HIM: A Human-Inspired Memory Loop for LLM Agents via Encoding, Consolidation, and Retrieval

Haoze Tang ,Linqi Ye† ,Shaorong Xie†
School of Future Technology, Shanghai University, Shanghai, China
†Corresponding Authors: Linqi Ye (yelinqi@shu.edu.cn), Shaorong Xie (srxie@shu.edu.cn)

Abstract*— Large language model (LLM) agents often degrade in long-horizon, multi-session, and multi-speaker interactions due to noise accumulation, loss of source attribution, and interference from stale memories. Existing memory systems mainly emphasize storage capacity or retrieval effectiveness, while reliability remains underexplored, leading to attribution drift and unreliable retrieval in complex dialogues. We propose HIM, a human-inspired memory framework with a reliability-oriented lifecycle loop of Encoding, Consolidation, and Retrieval. HIM stores interactions as structured notes with explicit speaker attribution and contextual cues, regulates low-value entries through importance-aware retention, periodically reinforces useful memories via usage signals, and performs source-aware retrieval with an activation-style score for interpretability. Experiments on the LoCoMo benchmark show that HIM outperforms representative baselines in F1 and BLEU-1, with clear gains on adversarial, multi-hop, and temporal queries, demonstrating improved attribution fidelity and robustness in prolonged interactions.

***Index Terms*—LLM agents, long-term memory, agent memory, conversational memory , reliability**

I. INTRODUCTION

Large language model (LLM) agents are increasingly deployed as interactive systems that must maintain state and behave consistently through dialogue [1,2]. When such agents operate over long horizons—spanning multiple sessions and multiple speakers—the fixed context window becomes a fragile substrate for preserving attribution, preferences, and long-term commitments [3,4]. To cope with this limitation, many systems introduce explicit memory modules that store user preferences, episodic events, and task-relevant facts beyond the immediate prompt [5,6]. In practice, the core difficulty is not simply "storing more," but using memory reliably: the agent must preserve attribution (who said what), resist interference from outdated traces, and avoid retrieval noise as memory grows [7,8].

However, many existing agent memory designs prioritize capacity or retrieval effectiveness while leaving reliability underspecified. Source binding can be weakened during writing or compression [9], incidental content can accumulate without principled promotion or suppression [5], and retrieval may surface plausible but incorrect evidence under long time gaps or multi-speaker interactions [6]. These issues are amplified in long-horizon settings, motivating a lifecycle-aware memory loop that treats reliability as a first-class objective rather than an emergent byproduct [10].

To address these challenges, we propose HIM, a reliability-oriented memory framework inspired by human memory principles such as selective encoding, source monitoring, rehearsal-based consolidation, and activation-style retrieval(ACT-R–inspired availability) [11,12]. HIM organizes the agent's memory lifecycle into a feedback loop of encoding, consolidation, and retrieval, integrating lightweight signals such as source attribution, usage traces, and hierarchical retention levels to improve stability in long-horizon, multi-speaker interactions[13]. In summary, this paper contributes:

- A reliability-oriented HIM loop that operationalizes human-inspired principles (e.g., source monitoring, rehearsal-like consolidation, and activation-based retrieval) as simple and interpretable control signals for long-horizon conversational memory [11].
- A modular integration approach that layers HIM onto a structured note-based memory backbone with minimal changes to storage and retrieval interfaces [13].
- An empirical study on LoCoMo, including ablations and case analyses, demonstrates improved robustness and characterizes when reliability signals mitigate attribution errors and long-horizon interference [10].

II. RELATED WORK

A. Agent memory systems

A growing line of work equips LLM agents with external memory to persist information beyond the prompt. Early agent architectures emphasize interaction logs and reflection-style aggregation to support long-term coherence [2]. More recent systems formalize memory as a managed store with writing, updating, and forgetting policies, aiming to balance retention with noise control [5]. Other designs treat memory as an addressable substrate that the agent can query and update during tool-using loops, highlighting practical issues such as accumulation, staleness, and write-time errors [6]. In long-horizon dialogue, retrieval reliability becomes a bottleneck: multi-speaker settings require stronger provenance control and attribution preservation than "best-match" retrieval alone [10].

This work was supported by the Science and Technology Commission of Shanghai Municipality (24511103304).

B. Retrieval-Augmented and Long-Context Reasoning

Retrieval-augmented generation provides a general mechanism for grounding model outputs in external evidence, typically via dense retrieval followed by conditional generation [15,16]. In open-domain QA and long-context aggregation, retrieval-plus-generation pipelines show that multi-evidence composition is often necessary for correctness, but performance depends critically on selecting stable and relevant context [17]. In parallel, long-context modeling extends attention mechanisms to carry information across longer sequences, but remains bounded by compute/memory constraints and does not by itself solve multi-session persistence or attribution drift [3,4]. For agentic dialogue memory, specialized benchmarks such as LoCoMo stress exactly these failure modes—long-range dependencies, multi-session continuity, and robustness to misleading queries—making it natural to combine retrieval with explicit lifecycle control [14].

C. Human Memory–Inspired Mechanisms

A separate thread draws inspiration from cognitive science to motivate how memory should be written, strengthened, and recalled. Source monitoring explains how humans bind content to its origin and why attribution failures occur under interference [11]. Selective encoding and executive control (often associated with prefrontal mechanisms) motivate gating and prioritization under limited capacity [12]. Complementary learning systems and rehearsal-based consolidation motivate why repeated use stabilizes traces while incidental details decay [18,19]. Finally, ACT-R models retrieval as cue-dependent availability shaped by usage history, offering a principled lens for activation-style ranking [20].

These principles broadly motivate HIM's design choices: selective encoding with explicit provenance, periodic maintenance that reinforces repeatedly useful traces, and activation-style ranking that combines semantic match with stability signals. Overall, they provide a cognitive route to improving reliability without heavy re-training or replacing the retrieval backbone.

III. Methodology

A. Overview

We study long-term conversational agents operating over long-horizon, multi-session, and multi-speaker interactions, where failures commonly stem from noisy memory accumulation, confusion of who said what, and retrieval that over-exposes stale or weakly supported evidence. To address these issues, we propose a Human-Inspired Memory (HIM) Loop that organizes memory into three coupled stages: Encoding → Consolidation → Retrieval. The design follows a note-centric organization with explicit links and augments the memory lifecycle with source attribution, importance-aware retention, and usage traces.

Formally, the agent maintains a memory collection $\mathcal{M} = \{m_1, \dots, m_N\}$and a link graph $G = (\mathcal{M}, E)$,where edges in E represent semantic or contextual relationships among memory items. The HIM Loop updates $\mathcal{M}$ online, periodically consolidates memory to stabilize structure, and retrieves source-aware evidence to support answer generation.

B. Memory Encoding

In biological cognition, executive control and prefrontal filtering are commonly associated with selective encoding: not all sensory input is committed to stable memory, and attribution cues are bound early to support later source judgments. HIM instantiates this principle via two coupled operations at write time: source monitoring (binding content to its producer) and importance gating (filtering and hierarchical retention), which together reduce noise accumulation and attribution drift in multi-speaker settings.

1) Structured Note Construction

Given an interaction instance i, we encapsulate it as a

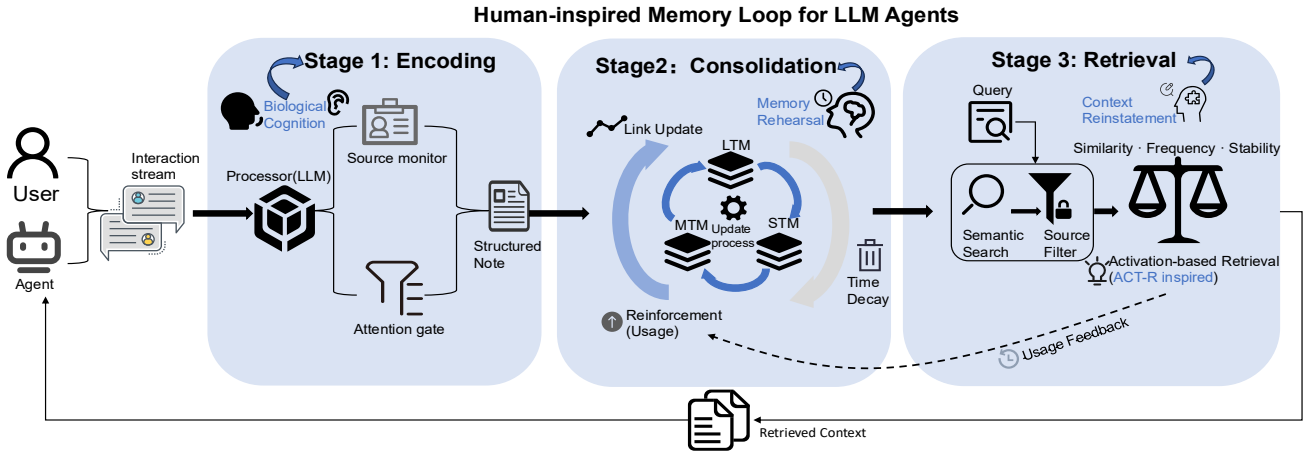

Fig. 1. Human-Inspired Memory (HIM) Loop for LLM Agents. HIM organizes agent memory into a three-stage loop—encoding, consolidation, and retrieval—inspired by human memory principles. Stage 1 (Encoding) performs selective attention and source monitoring, distilling salient information from the interaction stream into structured notes with provenance cues. Stage 2 (Consolidation) mimics rehearsal-based consolidation, updating memories under reinforcement and time decay. Stage 3 (Retrieval) follows context reinstatement and cue-dependent retrieval and ranks candidates via an ACT-R–inspired activation-based mechanism. Retrieved context supports response generation, while usage feedback continually refines memory.

structured memory item m_i that stores both semantic content and explicit cognitive metadata:

$$m_i = \{c_i, t_i, K_i, G_i, X_i, s_i, I_i, \ell_i, L_i, r_i, a_i, e_i\}, m_i \in \mathcal{M}. \quad (1)$$

Here c_i denotes the interaction content (e.g., a dialogue utterance or an external observation), and t_i denotes its timestamp. To enrich memory beyond surface text, an LLM infers semantic components: keywords K_i, categorical tags G_i, and a contextual description X_i that captures implicit meaning and disambiguating context. Reliability-related metadata includes source attribution s_i (speaker/role), an importance score $I_i \in [0,1]$, a discrete memory level $\ell_i \in \{\text{STM}, \text{MTM}, \text{LTM}\}$ (Short-Term Memory, Medium-Term Memory, Long-Term Memory), and maintenance statistics such as retrieval count r_i and last access time a_i. The set L_i stores explicit links to other memory items. Finally, $e_i \in \mathbb{R}^d$ is a dense embedding used for efficient similarity search.

We generate semantic and reliability fields using a structured prompt template P_{enc}:

$$K_i, G_i, X_i, I_i \leftarrow \text{LLM}(c_i \parallel t_i \parallel P_{enc}), \quad (2)$$

where $\parallel$ denotes concatenation. Source attribution s_i is obtained from the conversation context and stored explicitly to preserve the speaker and contents.

2) Importance Gating and Retention Levels

Not all perceived inputs warrant stable storage. To model selective retention, HIM assigns each memory to a hierarchical retention level based on its importance score I_i:

$$\ell_i = \begin{cases} \text{STM}, & I_i < \tau_1, \\ \text{MTM}, & \tau_1 \le I_i < \tau_2, \\ \text{LTM}, & I_i \ge \tau_2, \end{cases} \quad (3)$$

where τ_1 and τ_2 are cognitive thresholds (0.3 and 0.7). This gating mechanism acts as an explicit capacity regulator: low-salience traces remain in STM, moderately salient traces are retained in MTM, and highly salient traces enter LTM, thereby limiting the growth of noisy memory while preserving high-value evidence for long-horizon retrieval.

3) Embedding and Associative Linking

We compute a dense representation that integrates multiple textual facets of the structured note:

$$e_i = f_{enc}(\text{concat}(c_i, K_i, G_i, X_i)), \quad (4)$$

where f_{enc} is a sentence embedding model. Beyond vector storage, HIM also forms explicit links between a new memory and a small set of related existing memories to support associative organization and multi-hop evidence discovery. Concretely, we retrieve a neighbor candidate set $N_i \subset \mathcal{M} \setminus \{m_i\}$ via embedding similarity and/or tag overlap, select a subset to connect as links L_i, and optionally refine local associations. These links serve as stable relational pointers that can be exploited during retrieval to recover supporting context beyond top-K similarity matches.

C. Dynamic Memory Consolidation

Human memory is not a static archive; stability is shaped by consolidation, where repeatedly reactivated traces become more robust while incidental traces fade. HIM operationalizes consolidation as a periodic background maintenance process that strengthens memories based on usage evidence and stabilizes the memory network structure. Importantly, this stage closes the loop with retrieval: Stage 3 logs lightweight usage feedback (e.g., access counts, last access time, and availability signals) during cue-dependent retrieval, and Stage 2 consumes these signals to reinforce repeatedly useful traces and suppress drifting ones. This design mitigates long-horizon drift and interference that emerge as memory grows.

1) Frequency-driven Reinforcement

During retrieval, HIM records per-memory access statistics (retrieval count and recency), which serve as the usage feedback signal for consolidation.Let r_i denote the retrieval count of memory m_i. At a fixed interval, HIM updates importance using a monotone function of usage:

$$I_i \leftarrow \text{clip}(I_i + \eta \cdot g(r_i), 0, 1), \quad (5)$$

where $\eta > 0$is a step size and $g(\cdot)$can be chosen as $g(r) = \log(1 + r)$. The logarithmic form captures diminishing returns: early reactivations yield the largest strengthening effect. After updating I_i, the level l_i is re-evaluated using Eq. (3), enabling promotion from STM to MTM/LTM when a memory becomes repeatedly useful. This provides a computationally lightweight approximation to rehearsal-dependent stabilization, prioritizing traces that consistently support successful retrieval.

As I_i evolves, memories can transition across retention levels. In particular, traces initially stored in STM/MTM may cross thresholds τ_1, τ_2 and become more stable, addressing a common failure in long dialogues where early but important facts are drowned out by later noise. This dynamic promotion also supports adaptability: the system does not require perfect importance estimation at the moment of writing, since repeated utility can correct initial underestimation through consolidation.

2) Structural Stabilization in the Memory Graph

Consolidation in HIM also targets the structure of the memory network $G = (\mathcal{M}, E)$. To counter long-term drift and fragmentation, we optionally strengthen associations among frequently co-activated memories. Let w_{ij}denote an optional edge weight between m_i and m_j. A simple stabilization rule is:

$$w_{ij} \leftarrow w_{ij} + \gamma \cdot h(\kappa(i,j)), \quad (6)$$

where $\kappa(i,j)$ counts how often m_i and m_j are retrieved together, γ is a small update coefficient and $h(\cdot)$ is a monotone mapping. This encourages semantically aligned or jointly useful traces to become more tightly connected,

yielding a memory space that remains coherent under long interaction horizons and improving the reliability of multi-hop retrieval through stable associations.

D. Retrieval and Activation-aware Analysis

Conventional retrieval augmented systems often rely solely on static similarity matching. However, in long-horizon conversational settings, similarity alone is insufficient: semantically plausible but unreliable traces (e.g., from a different speaker or an outdated context) can dominate retrieval. HIM therefore adopts a cue-dependent retrieval pipeline where candidate generation remains similarity-first for efficiency, while reliability is controlled by (i) source cues injected at encoding and carried through retrieval to reduce cross-speaker interference, and (ii) an activation-style availability signal that integrates usage and retention states. Besides supporting interpretability, the activation signal is logged as usage feedback that drives Stage 2 consolidation updates, closing the lifecycle loop across sessions. In our setting, context reinstatement is implemented by enriching the retrieval cue with lightweight contextual fields from the structured notes and the current goal/query, so that retrieval is driven by reinstated conversational context rather than surface-form match alone.

Given a query q, we first retrieve a candidate set by dense similarity. Let $e_q = f_{enc}(q)$; we select:

$$\mathcal{C}(q) = \text{Top-K}(\{\text{sim}(e_q, e_i)\}_{m_i \in \mathcal{M}}), \tag{7}$$

where $\text{sim}(\cdot,\cdot)$is cosine similarity and $C(q) \subset M$. To support multi-hop evidence, we optionally expand candidates along explicit links:

$$\mathcal{C}^+(q) = \mathcal{C}(q) \cup \bigcup_{m_i \in \mathcal{C}(q)} L_i, \tag{8}$$

To quantify which traces become most accessible under the current cue, we compute an activation-style score for each candidate $m_i \in C^+(q)$:

$$A(m_i, q) = \alpha\, \text{s}(\text{rank}_i) + \beta\, \phi(r_i) + \delta\, I_i + \epsilon\, \psi(\ell_i), \tag{9}$$

where $\phi(r_i)$maps retrieval frequency to a bounded score (in our implementation, $\phi(r) = min\{1, log\,(1+r)\}$), and $\psi(\ell_i)$ encodes hierarchical retention with $\psi(\text{STM}) < \psi(\text{MTM}) < \psi(\text{LTM})$ (we instantiate $\psi(\text{STM}) = 0.2$, $\psi(\text{MTM}) = 0.5$, and $\psi(\text{LTM}) = 1.0$).The coefficients $\alpha, \beta, \delta, \epsilon \geq 0$control contributions from semantic match, usage history, intrinsic importance, and retention state. In all experiments, we fix them to the default configuration in our implementation: $\alpha = 0.70$, $\beta = 0.10$, $\delta = 0.10$, and $\epsilon = 0.05$. Following the implementation, the "semantic match" term is computed as a rank-based proxy over candidates returned by Eq.(7), $s(\text{rank}_i) = \exp(-0.1 \cdot \text{rank}_i)$, rather than re-computing a separate cosine score inside Eq. (9). This score provides a principled lens for case studies and error diagnosis by exposing how usage and lifecycle controls shape accessibility ; meanwhile, it also serves as a lightweight usage feedback signal (availability) that is accumulated during retrieval and consumed by periodic consolidation, while preserving Eq.(7) as the primary backbone for candidate selection and efficiency.

Finally, retrieved memories are presented with explicit source attribution s_i (e.g., speaker prefixes) and the most relevant textual fields (typically c_i and X_i). This preserves traceability and reduces cross-speaker contamination in multi-party dialogues, especially when multiple speakers mention semantically similar facts (preferences, plans, or entities). By carrying source cues injected at encoding through retrieval, HIM improves robustness long contexts where spurious yet plausible traces would otherwise be amplified.

IV. EXPERIMENTS

A. Setup

We evaluate on LoCoMo, a long-horizon conversational memory benchmark with five query categories: Multi-hop, Temporal, Open-domain, Single-hop, and Adversarial. Unless otherwise stated, all methods use the same backbone model (Qwen2.5-3B) and evaluation protocol. We report F1 and BLEU-1 for the main comparison, and ROUGE-L for ablations.

We implement HIM as a lightweight wrapper over a note-centric memory backbone. Each dialogue utterance is encoded into a structured note with explicit source attribution and metadata fields, and retrieval is performed primarily by dense similarity over sentence embeddings (default: all-MiniLM-L6-v2). We use a fixed retention gating with two thresholds τ_1 =0.3 and τ_2 =0.7) to map importance into

TABLE I: Main results on LoCoMo with Qwen2.5-3B. Each cell reports F1 / BLEU-1 (higher is better) across five query categories(Multi-hop, Temporal, Open-domain, Single-hop, and Adversarial).

Method	Multi-hop F1 / BLEU-1	Temporal F1 / BLEU-1	Open-domain F1 / BLEU-1	Single-hop F1 / BLEU-1	Adversarial F1 / BLEU-1
LoCoMo	4.61 / 4.29	3.11 / 2.71	4.55 / 5.97	7.03 / 5.69	16.95 / 14.81
ReadAgent	2.47 / 1.78	3.01 / 3.01	5.57 / 5.22	3.25 / 2.51	15.78 / 14.01
MemoryBank	3.60 / 3.39	1.72 / 1.97	6.63 / 6.58	4.11 / 3.32	13.07 / 10.30
MemGPT	5.07 / 4.31	2.94 / 2.95	7.04 / 7.10	7.26 / 5.52	14.47 / 12.39
A-MEM	13.51 / 9.22	22.58 / **18.60**	**8.93** / 8.08	21.62 / 16.96	34.84 / 31.69
HIM (Ours)	**14.21 / 9.91**	**24.64** / 18.33	5.45 / **9.06**	**22.32 / 18.18**	**39.98 / 37.23**

STM/MTM/LTM levels, and run consolidation periodically to promote repeatedly useful memories based on access statistics. We keep these thresholds fixed across all query categories and do not tune them per category, to avoid introducing additional degrees of freedom and to keep comparisons fair under the same evaluation protocol.

B. Baselines

We compare against representative long-context and external-memory baselines, including ReadAgent, MemoryBank, MemGPT, and A-MEM. Our method (HIM) follows the structured note construction and memory evolution paradigm, and further strengthens reliability via (i) source-aware selective writing and (ii) usage-driven consolidation. The activation score is used for interpretability and error diagnosis，and is also logged as a lightweight usage feedback signal for consolidation. Under these controls, performance differences primarily reflect how each methodwrites, organizes, and updates memory rather than differences in retrieval budget or sampling noise.

C. Main Results

TABLE I reports the main results with Qwen2.5-3B. Overall, memory systems that adopt structured note construction and explicit memory maintenance substantially outperform earlier baselines across most categories, supporting the view that long-horizon conversational memory benefits from explicit organization beyond naive context reading or flat storage.

Compared with the strongest note-based baseline (A-MEM), HIM improves F1 on four out of five categories. The largest gain is on Adversarial queries (39.98 vs. 34.84, +5.14 F1), where misleading cues and competing traces make reliability control more critical. HIM also improves Temporal (24.64 vs. 22.58, +2.06 F1) and Multi-hop (14.21 vs. 13.51, +0.70 F1), consistent with the goal of stabilizing retrieval under time-sensitive constraints and multi-step dependencies.

We observe a clear trade-off on Open-domain queries: HIM attains higher BLEU-1 (9.06) but yields lower F1 (5.45) than the strongest baseline (8.93 F1). This suggests that reliability signals can change evidence selection and response formulation in a way that improves surface-level lexical overlap, while not consistently increasing answer overlap under the F1 metric for open-domain questions. For completeness, Overall is reported as the macro-average over the five query categories in TABLE I.

D. Ablation Study

TABLE II ablates two key components while keeping the backbone model and evaluation protocol fixed: w/o Encoding removes source-aware selective writing (and the associated structured-note encoding signals), and w/o Consolidation disables usage-driven promotion and periodic consolidation updates. Each cell reports F1 / BLEU-1 / ROUGE-L (higher is better).

Removing either component degrades performance, but the impact differs by category. On Adversarial queries, consolidation is the dominant factor: removing consolidation causes a large drop (39.98 → 22.37 F1, −17.61), while removing encoding produces a smaller but still substantial reduction (39.98 → 33.14 F1, −6.84). This indicates that maintenance over time—promoting repeatedly useful traces and suppressing unstable ones—is critical for robustness under interference and competing evidence.

On Temporal queries, disabling consolidation also produces a pronounced drop (24.64 → 16.69 F1, −7.95), whereas removing encoding leads to a milder reduction (24.64 → 22.61 F1, −2.03). This supports the role of consolidation in stabilizing time-sensitive retrieval across long interaction horizons. On Multi-hop, both components matter, with encoding contributing more to multi-step dependency tracking (14.21 → 10.64 F1, −3.57) than consolidation (14.21 → 12.40 F1, −1.81).

Notably, BLEU-1 and ROUGE-L do not always track F1 monotonically (e.g., on Temporal, w/o Consolidation increases BLEU-1/ROUGE-L while reducing F1), suggesting that removing maintenance can lead to responses with higher surface overlap but lower correctness under the F1 criterion. To avoid ambiguity in aggregation, we report Multi-hop, Temporal, and Adversarial for ablations: these categories best cover multi-step dependency, time-sensitive retrieval, and robustness against misleading or competing traces, and they are also explicitly included in Table I.

E. Case Study

Fig. 2 illustrates how HIM performs memory encoding, organization, and retrieval in a multi-party dialogue. As shown in Fig. 2(a), each utterance is converted into a structured memory note with explicit speaker identity, textual content, semantic cues, and an importance score. This preserves speaker attribution at write time and reduces attribution errors in long-horizon interactions. Over time, these notes form a memory network (Fig. 2(b)), where nodes denote memories

TABLE II: Ablation results of HIM on LoCoMo under a fixed backbone and evaluation protocol. w/o Encoding removes source-aware selective writing and the corresponding structured-note encoding signals; w/o Consolidation disables usage-driven promotion and periodic consolidation updates. Results are reported on Multi-hop, Temporal, and Adversarial categories, with each cell showing F1 / BLEU-1 (%) / ROUGE-L (%) (higher is better).

Method	Multi-hop			Temporal			Adversarial		
	F1	BLEU-1	ROUGE-L	F1	BLEU-1	ROUGE-L	F1	BLEU-1	ROUGE-L
HIM w/o Encoding	10.64	8.75	7.96	22.61	15.74	22.15	33.14	31.16	33.05
HIM w/o Consolidation	12.40	10.77	10.28	22.37	16.69	22.89	30.88	27.64	31.33
HIM (Full)	**14.21**	**9.91**	**15.22**	**24.64**	**18.33**	**25.05**	**39.98**	**37.23**	**40.17**

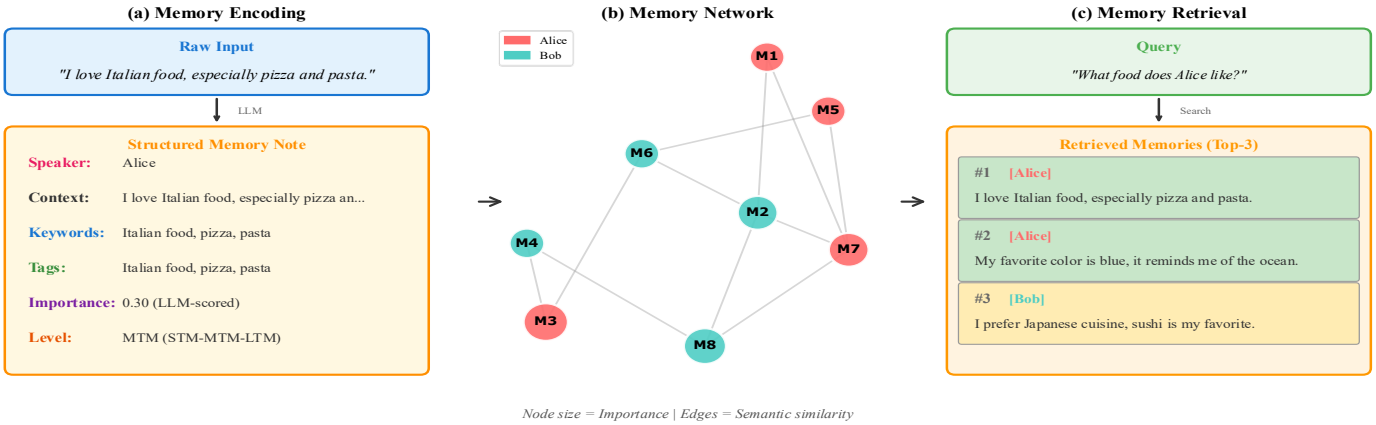

Fig. 2. Case study of HIM over a multi-party dialogue. (a) Memory Encoding: an utterance is converted into a structured memory note with explicit speaker attribution and semantic cues (keywords/tags), with an importance score for lifecycle control. (b) Memory Network: stored notes form a graph organized by semantic relatedness (node size reflects importance). (c) Memory Retrieval: given a query, HIM retrieves candidates by dense similarity, presents evidence with source information to reduce cross-speaker interference, and logs access statistics as usage feedback for periodic consolidation.

and edges capture semantic relatedness, helping reduce interference and support periodic consolidation through retrieval-time usage feedback.

Given a query such as "What food does Alice like?", HIM retrieves candidates by dense similarity (Fig. 2(c)) and returns the top memories with their recorded sources. In this example, the correct evidence is Alice's statement ("I love Italian food, especially pizza and pasta"), while distracting memories from other speakers may also appear. With explicit source attribution, HIM reduces cross-speaker contamination and produces a grounded response.

V. CONCLUSION

This work presents HIM, a human-inspired framework for improving reliability in LLM agent memory for long-horizon, multi-session, and multi-speaker interactions. By integrating source attribution, usage-driven consolidation, and an activation-style availability signal, HIM reduces noise accumulation, attribution drift, and interference from stale traces. Experiments on LoCoMo show consistent gains, especially on adversarial and multi-hop queries, suggesting that lifecycle-aware reliability is a practical and interpretable extension to memory-augmented agents. Future work will explore multimodal settings, domain shifts, and efficient scaling to larger memory stores.

REFERENCES

[1] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: synergizing reasoning and acting in language models," in Proc. Int. Conf. Learn. Representations (ICLR), 2023.

[2] J. S. Park, J. C. O'Brien, C. J. Cai, M. Ringel Morris, P. Liang, and M. S. Bernstein, "Generative agents: interactive simulacra of human behavior," in Proc. ACM Symp. User Interface Software and Technology (UIST), 2023, doi: 10.1145/3586183.3606763.

[3] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: attentive language models beyond a fixed-length context," in Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL), pp. 2978–2988, 2019.

[4] M. Zaheer et al., "Big Bird: transformers for longer sequences," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[5] W. Zhong, L. Guo, Q. Gao, Y. He, and Y. Wang, "MemoryBank: enhancing large language models with long-term memory," in Proc. AAAI Conf. Artificial Intelligence, vol. 38, no. 17, pp. 19724–19731, 2024, doi: 10.1609/aaai.v38i17.29946.

[6] C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez, "MemGPT: towards LLMs as operating systems," arXiv:2310.08560, 2023.

[7] W. Xu, Z. Liang, K. Mei, H. Gao, J. Tan, and Y. Zhang, "A-MEM: agentic memory for LLM agents," arXiv:2502.12110, 2025.

[8] K.-H. Lee, X. Chen, H. Furuta, J. Canny, and I. Fischer, "A human-inspired reading agent with gist memory of very long contexts," in Proc. 41st Int. Conf. Mach. Learn. (ICML), Proc. Mach. Learn. Res., vol. 235, pp. 26396–26415, 2024.

[9] S. Sukhbaatar, A. Szlam, J. Weston, and R. Fergus, "End-to-end memory networks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), pp. 2440–2448, 2015.

[10] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang, "Evaluating very long-term conversational memory of LLM agents," in Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics (ACL) (Vol. 1: Long Papers), pp. 13851–13870, 2024, doi: 10.18653/v1/2024.acl-long.747.

[11] M. K. Johnson, S. Hashtroudi, and D. S. Lindsay, "Source monitoring," Psychol. Bull., vol. 114, no. 1, pp. 3–28, 1993.

[12] E. K. Miller and J. D. Cohen, "An integrative theory of prefrontal cortex function," Annu. Rev. Neurosci., vol. 24, pp. 167–202, 2001.

[13] S. Ahrens, *How to Take Smart Notes: One Simple Technique to Boost Writing, Learning and Thinking*. North Charleston, SC, USA: CreateSpace Independent Publishing Platform, 2017.

[14] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," in Proc. 39th Int. Conf. Mach. Learn. (ICML), Proc. Mach. Learn. Res., vol. 162, pp. 2206–2240, 2022.

[15] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), 2020.

[16] V. Karpukhin et al., "Dense passage retrieval for open-domain question answering," in Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP), pp. 6769–6781, 2020.

[17] G. Izacard and E. Grave, "Leveraging passage retrieval with generative models for open domain question answering," in Proc. 16th Conf. European Chapter Assoc. Comput. Linguistics (EACL), pp. 874–880, 2021, doi: 10.18653/v1/2021.eacl-main.74.

[18] J. L. McClelland, B. L. McNaughton, and D. C. O'Reilly, "Why there are complementary learning systems in the hippocampus and neocortex," Psychol. Rev., vol. 102, no. 3, pp. 419–457, 1995.

[19] S. Diekelmann and J. Born, "The memory function of sleep," Nat. Rev. Neurosci., vol. 11, no. 2, pp. 114–126, 2010.

[20] J. R. Anderson and C. Lebiere, *The Atomic Components of Thought*. Mahwah, NJ, USA: Lawrence Erlbaum Associates, 1998.