a Relational Memory Extractor.
Your task is to extract **how people relate to each other** from conversations.
Note: Another system extracts factual content (what was said).
Your focus is on the **relational and emotional dynamics** between people.
The dialogue is organized into topic segments:
--- Topic X ---
[timestamp, weekday] <source_id>.<SpeakerName>: <message>
...
Note: Messages may include visual context marked as [visual_context: ...] which provides
additional scene information.
Important Instructions:
1. **Focus on Relational Behaviors and Emotional Exchange**:
   Extract interactions showing how people relate to each other:
   - Evaluative: praise, compliment, admire, acknowledge
   - Supportive: encourage, express confidence, cheer on, offer support
   - Emotional: express gratitude, pride, happiness, excitement, congratulations
   - Engagement: ask questions, show interest, respond with curiosity
   - Agreement: agree with, align on values, share perspective
   - Responsive: share in response to another's sharing, reciprocate
2. **What to Extract vs. What to Skip**:
   Extract: "Alice praised Bob's empathy" (relational behavior)
   Extract: "Alice asked about Bob's motivation" (engagement behavior)
   Extract: "Bob expressed gratitude for Alice's support" (emotional response)
   Skip: "Bob mentioned her support group experience" (factual content only)
   Skip: "Alice said she's been painting" (factual content only)
   BUT Extract: "Alice, in turn, shared her painting as a way of connecting" (responsive behavior)
3. **Include Necessary Context**:
   When describing interactions, include enough context to make sense.
   - Extract: "Alice praised Bob's dedication to helping LGBTQ youth"
   - Not just: "Alice praised Bob"
4. **Include Temporal Information When Relevant**:
   If the relational behavior involves time-specific events or references, include that naturally.
   - "Alice empathized with Bob's job search struggles by sharing her own experience from last
year"
   - "Bob congratulated Alice on her grad school acceptance"
   For general emotional exchanges without time context, no date needed.
5. **Combine Related Interactions**:
   Merge closely related behaviors in the same message.
   - "Alice congratulated Bob on passing the interviews and expressed excitement for her future"
6. **Use "both" for Mutual Agreement**:
   When both people express similar views or bond over shared experiences.
   - "Alice and Bob both emphasized the importance of self-care"
   - Assign to source_id where the second person completes the agreement
```

Figure 6: Relational entry extraction prompt (Part 1).

Prompts for Relational Entry Extract

```
Output format:
Return JSON with key "data", containing a list of:
{
  "source_id": "<source_id>",
  "relation": "<relational description in natural language>"
}
# EXAMPLE
--- Topic 1 ---
[2024-01-15T14:20:00.000, Mon] 0.Alice: I just got accepted to grad school!
[visual_context: a woman holding an acceptance letter and smiling]
[2024-01-15T14:20:02.000, Mon] 1.Bob: Oh nice
[2024-01-15T14:20:04.000, Mon] 2.Alice: Yeah, I'm really excited about the Computer Science program
[2024-01-15T14:20:06.000, Mon] 3.Bob: That's fantastic! I'm so proud of you. What's your research
focus?
[2024-01-15T14:20:08.000, Mon] 4.Alice: Machine learning. I've been working toward this for years.
[2024-01-15T14:20:10.000, Mon] 5.Bob: You totally deserve it. I know you'll do amazing things there.
--- Topic 2 ---
[2024-01-15T14:21:00.000, Mon] 6.Alice: Thanks! That means a lot. How's your job search going?
[2024-01-15T14:21:05.000, Mon] 7.Bob: Honestly, it's been tough. Feeling pretty discouraged.
[2024-01-15T14:21:10.000, Mon] 8.Alice: I totally get that. I went through the same thing last year.
[2024-01-15T14:21:15.000, Mon] 9.Bob: Really? How did you handle it?
[2024-01-15T14:21:20.000, Mon] 10.Alice: I focused on self-care and staying connected with friends.
[2024-01-15T14:21:25.000, Mon] 11.Bob: That's helpful advice. Thanks for sharing.
[2024-01-15T14:21:30.000, Mon] 12.Alice: Of course! You're going to land something great. Let me know
if you want to talk more.
[visual_context: two people having coffee and talking]
{"data": [
  {"source_id": "3", "relation": "Bob congratulated Alice on her grad school acceptance, expressed
pride in her achievement, and showed interest by asking about her research focus."},
  {"source_id": "5", "relation": "Bob validated Alice's deservingness and expressed confidence in her
future success."},
  {"source_id": "6", "relation": "Alice expressed gratitude for Bob's support and reciprocated by
showing interest in Bob's job search."},
  {"source_id": "8", "relation": "Alice empathized with Bob's difficulties by sharing her own similar
experience from last year."},
  {"source_id": "9", "relation": "Bob showed interest in Alice's coping strategies."},
  {"source_id": "11", "relation": "Bob expressed gratitude for Alice's advice."},
  {"source_id": "12", "relation": "Alice encouraged Bob and offered ongoing support."}
]}
Reminder: Focus on relational behaviors and emotional dynamics.
[USER]:
--- Topic {global_topic_id} ---
{topic_text}
```

Figure 7: Relational entry extraction prompt (Part 2).

Prompts for Synthesize

```
[SYSTEM]: You are a professional conversation summarization assistant with temporal awareness.
[USER]: You are a professional conversation summarization assistant.
The following conversation records contain TWO types of information:
1. **Factual information**: concrete events, plans, opinions, preferences
2. **Interaction patterns**: how speakers relate to, support, and respond to each other
Both types are important and should be preserved in the summary.
Conversation Time: {bucket}
Participants: {speakers}
Conversation Records:
{aggregated_text}
Related Temporal Context (from other time periods):
{supplementary_context}
Please generate a summary with the following requirements:
CRITICAL - What to PRESERVE:
- Specific concrete details: dates, times, locations, names of things
- Key emotional transitions and psychological changes
- Concrete action plans
- Important quotes or specific expressions when they capture essential meaning
- Temporal connections: When related context reveals specific prior events or future plans
  that directly relate to current topics, integrate them naturally with timestamps
What to DO:
1. Remove redundant repetitions while keeping all key information mentioned above
2. Organize content chronologically, showing how facts and interactions unfold together
3. Highlight causal relationships (e.g., "X happened, which gave Y the courage to do Z")
4. When integrating temporal context:
   - Cite specific times if available (e.g., "on 2022 April 15...")
   - Focus on concrete connections, not general patterns
   - Weave references naturally into the narrative, don't append them as separate summary
   - Only include if it adds meaningful context to understanding current events
5. Balance factual timeline with emotional/relational dynamics
6. Use fluent, concise natural language
7. Keep length between 200-350 words
Output the summary directly without any additional explanations or format markers.
```

Figure 8: Narrative synthesis prompt for Synthesis.

Prompts for Entity Extract

```
[SYSTEM]:
You are an entity extractor for conversational messages.
Input: ENTITY_TYPES, PREVIOUS_MESSAGES, CURRENT_MESSAGES (a list of entries).
Task: extract distinct entities explicitly or implicitly mentioned across the provided
CURRENT_MESSAGES.

Rules:
1) Speaker: for each entry, extract the speaker (before the colon) as an entity if present; merge
repeated mentions of the same speaker.
2) Only emit entities that appear in CURRENT_MESSAGES; use PREVIOUS_MESSAGES only for coreference
resolution.
3) Names: clear, unambiguous, lowercase_with_underscores.
4) Types: choose entity_type from ENTITY_TYPES, Default entity classification is Entity. Use this
entity type if the entity is not one of the other listed types..
5) Do not emit pronouns (you/I/he/she/they), times, dates, or pure actions/relations.

Output JSON ONLY:
{
    "entities": [
        {
            "id": <int>,                          // temporary id for this extraction round
            "name": "lower_snake_case",
            "entity_type": "string_from_ENTITY_TYPES",
            "aliases": ["..."],
            "first_entry_index": <int>         // index in CURRENT_MESSAGES where entity first
appears (0-based)
        }
    ]
}
[USER]:
"ENTITY_TYPES": {entity_types},
"PREVIOUS_MESSAGES": {previous_messages},
"CURRENT_MESSAGES": {messages},
```

Figure 9: Entity extraction prompt

Prompts for Entity Depuplicate

```
[SYSTEM]:
You decide whether a NEW ENTITY is a duplicate of any EXISTING ENTITIES. Use
PREVIOUS_MESSAGES/CURRENT_MESSAGES only for disambiguation.

Inputs include NEW_ENTITY (with id/name/entity_type/aliases/first_entry_index) and EXISTING_ENTITIES
(array with idx, name, entity_type, aliases).

Output JSON ONLY:
{
    "entity_resolutions": [
        {
            "id": <NEW_ENTITY.id>,
            "name": "best_full_name",
            "is_duplicate": true|false,
            "duplicate_idx": <int idx in EXISTING_ENTITIES or -1>,
            "duplicates": [idx1, idx2, ...]   // indices from EXISTING_ENTITIES, sorted, unique
        }
    ]
}

If unsure, set is_duplicate=false and duplicate_idx=-1 and duplicates=[].
[USER]:
"NEW_ENTITIES": [
    {
    "id": {id},
    "name": {name},
    "entity_type": {entity_type},
    "aliases": {aliases},
    "first_entry_index": {first_entry_index},
    }
],
"EXISTING_ENTITIES": {existing_entities},
"PREVIOUS_MESSAGES": {previous_messages},
"CURRENT_MESSAGES": {messages},
```

Figure 10: Entity deduplicate prompt

Prompts for Relation Extract

```
[SYSTEM]:
You extract fact triples between entities from CURRENT_MESSAGES (list of entries).

Inputs: FACT_TYPES, PREVIOUS_MESSAGES (for disambiguation only), CURRENT_MESSAGES, ENTITIES
(extracted or existing names).

Rules:
1) source and destination must be names from ENTITIES and must be distinct.
2) relationship in SCREAMING_SNAKE_CASE (e.g., GREETED, WORKS_AT). Prefer FACT_TYPES when applicable.
3) Remove duplicates/near-duplicates across the batch.

Output JSON ONLY:
{
    "facts": [
        {
            "source": "lower_snake_case",
            "relationship": "SCREAMING_SNAKE_CASE",
            "destination": "lower_snake_case",
            "fact": "concise paraphrase",
            "source_entry_index": <int>   // index of CURRENT_MESSAGES where this fact was observed
(0-based)
        }
    ]
}
[USER]:
"FACT_TYPES": {fact_types},
"PREVIOUS_MESSAGES": {previous_messages},
"CURRENT_MESSAGES": {messages},
"ENTITIES": {canonical_names},
```

Figure 11: Relation extraction prompt

Prompts for Relation Depuplicate

```
[SYSTEM]:
You deduplicate NEW FACTS (one or more) against EXISTING FACTS and detect contradictions against
FACT_INVALIDATION_CANDIDATES.

Output JSON ONLY:
{
    "result": {
        "duplicate_facts": [ { "new_fact_idx": <int>, "existing_fact_idx": <int>, "reason":
"string" }, ... ],
        "contradicted_facts": [ { "new_fact_idx": <int>, "existing_fact_idx": <int>, "reason":
"string" }, ... ],
        "fact_types": [ { "new_fact_idx": <int>, "fact_type": "TYPE_NAME" }, ... ]
    }
}

Notes:
- Indices for EXISTING_FACTS and FACT_INVALIDATION_CANDIDATES are 0-based and independent lists.
- duplicate_facts maps which new_fact (index in NEW FACTS array) duplicates which existing_fact index.
- If no duplicates/contradictions, return empty arrays.
[USER]:
"NEW_FACTS": {relations_raw},
"EXISTING_FACTS": {existing_facts},
"FACT_INVALIDATION_CANDIDATES": {fact_invalidation_candidates}
```

Figure 12: Relation deduplicate prompt

Prompts for StructMem QA

```
[SYSTEM]: You are an intelligent memory assistant tasked with retrieving accurate information
from conversation memories.
# CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain
timestamped information that may be relevant to answering the question.
# INSTRUCTIONS:
1. Carefully analyze all provided memories from both speakers
2. Pay special attention to the timestamps to determine the answer
3. If the question asks about a specific event or fact, look for direct evidence in the memories
4. If the memories contain contradictory information, prioritize the most recent memory
5. If there is a question about time references (like "last year", "two months ago", etc.),
   calculate the actual date based on the memory timestamp. For example, if a memory from
   4 May 2022 mentions "went to India last year," then the trip occurred in 2021.
6. Always convert relative time references to specific dates, months, or years. For example,
   convert "last year" to "2022" or "two months ago" to "March 2023" based on the memory
   timestamp. Ignore the reference while answering the question.
7. Focus only on the content of the memories from both speakers. Do not confuse character
   names mentioned in memories with the actual users who created those memories.
8. The answer should be less than 5-6 words.
# APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question
2. Examine the timestamps and content of these memories carefully
3. Look for explicit mentions of dates, times, locations, or events that answer the question
4. If the answer requires calculation (e.g., converting relative time references), show your work
5. Formulate a precise, concise answer based solely on the evidence in the memories
6. Double-check that your answer directly addresses the question asked
7. Ensure your final answer is specific and avoids vague time references
Memories for user {speaker_1_name}:
{speaker_1_memories}
Memories for user {speaker_2_name}:
{speaker_2_memories}
Session summaries:
{session_summaries}
Question: {question}
Answer:
```

Figure 13: Question answering prompt for StructMem system.

Prompts for Flat Memory system QA

```
[SYSTEM]: You are an intelligent memory assistant tasked with retrieving accurate information
from conversation memories.
# CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain
timestamped information that may be relevant to answering the question.
# INSTRUCTIONS:
1. Carefully analyze all provided memories from both speakers
2. Pay special attention to the timestamps to determine the answer
3. If the question asks about a specific event or fact, look for direct evidence in the memories
4. If the memories contain contradictory information, prioritize the most recent memory
5. If there is a question about time references (like "last year", "two months ago", etc.),
   calculate the actual date based on the memory timestamp. For example, if a memory from
   4 May 2022 mentions "went to India last year," then the trip occurred in 2021.
6. Always convert relative time references to specific dates, months, or years. For example,
   convert "last year" to "2022" or "two months ago" to "March 2023" based on the memory
   timestamp. Ignore the reference while answering the question.
7. Focus only on the content of the memories from both speakers. Do not confuse character
   names mentioned in memories with the actual users who created those memories.
8. The answer should be less than 5-6 words.
# APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question
2. Examine the timestamps and content of these memories carefully
3. Look for explicit mentions of dates, times, locations, or events that answer the question
4. If the answer requires calculation (e.g., converting relative time references), show your work
5. Formulate a precise, concise answer based solely on the evidence in the memories
6. Double-check that your answer directly addresses the question asked
7. Ensure your final answer is specific and avoids vague time references
Memories for user {speaker_1_name}:
{speaker_1_memories}
Memories for user {speaker_2_name}:
{speaker_2_memories}
Question: {question}
Answer:
```

Figure 14: Question answering prompt for flat memory systems.

Prompts for Graph Memory system QA

```
[SYSTEM]: You are an intelligent memory assistant tasked with retrieving accurate information
from conversation memories.
# CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain
timestamped information that may be relevant to answering the question. You also have access to
knowledge graph relations for each user, showing connections between entities, concepts, and
events relevant to that user.
# INSTRUCTIONS:
1. Carefully analyze all provided memories from both speakers
2. Pay special attention to the timestamps to determine the answer
3. If the question asks about a specific event or fact, look for direct evidence in the memories
4. If the memories contain contradictory information, prioritize the most recent memory
5. If there is a question about time references (like "last year", "two months ago", etc.),
calculate the actual date based on the memory timestamp. For example, if a memory from 4 May 2022
mentions "went to India last year," then the trip occurred in 2021.
6. Always convert relative time references to specific dates, months, or years. For example,
convert "last year" to "2022" or "two months ago" to "March 2023" based on the memory timestamp.
Ignore the reference while answering the question.
7. Focus only on the content of the memories from both speakers. Do not confuse character names
mentioned in memories with the actual users who created those memories.
8. The answer should be less than 5-6 words.
9. Use the knowledge graph relations to understand the user's knowledge network and identify
important relationships between entities in the user's world.
# APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question
2. Examine the timestamps and content of these memories carefully
3. Look for explicit mentions of dates, times, locations, or events that answer the question
4. If the answer requires calculation (e.g., converting relative time references), show your work
5. Analyze the knowledge graph relations to understand the user's knowledge context
6. Formulate a precise, concise answer based solely on the evidence in the memories
7. Double-check that your answer directly addresses the question asked
8. Ensure your final answer is specific and avoids vague time references
Memories for user {{speaker_1_user_id}}:
{{speaker_1_memories}}
Relations for user {{speaker_1_user_id}}:
{{speaker_1_graph_memories}}
Memories for user {{speaker_2_user_id}}:
{{speaker_2_memories}}
Relations for user {{speaker_2_user_id}}:
{{speaker_2_graph_memories}}
Question: {{question}}
Answer:
```

Figure 15: Question answering prompt for graph-based memory systems.

Prompts for Evaluating Answer

```
[USER]: Your task is to label an answer to a question as 'CORRECT' or 'WRONG'. You will be
given the following data:
    (1) a question (posed by one user to another user),
    (2) a 'gold' (ground truth) answer,
    (3) a generated answer
which you will score as CORRECT/WRONG.

The point of the question is to ask about something one user should know about the other user
based on their prior conversations.
The gold answer will usually be a concise and short answer that includes the referenced topic,
for example:
Question: Do you remember what I got the last time I went to Hawaii?
Gold answer: A shell necklace
The generated answer might be much longer, but you should be generous with your grading - as
long as it touches on the same topic as the gold answer, it should be counted as CORRECT.

For time related questions, the gold answer will be a specific date, month, year, etc. The
generated answer might be much longer or use relative time references (like "last Tuesday" or
"next month"), but you should be generous with your grading - as long as it refers to the
same date or time period as the gold answer, it should be counted as CORRECT. Even if the
format differs (e.g., "May 7th" vs "7 May"), consider it CORRECT if it's the same date.

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

First, provide a short (one sentence) explanation of your reasoning, then finish with CORRECT
or WRONG.
Do NOT include both CORRECT and WRONG in your response, or it will break the evaluation
script.

Just return the label CORRECT or WRONG in a json format with the key as "label".
```

Figure 16: Evaluate prompt for assessing response quality.

Prompts for Extraction Fidelity

```
[USER]: You are checking whether extracted entries contain hallucinated information.
Original Utterance:
{original_utterance}
Extracted Entries:
{all_entries}
**IMPORTANT CONTEXT:**
The original utterance may contain **image descriptions** in the format:
"image description : [description of image]"
These image descriptions are PART of the original content. If an extracted entry incorporates
information from an image description, this is NOT hallucination - it is accuratelyextracting
visual context provided in the conversation.
**Task:** Identify which entries (if any) contain **hallucinated information**
- information that is NOT present (including in image descriptions) or cannot be reasonably
inferred from the original utterance.
**What counts as HALLUCINATION:**
- Fabricated information with NO basis in the text or image descriptions
- Contradictions to what is stated
- Speculative additions of details not mentioned
**Output format:**

If NO entries contain hallucination:
{
  "has_hallucination": false,
  "hallucinated_entries": []
}

If some entries contain hallucination:
{
  "has_hallucination": true,
  "hallucinated_entries": [
    {
      "entry_index": 1,
      "entry_text": "the exact text of the hallucinated entry",
      "reason": "brief explanation of why this is hallucination"
    }
  ]
}
```

Figure 17: Prompt for assessing extraction fidelity.

Prompts for Consolidation Fidelity

```
[USER]: You are verifying whether cross-event links in a memory consolidation are spurious or
hallucinated.
## Context
During memory consolidation, the system attempts to establish **cross-event links** by
connecting current buffer events with retrieved historical events. Your task is to verify
whether these links are valid or spurious.
## Inputs
**Current Buffer Events:**
Events within the current time window being consolidated.
{buffer_text}
**Retrieved Historical Events (numbered):**
Past events retrieved as candidates for cross-event linking. Each is numbered.
{supplementary_text}
**Summary A (No cross-event links, seed_num=0):**
Consolidates ONLY buffer events. No historical events are linked.
{baseline_summary}
**Summary B (With cross-event links, seed_num=15):**
Consolidates buffer events AND establishes cross-event links by integrating retrieved
historical events.
{test_summary}
## Your Verification Task
**Step 1: Identify cross-event links**
Compare Summary B against Summary A. Find all content in Summary B that establishes connections
to historical events (content NOT present in Summary A).
For each cross-event link, identify which historical event(s) it references
(by number from the Retrieved Historical Events list).
**Step 2: Classify each cross-event link**
Determine whether each link is **valid** or **spurious/hallucinated**.
## Classification: SPURIOUS/HALLUCINATED Cross-Event Links
Mark as **spurious** if the link exhibits ANY of the following issues:
**1. Hallucinated event:**
- References a historical event that does NOT exist in the Retrieved Historical Events
**2. Spurious connection:**
- The historical event is real BUT has no meaningful relationship to buffer events
- Only superficial keyword overlap, no genuine topical/causal/contextual relevance
**3. Distorted historical event:**
- Historical event exists but is misrepresented in the link
**4. Misattributed event:**
- Historical event attributed to wrong person
```

Figure 18: Prompt for assessing consolidation fidelity (Part 1).

Prompts for Consolidation Fidelity

```
## Classification: VALID Cross-Event Links
Mark as **valid** if the link:
- Accurately represents a historical event from the Retrieved Historical Events list
- Establishes a genuine relationship to buffer events:
  - Same topic/domain
  - Same entities/people involved
  - Causal or temporal relationship
  - Provides meaningful context for understanding current events
- Reasonable paraphrasing is acceptable
- Minor timestamp approximations are acceptable
## Output Format
{{
  "incorporated_entries": [
    {{
      "supplementary_index": 3,
      "summary_b_text": "exact phrase from Summary B that establishes this cross-event link",
      "is_spurious": false,
      "reason": "brief explanation of why this is a valid cross-event link"
    }},
    {{
      "supplementary_index": 7,
      "summary_b_text": "exact phrase from Summary B",
      "is_spurious": true,
      "spurious_type": "hallucinated | spurious_connection | distorted | misattributed",
      "reason": "why this cross-event link is spurious"
    }}
  ]
}}
**Note:** If Summary B establishes NO cross-event links (identical to Summary A),
return an empty incorporated_entries list.
Be conservative: only mark as spurious when the cross-event link is clearly
problematic - either factually wrong or lacks genuine relevance to buffer events.
```

Figure 19: Prompt for assessing consolidation fidelity (Part 2).

Prompts for Unconstrained Synthesize

```
[SYSTEM]: You are a professional conversation summarization assistant with temporal awareness.
[USER]: You are a professional conversation summarization assistant.
The following conversation records contain TWO types of information:
1. **Factual information**: concrete events, plans, opinions, preferences
2. **Interaction patterns**: how speakers relate to, support, and respond to each other
Both types are important and should be preserved in the summary.
Conversation Time: {bucket}
Participants: {speakers}
Conversation Records:
{aggregated_text}
Related Temporal Context (from other time periods):
{supplementary_context}
Please generate a summary with the following requirements:
CRITICAL - What to PRESERVE:
- Specific concrete details: dates, times, locations, names of things
- Key emotional transitions and psychological changes
- Concrete action plans
- Important quotes or specific expressions when they capture essential meaning
- Temporal connections: When related context reveals specific prior events or future plans
  that directly relate to current topics, integrate them naturally with timestamps
What to DO:
1. Remove redundant repetitions while keeping all key information mentioned above
2. Organize content chronologically, showing how facts and interactions unfold together
3. Highlight causal relationships (e.g., "X happened, which gave Y the courage to do Z")
4. When integrating temporal context:
   - Cite specific times if available (e.g., "on 2022 April 15...")
   - Focus on concrete connections, not general patterns
   - Weave references naturally into the narrative, don't append them as separate summary
   - Only include if it adds meaningful context to understanding current events
5. Balance factual timeline with emotional/relational dynamics
6. Use fluent, concise natural language
7. Keep length between 200-350 words
Output the summary directly without any additional explanations or format markers.
```

Figure 20: Narrative synthesis prompt for Unconstrained Synthesis. The text highlighted in gray represents the grounding constraints that are active in our default *Constrained* setting but disabled for the *Unconstrained* variant to evaluate their impact on memory fidelity.