Back to Basics: Let Conversational Agents Remember with Just Retrieval and Generation

Yuqian Wu[1,*], Wei Chen[1,2,*,†], Zhengjun Huang[2], Junle Chen[2], Qingxiang Liu[1], Kai Wang[4], Xiaofang Zhou[2,3], Yuxuan Liang[1,†]

The Hong Kong University of Science and Technology (Guangzhou)[1]
The Hong Kong University of Science and Technology[2]
Guangzhou HKUST Fok Ying Tung Research Institute[3], National University of Singapore[4]

ywu188@connect.hkust-gz.edu.cn, †*onedeanxxx@gmail.com*, †*yuxliang@outlook.com*
Code available at: `https://github.com/yuqian2003/Nano-Memory`

Abstract

Existing conversational memory systems rely on complex hierarchical summarization or reinforcement learning to manage long-term dialogue history, yet remain vulnerable to context dilution as conversations grow. In this work, we offer a different perspective: the primary bottleneck may lie not in memory architecture, but in the *Signal Sparsity Effect* within the latent knowledge manifold. Through controlled experiments, we identify two key phenomena: *Decisive Evidence Sparsity*, where relevant signals become increasingly isolated with longer sessions, leading to sharp degradation in aggregation-based methods; and *Dual-Level Redundancy*, where both inter-session interference and intra-session conversational filler introduce large amounts of non-informative content, hindering effective generation. Motivated by these insights, we propose `Nano-Memory`, a minimalist framework that brings conversational memory back to basics, relying solely on retrieval and generation via Turn Isolation Retrieval (TIR) and Query-Driven Pruning (QDP). TIR replaces global aggregation with a max-activation strategy to capture turn-level signals, while QDP removes redundant sessions and conversational filler to construct a compact, high-density evidence set. Extensive experiments on multiple benchmarks demonstrate that `Nano-Memory` achieves robust performance across diverse settings, consistently outperforming strong baselines while maintaining high efficiency in tokens and latency, establishing a new minimalist baseline for conversational memory.

1 Introduction

Large language models enable conversational agents capable of long-term user interaction. A key requirement for such agents is the ability to remember past interactions and use them to inform future responses. This capability, often referred to as conversational memory (Hu et al., 2025b), has become a fundamental component of personalized dialogue systems.

Existing research explores two primary paradigms for implementing conversational memory, as illustrated in Fig. 1. One paradigm, implicit memory, integrates memory directly into the parameters or internal states of the language model. Representative approaches include updating full model weights (Liu et al., 2024), maintaining latent states (Zhang et al., 2025b), introducing dedicated memory layers (Berges et al., 2024), or relying on KV-cache (Eyuboglu et al., 2025; Zweiger et al., 2026) based persistence. These methods typically depend on continual learning mechanisms to encode user history into model parameters. Although promising, such approaches usually require additional training procedures, incur significant computational overhead, and may suffer from issues such as catastrophic forgetting or training instability (Kemker et al., 2018) inherent in neural networks.

[*]Equal contribution. [†]Corresponding authors.

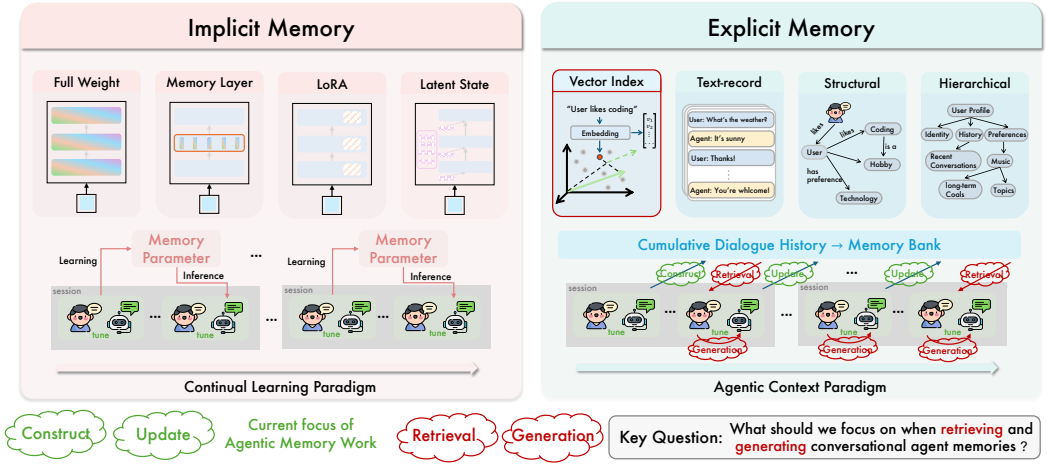


Figure 1: Different types of memory methods for personalized conversational agents

As a result, another paradigm known as explicit memory remains the de facto solution in practice (Gupta, 2025a;b). In this setting, dialogue history is externalized into a memory bank that can be retrieved during inference. Recent studies (categorized in Appendix A.2, Tab. 6) have proposed increasingly sophisticated mechanisms for memory construction, organization, and updating, including vector indices, structured memories, hierarchical profiles, and text-based records. While these designs the usability of memory management, they also introduce considerable system complexity (Anthropic, 2025; Lopopolo, 2026).

However, despite the growing sophistication of recent explicit memory designs, a fundamental question remains underexplored: *what truly decisive evidence should both retrieval and generation focus on when serving as the core primitives for conversational memory?*

In this work, we provide a complementary perspective by examining how useful signals are distributed and utilized in conversational data. Our analysis reveals a *Signal Sparsity Effect* in the latent knowledge manifold: only a small subset of turns contains evidence relevant to a query, while the majority of content is redundant. As conversations grow longer, this imbalance becomes increasingly severe, posing challenges to both retrieval and generation.

Through controlled experiments, we identify two key phenomena. First, *Decisive Evidence Sparsity*: relevant signals become increasingly isolated as session length increases, making them difficult to capture with conventional aggregation-based retrieval. Second, *Dual-Level Redundancy*: redundancy arises at both the inter-session level, where irrelevant sessions introduce interference, and the intra-session level, where conversational filler dominates most turns and dilutes useful information. Together, these significantly effects degrade both retrieval precision and generation quality. These findings highlight a simple yet overlooked principle: when using retrieval and generation as the foundation of conversational memory, the key is to *accurately localize decisive signals and eliminate redundant context*.

Based on this insight, we back to the basics and propose `Nano-Memory`, a minimalist framework built solely upon retrieval and generation. Specifically, we introduce *Turn Isolation Retrieval* (TIR), which replaces coarse aggregation with a max-activation strategy to capture turn-level signals, and *Query-Driven Pruning* (QDP), which removes redundant sessions and intra-session conversational filler to construct a compact, high-density evidence set for generation. Our contributions are summarized as follows:

- We identify the *Signal Sparsity Effect* as a fundamental bottleneck in conversational memory and characterize it through two key phenomena: Decisive Evidence Sparsity and Dual-Level Redundancy.
- We propose `Nano-Memory`, a minimalist framework that returns conversational memory to the basic paradigm of retrieval and generation, instantiated via TIR and QDP.
- We demonstrate through extensive experiments that a simple, well-designed retrieval-generation pipeline can outperform more complex memory systems, establishing a new minimalist baseline for long-term conversational memory.

2 Preliminary

Definition 1. *(Conversational Interaction)*. A conversational agent interacts with a user through a sequence of dialogue sessions: $\mathcal{H} = \{c_i\}_{i=1}^{C}$, where C denotes the number of sessions in the dialogue history. Each session c_i is composed of a sequence of interaction turns: $c_i = \{t_j\}_{j=1}^{T_i}$, where T_i is the number of turns in session i. Each turn is defined as a pair: $t_j = (u_j, r_j)$, where u_j denotes the user request r_j denotes the agent response. The complete interaction history $\mathcal{H}$ therefore contains all previously observed user–agent interactions accumulated across sessions.

Definition 2. *(Explicit Conversational Memory)*. Given a dialogue history $\mathcal{H}$, an explicit conversational memory is an external memory bank constructed from historical interactions without modifying the parameters of the response model. Formally, the memory bank is defined as $\mathcal{M} = \{m_k\}_{k=1}^{|\mathcal{M}|}$, where each memory unit m_k represents a textual fragment derived from the dialogue history $\mathcal{H}$. The initial memory bank is produced by a *memory construction function* $f_c : \mathcal{H} \rightarrow \mathcal{M}$, which transforms the dialogue history into a collection of retrievable memory entries. Furthermore, as the conversation progresses dynamically, this memory bank requires continuous maintenance governed by a *memory update function* $f_u : (\mathcal{M}, \Delta h) \rightarrow \mathcal{M}'$, where Δh denotes newly arriving dialogue turns or newly extracted user insights. The design of $f_c(\cdot)$ and $f_u(\cdot)$ collectively determines how conversational history is partitioned, organized, indexed, and maintained over time.

Existing studies primarily focus on complex construction and update functions, employing structured organizations like hierarchical (Hu et al., 2025a; Rezazadeh et al., 2025) or graph-based memories (Zhang et al., 2025a), alongside update strategies such as summarization (Xu et al., 2025; Tan et al., 2025b; Yao et al., 2026) or dynamic forgetting (Packer et al., 2024; Zhang et al., 2025c; Yue et al., 2026). However, maintaining these intricate structures introduces significant computational overhead and risks losing fine-grained contextual nuances. Diverging from these heavily engineered paradigms, we simply append raw historical interactions, allowing the memory bank $\mathcal{M}$ to grow naturally in an unstructured form. By preserving unaltered conversational contexts, we pivot to a query-centric perspective, focusing entirely on the challenges of the retrieval and augmentation stages.

Problem. *(Explicit Memory-Augmented Conversational Agent)*. Given a user request u^*, a budget of N context units, and a memory bank $\mathcal{M}$, the goal is to generate an accurate response r^* by leveraging the external memory bank in two stages: *(i) Retrieval*, where a retrieval function f_r identifies the N most relevant memory units $\{m_n\}_{n=1}^{N} \leftarrow f_r(u^*, \mathcal{M}, N)$ based on the query u^*; and *(ii) Generation*, where the retrieved units are organized chronologically and fed into a language agent to synthesize the final response $r^* = f_g(u^*, \{m_n\}_{n=1}^{N})$.

This formulation aligns with the general paradigm of retrieval-augmented generation (Lewis et al., 2020), where an external memory source provides additional grounding information for the language agent's responses, but differs in that the retrieval corpus consists of the agent's accumulated historical interactions rather than a static external knowledge base.

3 Methodology

3.1 Associate Deep with Turn Isolation Retrieval

Intuition. As mentioned above, we argue that the bottleneck of long-term memory lies not in organization, but in the activation mechanism within the latent knowledge manifold. As the conversation history $\mathcal{H}$ expands, it forms an increasingly dense manifold of interactions. We hypothesize that within this manifold, the information critical to a specific query is not distributed uniformly or densely, but rather exhibits pronounced *decisive evidence sparsity*.

Empirical Study. To validate this hypothesis, we conduct a statistical analysis on the LoCoMo dataset, focusing on the semantic relevance between queries u^* and individual turns $t_{i,j}$ across varying session scales. We observed the following findings from the experiment:

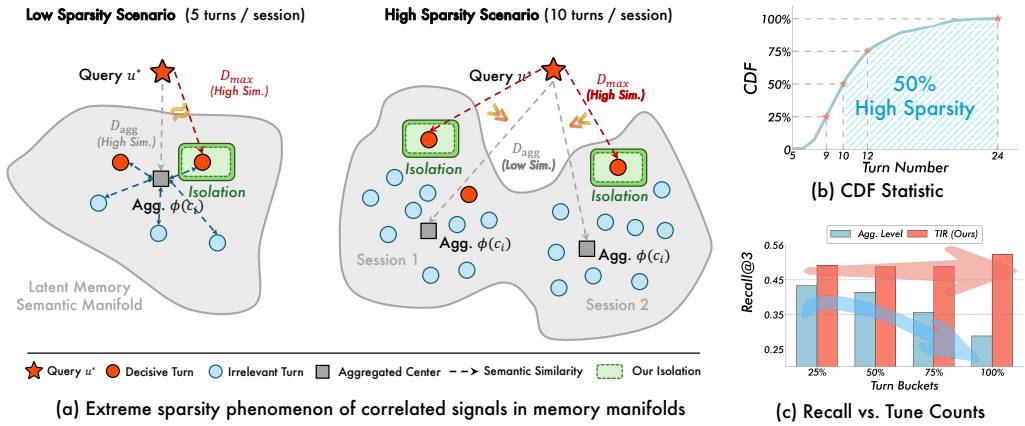


Figure 2: Illustration of *Signal Sparsity Effect* in latent memory manifolds and the robust retrieval performance of *TIR* across varying session lengths.

✷ *Finding I: The Sparsity of Decisive Evidence.* We first examined overall session length, as shown in Fig.2 (b), which illustrates the cumulative distribution function of turn number. While a small percentage of sessions contain fewer than 5 turns, a significant proportion (up to 50%) have 10 or more turns. As shown in Fig.2 (a), in compact sessions, relevant signals are relatively concentrated. However, in extended sessions, the decisive turns (high-relevance points) become geographically isolated within a sea of irrelevant contexts.

✷ *Finding II: The Failure of Global Aggregation.* A direct consequence of sparsity is context dilution. Traditional paradigms, based on aggregation (Xu et al., 2026; Pan et al., 2025), attempts to represent session c_i by compressing the entire content of the session into a vector $\phi(c_i)$, as shown in Fig.2 (a). In compact sessions, this remains close to decisive evidence. However, in longer sessions, this naturally diverges and becomes trivial. Furthermore, as shown in Fig.2 (c), while aggregation-based "turning point level" retrieval performs well in short sessions (25th percentile), its Recall@3 drops sharply with increasing session length (from approximately 0.44 to <0.30). This confirms that the global mean deviates significantly from the target due to the increase in irrelevant words, leading to a failure in retrieval scope.

Implementation Details. Motivated by the verified sparsity, we move away from "over-engineered" memory structures and propose the *Turn Isolation Retrieval (TIR)* mechanism. Instead of attempting to suppress noise through complex summarization, TIR bypasses dilution by directly isolating high-evidence signals within the manifold.

Specificially, Given a memory bank $\mathcal{M}$ and a query u^*, we employ a embedding model ϕ to project inputs into the latent semantic space. For each constituent turn $t_{i,j}$ within a candidate session c_i, we calculate its fine-grained local relevance: $s(u^*, t_{i,j}) = \langle \phi(u^*), \phi(t_{i,j}) \rangle$. To prevent the decisive signal from being obscured by the intra-session context, we define the overall session relevance $\mathcal{S}$ through a max-activation strategy: $\mathcal{S}(c_i, u^*) = \max_{1 \le j \le T_i} s(u^*, t_{i,j})$. Unlike mean-pooling, which is sensitive to session length T_i, this max-isolation ensures that the retrieval score is anchored solely to the most relevant evidence, maintaining a robust retrieval margin even in extremely long sessions. Finally, the retrieval system f_r ranks all candidate sessions and selects the Top-k units:$\{m_n\}_{n=1}^{k} = \text{Top-}k_{c_i \in \mathcal{M}} \mathcal{S}(c_i, u^*)$.

As shown in Fig. 2(c), our TIR mechanism exhibits remarkable stability across all session length percentiles, effectively neutralizing the impact of context expansion.

3.2 Reply Sharp with Query Driven Pruning

Intuition. During retrieval, the TIR mechanism identifies high-gain signal points (Top-k units) in the implicit memory manifold; however, following default settings, the retrieved content is typically the Top-k conversation sessions corresponding to the Top-k units. There-

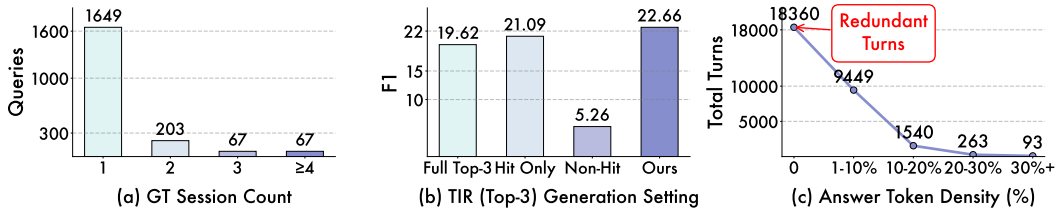


Figure 3: (a) Distribution of the Ground Truth (GT) session count required per query. (b) Downstream QA performance (F1) across different Top-3 generation settings. (c) Distribution of F1 support scores for individual turns within the retrieved GT sessions.

fore, our key intuition is that retrieval solves the "accessibility" problem in the generation phase, but not entirely the "understanding" problem: directly inputting these raw fragments hinders the efficient integration of sparse signals distributed across time slices into a globally coherent response. Furthermore, each dialogue unit (c_i) still contains a significant amount of discourse markers, verbal redundancy, and background information irrelevant to the current query (u^*). Injecting such content into the lengthy context generation process exacerbates noise and further degrades performance, constituting *dual-level redundancy*.

Empirical Study. To validate context refinement, we analyze how retrieved sessions impact response quality to guide our generation strategy, identifying two distinct sources of noise.

✷ *Finding III: The Inter-Session Interference Effect.* We first examine the noise across sessions. As shown in Fig. 3(a), the vast majority of user queries inherently require evidence from only a single Ground Truth (GT) session. Consequently, a standard top-k retrieval setup ($k > 1$) inevitably introduces irrelevant sessions into the generation stage. Fig. 3(b) explicitly demonstrates this negative impact: generating answers using only the correctly retrieved GT session ("Hit Only") achieves an F1 score of 21.09, which significantly outperforms the score obtained using the complete retrieved sessions ("Full Top-3"). Furthermore, forcing the LLM to generate using only the irrelevant Top-3 sessions ("Non-Hit") decreases the F1 score to 5.26, demonstrating that such redundant context acts as misleading semantic noise.

✷ *Finding IV: Intra-Session Conversational Redundancy.* Even when f_r successfully isolates GT sessions, significant redundancy still exists. To illustrate this, we aggregate all constituent turns from ground truth sessions in LoCoMo (totaling 29,705 turns) and evaluate their individual information density. As shown in Fig. 3(c), we measure the *Answer Token Density* by computing the token-level F1 percentage between each raw turn and the ground-truth answer to quantify the exact word-match evidence.The distribution reveals that out of all turns within the ground truth sessions, a massive 18,360 turns yield zero overlap. Yielding no semantic contribution, these turns act purely as redundant conversational padding (e.g., greetings, topic transitions) that misguides generation. Furthermore, even the minority of turns that contain answer tokens predominantly fall into the lowest density bin (1%-10%), as the critical information is heavily diluted by conversational filler. This heavily skewed distribution intuitively demonstrates that ground truth sessions remain overwhelmingly dominated by off-topic conversational inertia rather than query-relevant facts.

Implementation Details. Motivated by these findings, we identify intra- and inter-session redundancy as the primary semantic bottlenecks for generation. To resolve this, we propose *Query Driven Pruning (QDP)* mechanism. Following the retrieval top-(k) sessions $\mathcal{C}_{ret} = \{c_1, \ldots, c_k\}$, we consolidate them into a unified context $\mathcal{E}_{fused}$, then employ a filtering language model $\mathcal{M}_f$ to process. Specifically, $\mathcal{M}_f$ is prompted (see Appendix C) to extract only the query-relevant fragments from $\mathcal{E}_{fused}$.This operation yields the distilled context $\hat{\mathcal{H}} = \text{Filter}(\mathcal{M}_f, \mathcal{E}_{fused}, u^*)$, explicitly purging both inter- and intra-session noise. Finally, the response a is synthesized by the generator f_g using only the pruned context: $a = f_g(\tilde{\mathcal{H}}, u^*)$.

As shown in Fig 3(b), QDP achieves an F1 score of 22.66, significantly outperforming the "Hit Only" setting. This confirms that beyond merely isolating the ground truth session, QDP effectively mitigates intra-session redundancy and maximizes memory utility.

4 Experiments

In this section, we conduct extensive experiments to study the following research questions:

- **RQ1: (Effectiveness)** Can `Nano-Memory` outperform existing agent memory methods?
- **RQ2: (Universality)** Can `Nano-Memory` maintain consistent improvements across different language backbones, retrieval models and diverse query types?
- **RQ3: (Efficiency)** How does `Nano-Memory` compare to baselines in terms of computational overhead, including construction time (offline), inference latency (online), and overall token consumption?
- **RQ4: (Ablation & Mechanism)** How does each component of `Nano-Memory` contribute to its overall performance?

Benchmarks We conduct experiments on four publicly available benchmarks: LoCoMo (Maharana et al., 2024), Long-MT-Bench+ (Pan et al., 2025), LongMemEval-s (Wu et al., 2024) and LongMemEval-m (Wu et al., 2024), which are specifically designed to evaluate agent capabilities in managing long-term conversational dependencies. We utilize the complete set of QA pairs for evaluation, with detailed statistics listed in Appendix A.1.

Baselines We compare `Nano-Memory` with Full History and representative memory baselines: MPNet (Song et al., 2020), Contriever (Izacard et al., 2021),MPC (Lee et al., 2023),RecurSum (Wang et al., 2025),SeCom (Pan et al., 2025),HippoRAG 2 (Gutiérrez et al., 2025),RAPTOR (Sarthi et al., 2024),A-Mem (Xu et al., 2025) and MemGAS (Xu et al., 2026). More baseline details in Appendix A.2

Protocol. We evaluate model performance using F1 (Maharana et al., 2024), BLEU (Papineni et al., 2002), ROUGE (Chin-Yew, 2004), BERTScore (Zhang et al., 2020) and GPT-4o-as-Judge (4o-J) (Zheng et al., 2023). Following the standard experimental setup (Xu et al., 2026),we use `gpt-4o-mini-2024-07-18` (temperature=0) as generator and Contriever as retriever. All baselines operate with a consistent top-k ($k = 3$) retrieval and unified prompts (see Appendix C). We exclude LongMTBench+ due to missing retrieval ground truth and the retrieval performance of RAPTOR and A-Mem cannot be evaluated in all tasks. Results for RAPTOR, A-Mem, and HippoRAG v2 on LongMemEval-m are omitted due to high computational costs. The best and second-best results are in **bold** and underlined, respectively.

4.1 Effectiveness Analysis

Table 1: Retrieval Performance on LoCoMo. (Full results can be found in Appendix Table 7.)

Model	Recall@3	NDCG@3	Recall@5	NDCG@5	Recall@10	NDCG@10	Time
MPNet (2020)	45.92	47.68	53.98	51.76	68.58	56.85	1.42s
Contriever (2021)	49.95	52.15	58.31	56.29	71.80	60.91	1.24s
MPC (2023)	49.50	51.47	57.45	55.53	71.85	60.47	0.81s
RecurSum (2025)	47.23	48.99	59.01	54.58	74.97	60.07	1.26s
SeCom (2025)	52.97	54.03	64.05	59.37	78.25	64.28	2.17s
HippoRAG 2 (2025)	56.60	58.37	65.06	62.50	78.05	66.79	3.70s
MemGAS (2026)	56.85	59.22	67.98	64.25	81.22	68.71	2.64s
`Nano-Memory`	**69.39** (↑22.06%)	**71.19** (↑20.21%)	**76.33** (↑12.28%)	**74.58** (↑16.08%)	**86.46** (↑6.45%)	**77.99** (↑13.51%)	**0.74s** (↓71.97%)

To evaluate the overall effectiveness of our proposed method, we conducted a comprehensive comparison with various ten baselines on four long-term conversation benchmarks. The performance of QA generation and retrieval are presented in Tab. 1 and Tab. 2, respectively. We observe ❶ *Outstanding QA Generation*: `Nano-Memory` achieves competitive generation performance compared to existing baselines. For example, on the LoCoMo dataset, `Nano-Memory` delivers a significant 28.3% improvement in F1 score over the strongest baseline. ❷ *Competitive Retrieval*: Tab. 1 indicates the competitive performance advantages of `Nano-Memory` in the retrieval stage, outperforming the strongest baseline by 22.05% in Recall@3 on LoCoMo dataset. Notably, on LongMemEval-s dataset (detailed in Appendix

Table 2: QA performance. Contriever is the retrieval backbone (excluding Full History and MPNet), with `gpt-4o-mini-2024-07-18` as the generator. *Avg.Tokens* denotes the average token consumption during the generation phase.

Model	4o-J	F1	BLEU4	ROUGE1	ROUGE2	ROUGEL	BERTScore	Avg. Tokens
LoCoMo								
Full History	33.43	12.23	1.84	12.70	5.66	11.73	84.07	20,078
MPNet (2020)	38.07	14.52	2.36	14.97	6.82	13.84	84.46	2,474
Contriever (2021)	40.33	15.76	2.77	16.08	7.75	15.10	84.70	2,348
MPC (2023)	40.38	14.81	1.99	15.10	6.83	14.13	84.43	2,683
RAPTOR (2024)	29.46	14.44	2.85	14.97	7.37	14.02	84.39	1,931
RecurSum (2025)	22.56	9.14	0.99	9.82	3.38	8.98	83.45	3,074
HippoRAG 2 (2025)	45.62	14.95	2.76	15.38	7.52	14.35	84.50	2,991
A-Mem (2025)	40.81	14.72	2.83	16.22	7.71	14.89	84.72	3,042
SeCom (2025)	43.45	15.28	3.12	17.16	8.52	16.01	84.84	1,021
MemGAS (2026)	41.07	17.66	3.61	18.00	8.93	16.99	85.13	2,825
`Nano-Memory`	**48.84**	**22.66**	**5.33**	**22.92**	**12.04**	**21.68**	**85.94**	1,403
LongMTBench+								
Full History	67.44	36.07	11.32	37.90	20.51	29.25	87.81	19,194
MPNet (2020)	63.89	35.32	11.10	37.32	20.45	28.57	87.73	12,187
Contriever (2021)	63.54	31.82	7.51	33.42	18.91	26.53	87.01	12,045
MPC (2023)	61.81	31.52	7.97	33.56	17.27	25.20	86.51	12,289
RAPTOR (2024)	59.72	37.69	13.47	40.08	21.68	30.88	88.38	10,631
RecurSum (2025)	24.65	26.58	6.91	29.23	11.93	20.90	86.11	13,527
HippoRAG 2 (2025)	63.54	35.64	11.05	37.61	20.37	28.76	87.70	13,583
A-Mem (2025)	65.73	36.82	11.36	38.92	20.88	29.14	87.92	13,735
SeCom (2025)	64.58	38.89	13.40	40.87	22.81	31.90	88.45	4,714
MemGAS (2026)	**67.71**	41.49	15.78	43.69	24.47	34.64	88.90	12,873
`Nano-Memory`	64.15	**42.40**	**16.83**	**44.55**	**25.45**	**35.57**	**89.34**	6,174
LongMemEval-s								
Full History	50.60	11.48	1.40	12.10	5.47	10.85	83.07	103,137
MPNet (2020)	53.20	13.96	2.21	14.49	6.78	12.93	83.72	8,173
Contriever (2021)	55.40	13.37	2.15	13.90	6.72	12.43	83.64	8,471
MPC (2023)	53.80	13.60	1.74	14.27	6.49	12.95	83.49	8,457
RAPTOR (2024)	32.20	12.08	1.90	12.73	5.82	11.25	83.50	6,254
RecurSum (2025)	35.40	12.29	2.09	13.01	5.55	11.52	83.60	8,853
HippoRAG 2 (2025)	57.60	14.73	2.15	15.30	7.36	13.83	83.86	8,530
A-Mem (2025)	55.60	13.73	2.11	14.82	6.81	12.98	83.88	9,018
SeCom (2025)	56.00	12.95	2.25	13.80	6.09	11.93	83.51	2,741
MemGAS (2026)	**60.20**	20.74	4.48	21.36	10.47	19.80	85.29	8,829
`Nano-Memory`	57.20	**21.06**	**4.89**	**21.80**	**10.77**	**20.16**	**85.54**	4,302
LongMemEval-m								
Full History	12.20	5.70	0.78	6.27	2.08	5.28	81.62	128,000
MPNet (2020)	37.80	10.97	1.65	11.57	4.99	10.07	83.06	7,989
Contriever (2021)	42.80	11.88	1.66	12.63	5.51	11.11	83.31	8,274
MPC (2023)	37.80	11.28	1.37	11.93	5.12	10.57	82.98	8,428
RecurSum (2025)	23.80	10.04	1.70	10.89	4.26	9.21	83.12	8,927
SeCom (2025)	42.80	11.33	1.79	12.03	5.07	10.49	83.36	2,821
MemGAS (2026)	45.40	16.85	3.39	17.60	8.25	16.14	84.69	8,852
`Nano-Memory`	**46.60**	**18.09**	**3.96**	**18.88**	**9.03**	**17.40**	**85.10**	4,261

Tab. 7), although MemGAS achieves higher retrieval recall, its multi-granularity mechanism inherently introduces broad conversational noise during the retrieval stage, which ultimately degrades its final QA performance. In contrast, `Nano-Memory` utilizes QDP during the generation stage to filter out intre- and intra-session noise, effectively leveraging raw retrieval results to achieve superior downstream QA performance.

4.2 Universality Analysis

To comprehensively evaluate the robustness of `Nano-Memory`, we conduct a universality analysis across three key dimensions: ❶*Varying LLM backbones:* As shown in Fig. 4, `Nano-Memory` demonstrates strong universality by achieving substantial performance gains

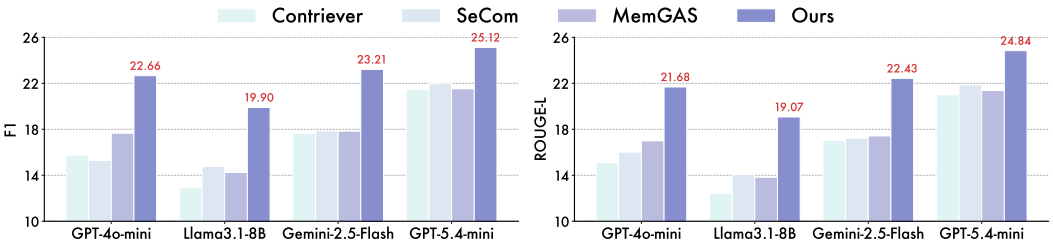


Figure 4: Comparison of different LLM backbones on LoCoMo. (Full results in Tab. 8.)

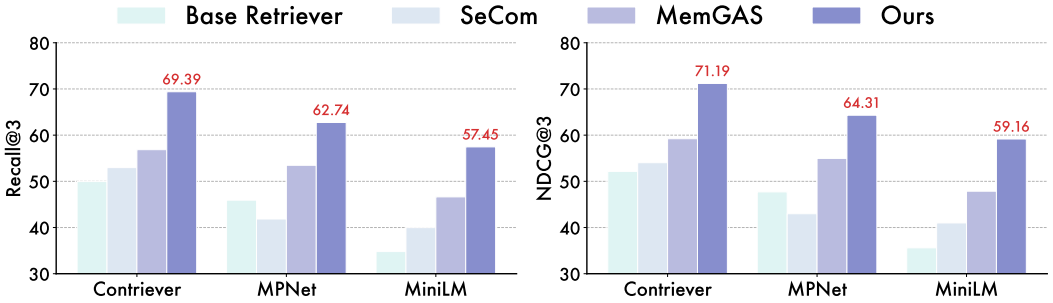


Figure 5: Comparison of different retrievers on LoCoMo. *Base Retriever* refers to the vanilla Contriever, MPNet, and MiniLM. (Full results can be found in Appendix Tab. 9.)

Table 3: Time Cost on LoCoMo under the same setting.

Model	**Consturct**	**Retrieval**	**Generation**	**Total Time**	**F1**
Contriever	\	1.24s	381.71s.	382.95s	15.76
MPNet	\	1.42s	317.28s	318.70s	14.52
Secom	294.62s	2.17s	111.81s	316.17s	15.28
MemGAS	101.42s	2.64s	281.93s	361.78s	17.66
`Nano-Memory`	\	**0.74s**	**187.2s**	**187.96s**	**22.66**

Table 4: Ablation Study on LoCoMo. (Full results can be found in Appendix Tab. 11.)

Model	**4o-J**	**F1**	**BLEU4**	**ROUGE1**	**ROUGE2**	**ROUGEL**	**BERTScore**	**Avg. Tokens**
Baseline	40.33	15.76	2.77	16.08	7.75	15.10	84.70	2,348
w/ *TIR*	43.42	19.62	3.85	19.86	10.43	18.68	85.33	2,685
w/ *TIR + QDP*	**48.84**	**22.66**	**5.33**	**22.92**	**12.04**	**21.68**	**85.94**	1,403

with both closed-source and open-source LLMs. ❷ *Different retrievers:*Fig. 5 demonstrates that `Nano-Memory` consistently achieves outstanding retrieval performance across various retrieval backbones. Specifically, it yields relative Recall@3 improvements of 22.06%, 17.34% and 23.20% over the strongest baseline when paired with Contriever, MPNet, and MiniLM, respectively. The corresponding final QA generation performance for each retriever is detailed in Appendix B.1 (Tab. 10). ❸ *Diverse query types* Fig. 6 highlights `Nano-Memory` achieves the best performance across all query types. Notably, the largest relative gain (39.7%) is observed on *Temporal* queries, directly exposing the inherent flaw of existing retrieval mechanisms that lose hard temporal constraints during context compression. Results for different retrievers under various query settings are detailed in Appendix B.2 (Fig. 10).

4.3 Efficiency Analysis

We evaluate the end-to-end computational overhead on LoCoMo dataset, covering time costs (Tab. 3) and token consumption (Fig. 7). ❶ *Time Efficiency*: Tab. 3 demonstrates that

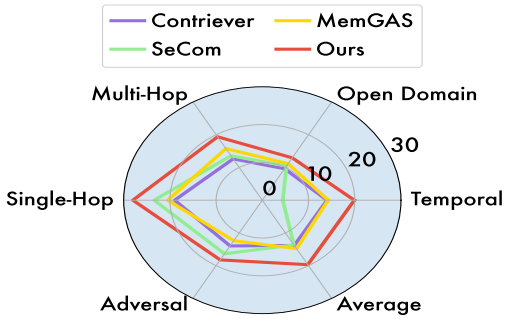


Figure 6: Comparison of F1 scores across different query types in LoCoMo.

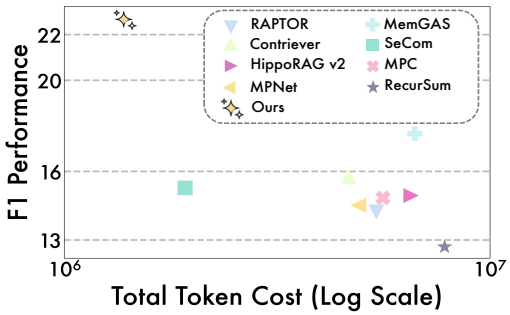


Figure 7: Trade-off between F1 score and total token consumption on LoCoMo.

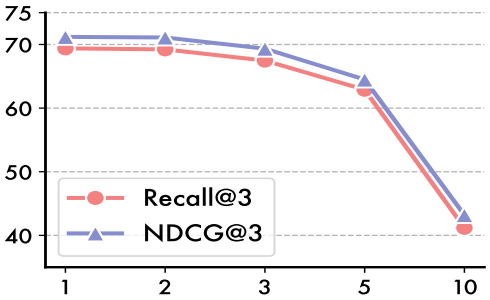


Figure 8: Retrieval performance when aggregating the top-k relevant turns.

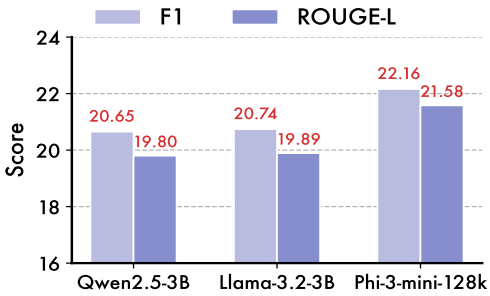


Figure 9: Generation performance when using different language models for pruning.

`Nano-Memory` reduces total execution time by 48.05% compared to the strongest baseline. Traditional memory systems face a strict latency trade-off: they either incur heavy offline computation costs for structural maintenance, or suffer from severe online generation delays due to processing redundant historical contexts. `Nano-Memory` effectively breaks this bottleneck. By operating directly on unstructured dialogue history, it mitigates offline construction overhead. Moreover, filtering conversational noise via QDP significantly accelerates online generation, confirming that our paradigm optimizes end-to-end efficiency without structural maintenance overhead. ❷ *Token Efficiency*: As illustrated in Fig. 7, `Nano-Memory` achieves an outstanding trade-off between F1 performance and token consumption. Baselines typically incur heavy token overhead by requiring iterative LLM inferences for offline memory construction or directly processing noisy top-k contexts. By bypassing offline construction and explicitly pruning redundant contexts, `Nano-Memory` reduces overall token consumption by 3–5$\times$ while increasing response accuracy. This exposes a fundamental inefficiency in existing memory paradigms: they allocate substantial computational budget to processing conversational filler, rather than focusing strictly on sparse query-relevant signals.

4.4 Ablation & Mechanism Analysis

Ablation Study. Tab. 4 shows that the integration of TIR and QDP is essential to mitigate the two facets of the *Signal Sparsity Effect*. At the retrieval stage, TIR successfully overcomes Decisive Evidence Sparsity by directly isolating high-evidence signals within the manifold, resulting in an initial improvement of F1 to 19.62 on LoCoMo. However, feeding these raw retrieved sessions directly into the generator degrades response quality due to conversational noise. At the generation stage, `Nano-Memory` mitigates *dual-level redundancy* by pruning this redundant context via QDP, which further boosts the F1 score to 22.66 while sharply reducing token consumption from 2,685 to 1,403. This suggests that improving signal density can serve as a highly effective alternative to increasing architectural complexity. Complete ablation results across varying query settings are provided in Appendix B.3.

TIR Mechanism. We determine a candidate session's relevance score by aggregating the query-turn similarities of its top-k most relevant turns. As illustrated in Fig. 8, increasing the number of retained turns (k) sharply degrades overall performance, because the aggregation of extra turns introduces conversational noise that dilutes the decisive matching signal.

QDP Mechanism. To examine whether QDP relies on specific generator architectures, we evaluate the pruning effectiveness using lightweight models including Qwen2.5-3B, Llama-3.2-3B, and Phi-3-mini-128k. As shown in Fig. 9, `Nano-Memory` maintains robust performance across all pruners, demonstrating that the pruning mechanism is compatible with various lightweight backbones and does not require architectural alignment with the generator.

5 Related Work

5.1 The Representation Ability for Better Retrieval.

Long-term conversational memory inherently requires segmentation strategies to manage growing contexts (Maharana et al., 2024; Wu et al., 2024; Pan et al., 2025). Existing methods (categorized in Appendix A.2, Tab. 6) predominantly rely on *single-granularity* representations, utilizing discrete conversational units (e.g., *turn-level* (Yuan et al., 2024) and *session-level* (Li et al., 2025; Lu et al., 2023; Lee et al., 2023)), abstractive compressions (e.g., *summary-level* (Zhong et al., 2024; Chen et al., 2025; Lu et al., 2023), *keyword-level* (Xu et al., 2025; Chhikara et al., 2025; Zhou & Han, 2025; Ke et al., 2025) or semantic boundary *topic-level* (Pan et al., 2025; Fang et al., 2025; Fountas et al., 2025)). However, these approaches face a fundamental limitation: they either fragment episodic context or induce semantic dilution, losing decisive evidence during retrieval. To address this, *Multi-granularity* paradigm (Xu et al., 2026; Jiang et al., 2026; Li et al., 2026) fuses representations across multiple granularities yet introduces redundancy. Unlike these paradigms, our retrieval mechanism directly retrieves the single most relevant turn, preserving the core matching signal from raw dialogue, effectively avoiding context dilution without compromising precision.

5.2 The Information Feature for Better Responses.

Existing memory systems (Xu et al., 2025; Packer et al., 2024; Chhikara et al., 2025) primarily focus on memory construction and retrieval, ignoring noise within the retrieved contexts. Post-retrieval reranking methods reorder retrieval contexts with heuristic rules (Liu et al., 2025) or parametric learning Tan et al. (2025a); Du et al. (2025), however, they fail to remove redundant information from the retrieved contexts. Constraint-based approaches (Rasmussen et al., 2025; Han et al., 2025) use fixed rules to filter retrieved contexts based on temporal/logical conflicts, but they depend on hand-crafted metadata and cannot adapt to different queries. Multi-granularity (Xu et al., 2026; Li et al., 2026) enrich the generation phase by appending vary representations(e.g., keywords, summaries) from the retrieved contexts, introducing representational redundancy and computational overhead. In contrast, our QDP filters the retrieved contexts, ensuring robust generation while reducing noise.

6 Conclusion, Limitations and Future Work

In this paper, we proposed `Nano-Memory` , a nonparametric memory for conversational agent. By designing TIR that scores each past session by its single most relevant turn and QDP that filters redundant retrieved context, our method effectively preserves sparse critical information while eliminating noise. Extensive experiments demonstrate its effectiveness, universality, and efficiency. However, `Nano-Memory` 's passive, unstructured paradigm limits the agent's capacity for proactive self-evolution and continuous knowledge consolidation. Furthermore, the reactive nature of online query-driven pruning introduces sequential inference latency. Future research will focus on transitioning this framework into a proactive, self-evolving memory system capable of autonomous, offline knowledge restructuring.

Ethics Statement

This paper presents work whose goal is to advance the field of language models. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62402414), the Guangdong Basic and Applied Basic Research Foundation (No. 2025A1515011994), Guangdong Provincial Project 2025D03J0014, Guangzhou Municipal Science and Technology Project (No. 2023A03J0011), the Guangzhou Industrial Information and Intelligent Key Laboratory Project (No. 2024A03J0628), and Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things (No. 2023B1212010007). Additional support came from Guangzhou HKUST Fok Ying Tung Research Institute (grant No. 2023ZD007), the HKUST-MetaX Joint Laboratory for Advanced AI Computing (grant No. METAX24EG01), and the JC STEM Lab of Data Science Foundations, which is funded by The Hong Kong Jockey Club Charities Trust (where part of this work was conducted).

References

Anthropic. Effective harnesses for long-running agents. *Engineering at Anthropic*, 2025.

Vincent-Pierre Berges, Barlas Oğuz, Daniel Haziza, Wen-tau Yih, Luke Zettlemoyer, and Gargi Ghosh. Memory layers at scale. *arXiv preprint arXiv:2412.09764*, 2024.

Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 755–773, 2025.

Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Lin Chin-Yew. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out, 2004*, 2004.

Yiming Du, Bingbing Wang, Yang He, Bin Liang, Baojun Wang, Zhongyang Li, Lin Gui, Jeff Z Pan, Ruifeng Xu, and Kam-Fai Wong. Memguide: Intent-driven memory selection for goal-oriented multi-session llm agents. *arXiv preprint arXiv:2505.20231*, 2025.

Sabri Eyuboglu, Ryan Ehrlich, Simran Arora, Neel Guha, Dylan Zinsley, Emily Liu, Will Tennien, Atri Rudra, James Zou, Azalia Mirhoseini, et al. Cartridges: Lightweight and general-purpose long context representations via self-study. *arXiv preprint arXiv:2506.06266*, 2025.

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. Lightmem: Lightweight and efficient memory-augmented generation, 2025. URL https://arxiv.org/abs/2510.18866.

Zafeirios Fountas, Martin Benfeghoul, Adnan Oomerjee, Fenia Christopoulou, Gerasimos Lampouras, Haitham Bou Ammar, and Jun Wang. Human-inspired episodic memory for infinite context llms. In *The Thirteenth International Conference on Learning Representations*, 2025.

Manthan Gupta. I reverse engineered claude's memory system, and here's what i found! https://manthanguptaa.in/posts/claude_memory/, 2025a.

Manthan Gupta. I reverse engineered chatgpt's memory system, and here's what i found! https://manthanguptaa.in/posts/chatgpt_memory/, 2025b.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From rag to memory: Non-parametric continual learning for large language models. *arXiv preprint arXiv:2502.14802*, 2025.

Jiale Han, Austin Cheung, Yubai Wei, Zheng Yu, Xusheng Wang, Bing Zhu, and Yi Yang. Rag meets temporal graphs: Time-sensitive modeling and retrieval for evolving knowledge. *arXiv preprint arXiv:2510.13590*, 2025.

Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. Hiagent: Hierarchical working memory management for solving long-horizon agent tasks with large language model. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32779–32798, 2025a.

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

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*, 2021.

Dongming Jiang, Yi Li, Guanpeng Li, and Bingzhe Li. Magma: A multi-graph based agentic memory architecture for ai agents, 2026. URL https://arxiv.org/abs/2601.03236.

Cai Ke, Yiming Du, Bin Liang, Yifan Xiang, Lin Gui, Zhongyang Li, Baojun Wang, Yue Yu, Hui Wang, Kam-Fai Wong, et al. Flexibly utilize memory for long-term conversation via a fragment-then-compose framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 21130–21147, 2025.

Ronald Kemker, Marc McClure, Angelina Abitino, Tyler Hayes, and Christopher Kanan. Measuring catastrophic forgetting in neural networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Gibbeum Lee, Volker Hartmann, Jongho Park, Dimitris Papailiopoulos, and Kangwook Lee. Prompted llms as chatbot modules for long open-domain conversation. In *Findings of the association for computational linguistics: ACL 2023*, pp. 4536–4554, 2023.

Songxin Lei, Huijun Tang, Chuangyi Li, Xueying Zhang, Chenli Xu, and Huaming Wu. Federated maddpg-based collaborative scheduling strategy in vehicular edge computing. *IEEE Transactions on Mobile Computing*, 2025a.

Songxin Lei, Qiongyan Wang, Yanchen Zhu, Hanyu Yao, Sijie Ruan, Weilin Ruan, Yuyu Luo, Huaming Wu, and Yuxuan Liang. A game-theoretic spatio-temporal reinforcement learning framework for collaborative public resource allocation, 2025b. URL https://arxiv.org/abs/2510.26184.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. Hello again! llm-powered personalized agent for long-term dialogue. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5259–5276, 2025.

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. Timem: Temporal-hierarchical memory consolidation for long-horizon conversational agents, 2026. URL https://arxiv.org/abs/2601.02845.

Jun Liu, Zhenglun Kong, Changdi Yang, Fan Yang, Tianqi Li, Peiyan Dong, Joannah Nanjekye, Hao Tang, Geng Yuan, Wei Niu, et al. Rcr-router: Efficient role-aware context routing for multi-agent llm systems with structured memory. *arXiv preprint arXiv:2508.04903*, 2025.

Na Liu, Liangyu Chen, Xiaoyu Tian, Wei Zou, Kaijiang Chen, and Ming Cui. From llm to conversational agent: A memory enhanced architecture with fine-tuning of large language models. *arXiv preprint arXiv:2401.02777*, 2024.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Rocco Lopopolo. Harness engineering: leveraging codex in an agent-first world. *OpenAI engineering note*, 2026.

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

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 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)*, pp. 13851–13870, 2024.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Menglin Xia, Xufang Luo, Jue Zhang, Qingwei Lin, Victor Rühle, Yuqing Yang, Chin-Yew Lin, et al. Llmlingua-2: Data distillation for efficient and faithful task-agnostic prompt compression. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 963–981, 2024.

Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H Vicky Zhao, Lili Qiu, et al. On memory construction and retrieval for personalized conversational agents. *arXiv preprint arXiv:2502.05589*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

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

Alireza Rezazadeh, Zichao Li, Wei Wei, and Yujia Bao. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for llms. In *The Thirteenth International Conference on Learning Representations*, 2025.

Parth Sarthi, Salman Abdullah, Aditi Tuli, Shubh Khanna, Anna Goldie, and Christopher D Manning. Raptor: Recursive abstractive processing for tree-organized retrieval. In *The Twelfth International Conference on Learning Representations*, 2024.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

Haoran Sun and Shaoning Zeng. Hierarchical memory for high-efficiency long-term reasoning in llm agents. *arXiv preprint arXiv:2507.22925*, 2025.

Xingyu Tan, Xiaoyang Wang, Qing Liu, Xiwei Xu, Xin Yuan, Liming Zhu, and Wenjie Zhang. Memotime: Memory-augmented temporal knowledge graph enhanced large language model reasoning. *arXiv preprint arXiv:2510.13614*, 2025a.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, et al. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8416–8439, 2025b.

Qingyue Wang, Yanhe Fu, Yanan Cao, Shuai Wang, Zhiliang Tian, and Liang Ding. Recursively summarizing enables long-term dialogue memory in large language models. *Neurocomputing*, 639:130193, 2025.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL https://arxiv.org/abs/2002.10957.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *arXiv preprint arXiv:2410.10813*, 2024.

Yuqian Wu, Yuhong Peng, Jiapeng Yu, and Raymond Lee. Mas4poi: a multi-agents collaboration system for next poi recommendation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 356–367. Springer, 2025.

Derong Xu, Yi Wen, Pengyue Jia, Yingyi Zhang, Yichao Wang, Huifeng Guo, Ruiming Tang, Xiangyu Zhao, Enhong Chen, Tong Xu, et al. From single to multi-granularity: Toward long-term memory association and selection of conversational agents. *The Fourteenth International Conference on Learning Representations.*, 2026.

Wujiang Xu, Kai Mei, Hang Gao, Juntao Tan, Zujie Liang, and Yongfeng Zhang. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*, 2025.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Yilun Yao, Shan Huang, Elsie Dai, Zhewen Tan, Zhenyu Duan, Shousheng Jia, Yanbing Jiang, and Tong Yang. Arc: Active and reflection-driven context management for long-horizon information seeking agents. *arXiv preprint arXiv:2601.12030*, 2026.

Jiapeng Yu, Yuqian Wu, Yajing Zhan, Wenhao Guo, Zhou Xu, and Raymond Lee. Colearning: code learning for multi-agent reinforcement collaborative framework with conversational natural language interfaces. *Frontiers in Artificial Intelligence*, 8:1431003, 2025.

Ruifeng Yuan, Shichao Sun, Yongqi Li, Zili Wang, Ziqiang Cao, and Wenjie Li. Personalized large language model assistant with evolving conditional memory, 2024. URL https://arxiv.org/abs/2312.17257.

Yanwei Yue, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. Mem-t: Densifying rewards for long-horizon memory agents. *arXiv preprint arXiv:2601.23014*, 2026.

Guibin Zhang, Muxin Fu, Kun Wang, Guancheng Wan, Miao Yu, and Shuicheng YAN. G-memory: Tracing hierarchical memory for multi-agent systems. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Guibin Zhang, Muxin Fu, and Shuicheng Yan. Memgen: Weaving generative latent memory for self-evolving agents. *The Fourteenth International Conference on Learning Representations.*, 2025b.

Guibin Zhang, Haotian Ren, Chong Zhan, Zhenhong Zhou, Junhao Wang, He Zhu, Wangchunshu Zhou, and Shuicheng Yan. Memevolve: Meta-evolution of agent memory systems. *arXiv preprint arXiv:2512.18746*, 2025c.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL https://arxiv.org/abs/1904.09675.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

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

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

Adam Zweiger, Xinghong Fu, Han Guo, and Yoon Kim. Fast kv compaction via attention matching. *arXiv preprint arXiv:2602.16284*, 2026.

SUPPLEMENTARY MATERIAL

BACK TO BASICS: LET CONVERSATIONAL AGENTS REMEMBER WITH JUST RETRIEVAL AND GENERATION

TABLE OF CONTENTS

A Experimental Details **16**

A.1 Datasets Details 16

A.2 Baseline Details 17

A.3 Implementation Details 19

B More Experimental Analysis **19**

B.1 Effectiveness Analysis 19

B.2 Universality Analysis 20

B.3 Ablation and Mechanism Analysis 21

C Prompts **23**

D More Discussion **23**

D.1 Current Limitation 23

D.2 Future Direction 24

E LLM Disclosure **24**

A Experimental Details

A.1 Datasets Details

Table 5: Detailed statistics of all evaluated datasets. The term "Avg." (e.g., Avg. Queries) denotes the average number corresponding to each conversation.

Dataset	LoCoMo	Long-MT-Bench+	LongMemEval-s	LongMemEval-m
Conversation Subject	User-User	User-AI	User-AI	User-AI
Session Dates	✓	✗	✓	✓
Retrieval Ground-Truth	✓	✗	✓	✓
QA Ground-Truth	✓	✓	✓	✓
Total Conversations	10	11	500	500
Avg. Sessions	27.2	4.9	47.7	475.3
Avg. Queries	198.6	26.2	1.0	1.0
Avg. Turns	301.1	65.5	248.7	2466.4
Avg. Tokens	20,756.2	19,287.5	104,315.2	1,030,374.7

LoCoMo (Maharana et al., 2024) is designed to evaluate long-term memory in long-context LLMs and RAG systems in multi-turn question-answering tasks. In this study, we employ its publicly available subset, which contains 10 high-quality, long conversations with 27.2 sessions and 20,756.2 tokens on average. The evaluation tasks consist of 1,986 queries

Table 6: Comparison of various retrieval granularities. **Efficiency** is measured by Tokens (generation consumption) and Latency. **Simplicity** is quantified by direct inference over raw dialogue history and the avoidance of representational redundancy (e.g., multi-granularity clustering).✓ : High, △ : Moderate, ✗ : Low. ∗ indicates a specific variant that isolates the decisive turn to mitigate context dilution.

Category	Example Works	Operation	Effectiveness	Simplicity	Tokens	Latency
Keyword-Level	A-Mem (2025), H-MEM (2025)	Extract turn/session details	✗	△	△	△
Turn-Level	Conditional Memory (2024)	Raw Turn Retrieval	✗	✓	✓	✓
Session-Level	MemoChat (2023), LD-Agent (2025)	Raw Session Retrieval	✗	✓	✓	✓
Summary-Level	Memorybank (2024), Recursum (2025)	Recurrent summary	✗	△	△	△
Topic-Level	Secom (2025), LightMem (2025)	Semantic segmentation	△	△	△	△
Multi-Granularity	MemGAS (2026), TiMem (2026)	Adaptive Fusion	✓	✗	✗	✗
Turn − Level∗	**Nano-Memory (Ours)**	Mitigate Sparsity	✓	✓	✓	✓

covering five problem types: (i) single-hop retrieval, which sources answers from a specific isolated session; (ii) multi-hop retrieval, which necessitates synthesizing details scattered across multiple sessions; (iii) temporal reasoning, demanding the deduction of chronological sequences and time intervals; (iv) open-domain knowledge, which requires grounding speaker information in external commonsense or world knowledge ; and (v) adversarial, evaluating the model's ability to detect and refuse to answer misleading or unanswerable prompts.

Long-MT-Bench+ (Pan et al., 2025) is an enhanced long-dialogue benchmark reconstructed from MT-Bench+ (Lu et al., 2023). It addresses scarce QA pairs and short dialogues by merging five consecutive sessions into a long-form conversation. The dataset contains 11 conversations, averaging 4.9 sessions and approximately 19287.5 tokens. Unlike LoCoMo, it focuses exclusively on user-AI interactions and does not provide session timestamps or retrieval ground truth.

LongMemEval-s (Wu et al., 2024) follows the needle-in-a-haystack paradigm, focusing on dynamically tracking and updating information in task-oriented conversations. The benchmark includes 500 manually designed questions, each accompanied by a chat history of about 47.7 sessions and averaging 104,315.2 tokens. The queries are systematically categorized to evaluate diverse memory dimensions: (i) single session user, recalling specific details provided by the user within an isolated interaction; (ii) single session assistant, retrieving information previously generated by the agent in a single session; (iii) single session preference, personalizing responses based on user traits shared locally; (iv) multi-session, synthesizing dispersed evidence across multiple sessions to address complex questions; (v) temporal reasoning, resolving chronological dependencies, including explicit dates and inferred time references and (vi) knowledge update, tracking dynamically evolving user states and reasoning with overridden personal information over time.

LongMemEval-m (Wu et al., 2024) is an extended version of LongMemEval-s, scaling each dialogue history to an average of 500 sessions and 1030374.7 tokens. It is designed to detect lost-in-the-middle phenomenon in long-context models and to assess their ability to maintain temporal consistency and perform causal reasoning over extended contexts. Both versions challenge AI systems to process lengthy user-assistant interactions, maintain dynamically evolving memory, and ensure consistency throughout conversational history.

A.2 Baseline Details

In this appendix, we provide detailed descriptions of the advanced memory methods and different RAG systems used in our default evaluation.

- *Full History* is a retrieval-free baseline that feeds the most recent 128k tokens of dialogue history directly into the LLM's context window, allowing the model to process the complete, unedited history.

- *MPNet* (Song et al., 2020) is a pre-training approach for LLMs that combines the strengths of masked language modeling from BERT (Liu et al., 2019) and permuted language modeling from Xlnet (Yang et al., 2019), while addressing their key limitations. It captures dependencies among predicted tokens through permutation-based objectives and incorporates auxiliary position information to align pre-training more closely with downstream tasks, enabling better utilization of full sentence context. This method particularly excels in natural language understanding and downstream performance across a range of tasks. https://github.com/microsoft/MPNet
- *Contriever* (Izacard et al., 2021) is an unsupervised dense retrieval model trained via contrastive learning, enabling effective representation of queries and documents without labeled data. It demonstrates strong capabilities in zero-shot retrieval across diverse domains, as well as in-domain performance when further fine-tuned, and particularly excels in multilingual and cross-lingual retrieval settings, including challenging scenarios involving different scripts. https://github.com/facebookresearch/contriever
- *MPC* (Lee et al., 2023) is a modular approach to building conversational agents that leverages pre-trained large language models as independent components to achieve long-term consistency and flexibility. It employs techniques such as few-shot prompting, chain-of-thought reasoning, and external memory, enabling the creation of consistent and engaging chatbots without requiring model fine-tuning. https://github.com/krafton-ai/MPC
- *RAPTOR* (Sarthi et al., 2024) is a retrieval-augmented framework that recursively embeds, clusters, and summarizes text chunks to build a multi-layer tree structure with increasing levels of abstraction from bottom-up summaries. At inference time, it retrieves relevant information across different abstraction levels, enabling effective integration of context from long documents. RAPTOR performs strongly on long-context QA tasks involving single-hop and multi-hop reasoning. https://github.com/parthsarthi03/raptor
- *RecurSum* (Wang et al., 2025) is a recursive summarization approach that enhances long-term dialogue capabilities in large language models by progressively generating updated memories. It begins by having the LLM memorize short dialogue contexts, then iteratively produces new summaries incorporating prior memories and additional conversation segments, enabling the chatbot to generate responses grounded in the most recent consolidated memory. This method particularly excels in maintaining consistency across extended conversations. https://github.com/qingyue2014/Rsum
- *HippoRAG 2* (Gutiérrez et al., 2025) is a retrieval-augmented framework that advances long-term memory capabilities in LLMs by leveraging an enhanced Personalized PageRank algorithm, incorporating deeper passage integration and efficient online LLM utilization. It excels in factual knowledge recall, sense-making, and associative memory tasks.https://github.com/OSU-NLP-Group/HippoRAG
- *A-Mem* (Xu et al., 2025) is an agentic memory system for LLM agents that dynamically organizes memories by drawing inspiration from the Zettelkasten method to build interconnected knowledge networks. When adding a new memory, a-mem generates structured notes with contextual descriptions, keywords, and tags, then identifies connections with historical memories to establish links, while enabling memory evolution through updates to existing representations as new information is integrated. https://github.com/WujiangXu/A-mem
- *SeCom* (Pan et al., 2025) is a memory construction framework that builds a segment-level memory bank using a conversation segmentation model to divide long-term conversations into topically coherent units, while applying LLMLingua-2 (Pan et al., 2024) compression to denoise memory entries and improve retrieval performance. https://github.com/microsoft/SeCom
- *MemGAS* (Xu et al., 2026) is a memory consolidation framework that organizes memories into units of varying granularity and uses Gaussian Mixture Models to cluster

Table 7: Retrieval Performance. Contriever is used as the default retriever for all methods except MPNet.

Model	Recall@3	NDCG@3	Recall@5	NDCG@5	Recall@10	NDCG@10	Time
LoCoMo							
MPNet (2020)	45.92	47.68	53.98	51.76	68.58	56.85	1.42s
Contriever (2021)	49.95	52.15	58.31	56.29	71.80	60.91	1.24s
MPC (2023)	49.50	51.47	57.45	55.53	71.85	60.47	0.81s
RecurSum (2025)	47.23	48.99	59.01	54.58	74.97	60.07	1.26s
SeCom (2025)	52.97	54.03	64.05	59.37	78.25	64.28	2.17s
HippoRAG 2 (2025)	56.60	58.37	65.06	62.50	78.05	66.79	3.70s
MemGAS (2026)	56.85	59.22	67.98	64.25	81.22	68.71	2.64s
Nano-Memory	**69.39**	**71.19**	**76.33**	**74.58**	**86.46**	**77.99**	0.74s
LongMemEval-s							
MPNet (2020)	66.60	75.86	77.02	78.72	85.96	81.07	0.47s
Contriever (2021)	71.06	79.84	81.28	82.56	90.00	84.38	0.39s
MPC (2023)	75.32	87.47	75.96	86.79	77.23	87.17	0.71s
RecurSum (2025)	60.00	70.90	68.09	73.27	80.00	76.59	0.57s
SeCom (2025)	71.06	80.88	80.43	83.08	89.15	85.11	0.79s
HippoRAG 2 (2025)	75.53	85.44	84.68	87.32	91.28	88.73	21.46s
MemGAS (2026)	**78.51**	**86.83**	**88.30**	**88.86**	92.55	89.79	2.19s
Nano-Memory	77.23	85.77	87.02	88.33	**94.26**	**89.81**	0.25s
LongMemEval-m							
MPNet (2020)	38.09	49.91	47.23	53.81	61.70	57.51	2.16s
Contriever (2021)	45.32	57.18	54.26	60.42	66.81	63.63	1.78s
MPC (2023)	35.96	47.51	42.55	50.36	54.26	53.88	2.34s
SeCom (2025)	51.49	63.31	**63.62**	67.14	68.72	68.68	3.10s
MemGAS (2026)	51.06	61.36	**63.62**	66.07	**77.02**	69.46	8.88s
Nano-Memory	**52.77**	**64.61**	**63.62**	**68.23**	75.74	**71.42**	2.08s

and associate new memories with existing ones. An entropy-based router dynamically selects the most appropriate granularity for each query by assessing relevance distributions, ensuring a balance between information completeness and minimal noise. https://github.com/quqxui/MemGAS

A.3 Implementation Details

Across all tasks, we utilize *gpt-4o-mini-2024-07-18* as the default generator. To ensure strict experimental reproducibility and eliminate generative randomness, we set the LLM temperature to 0 and cap the maximum response length at 4000 tokens. All evaluations are conducted on a Linux server featuring an AMD EPYC 7763 128-Core Processor with 256GB of RAM, accelerated by 4 $\times$ NVIDIA RTX A6000 GPUs, each with 48GB memory.

B More Experimental Analysis

B.1 Effectiveness Analysis

We provide complete retrieval performance across multiple metrics on three long-term conversational benchmarks, as presented in Tab. 7. Note that LongMTBench+ is excluded from this evaluation because it lacks retrieval ground truth.

Table 8: QA Performance Comparison of Different Generators on LoCoMo with Contriever.

Model	4o-J	F1	BLEU4	Rouge1	Rouge2	RougeL	BERTScore
GPT-4o-mini-2024-07-18							
Contriever (2021)	40.33	15.76	2.77	16.08	7.75	15.10	84.70
SeCom (2025)	43.45	15.28	3.12	17.16	8.52	16.01	84.84
MemGAS (2026)	41.07	17.66	3.61	18.00	8.93	16.99	85.13
Nano-Memory	**48.84**	**22.66**	**5.33**	**22.92**	**12.04**	**21.68**	**85.94**
Llama3.1-8b-instruct							
Contriever (2021)	37.97	12.91	1.70	13.19	6.38	12.41	83.90
SeCom (2025)	40.58	14.78	2.06	15.06	7.55	14.08	84.27
MemGAS (2026)	34.69	14.24	1.66	14.64	7.20	13.83	84.14
Nano-Memory	**44.66**	**19.90**	**3.11**	**20.03**	**10.84**	**19.07**	**85.19**
gpt-5.4-mini							
Contriever (2021)	46.68	21.44	3.79	22.33	10.84	21.01	85.33
SeCom (2025)	47.99	21.97	4.12	21.06	11.68	21.85	85.47
MemGAS (2026)	44.46	21.52	4.56	22.47	11.29	21.37	85.44
Nano-Memory	**53.58**	**25.12**	**5.59**	**25.97**	**14.01**	**24.84**	**85.82**
Gemini-2.5-flash							
Contriever (2021)	33.23	17.65	3.35	18.06	8.86	17.03	84.82
SeCom (2025)	35.55	21.97	4.12	21.06	11.68	21.85	85.47
MemGAS (2026)	38.37	17.84	3.33	18.36	8.92	17.42	85.00
Nano-Memory	**43.47**	**23.21**	**4.81**	**23.59**	**12.44**	**22.43**	**85.81**

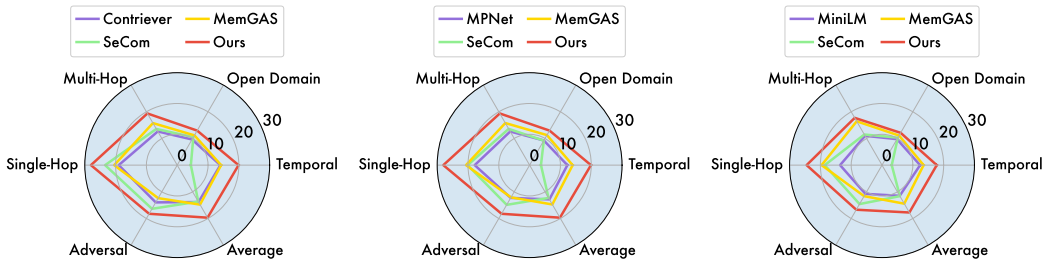


Figure 10: Comparison of F1 scores across different query types in LoCoMo using *gpt-4o-mini-2024-07-18* as the generator with Contriever(left), MPNet (middle), and MiniLM (right)

B.2 Universality Analysis

We provide more experimental results for the universality analysis. Specifically, Tab. 8 presents the QA generation performance across different LLM backbones. Tab. 9 and Tab. 10 detail the intermediate retrieval and final QA performance across different retrievers, respectively. Fig 10 illustrates the generation performance of all queries with different retrievers. We observe similar analysis as in the main text.

To further substantiate *Finding I* and the initial trends observed in Fig 2(c) of the main text, Fig 11 illustrates the retrieval performance across different session length percentiles using three datasets and three dense retrievers. The results reveal a consistent phenomenon: as session length increases, the performance of the Turn-level baseline (which relies on mean-pooling) degrades significantly, particularly in the longest quartile (75%-100%). This confirms that accumulating extended conversational history inherently dilutes the core matching signals. Conversely, our proposed retrieval mechanism maintains robust performance across all length buckets. By retaining the single most relevant turn, our approach effectively neutralizes the context dilution effect, demonstrating strong generalizability across varying dialogue lengths and retriever architectures.

Table 9: Retrieval Performance Comparison of Different Retrievers on LoCoMo with `gpt-4o-mini-2024-07-18`.

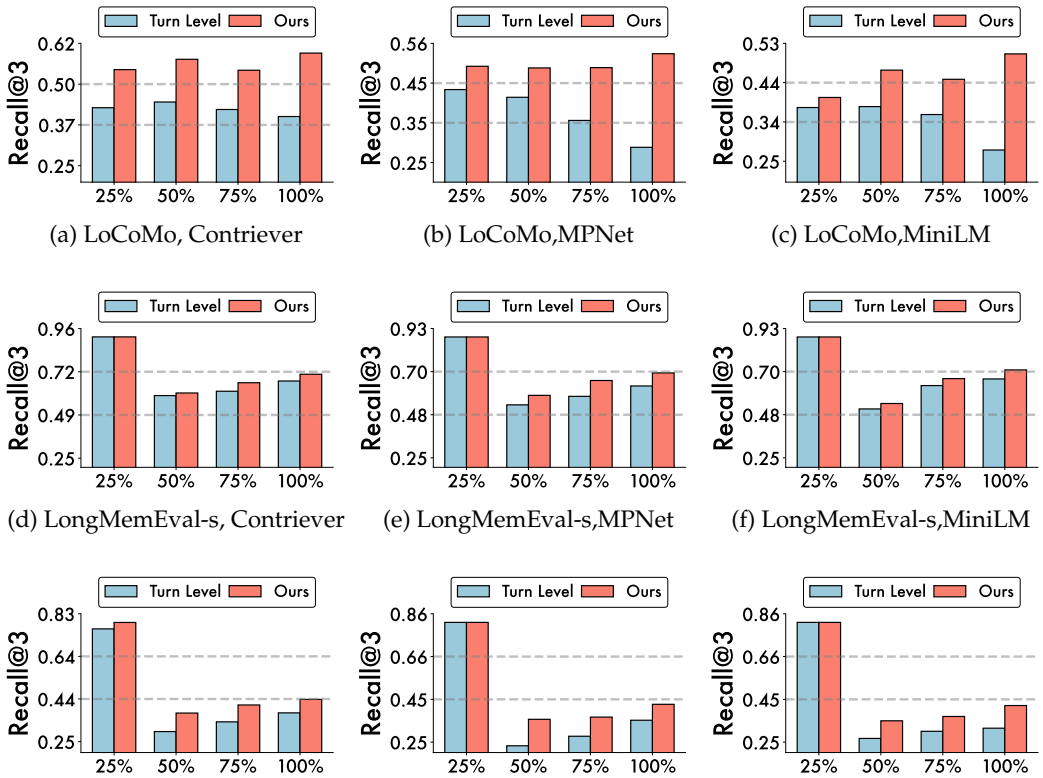
Model	Recall@3	NDCG@3	Recall@5	NDCG@5	Recall@10	NDCG@10
Contriever						
Contriever (2021)	49.95	52.15	58.31	56.29	71.80	60.91
MPC (2023)	53.37	55.34	57.75	57.33	66.52	60.28
RecurSum (2025)	47.23	48.99	59.01	54.58	74.97	60.07
SeCom (2025)	52.97	54.03	64.05	59.37	78.25	64.28
MemGAS (2026)	56.85	59.22	67.98	64.25	81.22	68.71
Nano-Memory	**69.39**	**71.19**	**76.33**	**74.58**	**86.46**	**77.99**
MPNet						
MPNet (2020)	45.92	47.71	53.98	51.79	68.58	56.88
MPC (2023)	45.47	47.35	54.08	51.68	68.28	56.59
RecurSum (2025)	49.50	51.15	59.47	56.16	76.64	61.99
SeCom (2025)	41.84	42.98	53.93	48.80	72.21	55.02
MemGAS (2026)	53.47	54.96	63.49	59.95	79.25	65.11
Nano-Memory	**62.74**	**64.31**	**70.80**	**68.31**	**82.48**	**72.21**
MiniLM						
MiniLM (2020)	34.79	35.60	41.84	39.15	56.55	44.59
MPC (2023)	42.30	43.59	51.21	48.08	68.03	54.03
RecurSum (2025)	44.76	46.82	54.73	51.64	72.16	57.52
SeCom (2025)	39.98	40.99	51.41	46.64	70.19	53.31
MemGAS (2026)	46.63	47.86	57.25	53.14	72.36	58.30
Nano-Memory	**57.45**	**59.16**	**67.17**	**63.67**	**80.61**	**68.19**

Table 10: QA Performance Comparison of Different Retrievers on LoCoMo with `gpt-4o-mini-2024-07-18`.

Model	4o-J	F1	BLEU4	ROUGE1	ROUGE2	ROUGEL	BERTScore
Contriever							
Contriever (2021)	40.33	15.76	2.77	16.08	7.75	15.10	84.70
SeCom (2025)	43.45	15.28	3.12	17.16	8.52	16.01	84.84
MemGAS (2026)	41.07	17.66	3.61	18.00	8.93	16.99	85.13
Nano-Memory	**48.84**	**22.66**	**5.33**	**22.92**	**12.04**	**21.68**	**85.94**
MPNet							
MPNet (2020)	38.07	14.52	2.36	14.97	6.82	13.84	84.46
SeCom (2025)	37.30	14.83	2.83	15.36	7.33	14.27	84.54
MemGAS (2026)	37.71	17.43	3.61	17.83	8.84	16.81	85.09
Nano-Memory	**45.67**	**21.00**	**4.86**	**21.29**	**11.49**	**20.12**	**85.61**
MiniLM							
MiniLM (2020)	29.05	12.76	1.75	11.72	5.34	12.02	84.05
SeCom (2025)	32.20	13.78	2.35	14.27	6.58	13.28	84.29
MemGAS (2026)	32.38	15.92	3.09	16.47	7.82	15.41	85.02
Nano-Memory	**43.45**	**20.09**	**4.68**	**20.39**	**10.62**	**19.24**	**85.47**

B.3 Ablation and Mechanism Analysis

We provide complete ablation results on four long-term conversational benchmarks, as presented in Tab. 11 and the different query ablation in Fig 12. We observe similar analysis as in the main text.



(a) LoCoMo, Contriever (b) LoCoMo,MPNet (c) LoCoMo,MiniLM

(d) LongMemEval-s, Contriever (e) LongMemEval-s,MPNet (f) LongMemEval-s,MiniLM

(g) LongMemEval-m, Contriever (h) LongMemEval-m,MPNet (i) LongMemEval-m, MiniLM

Figure 11: Recall@3 performance comparison across LoCoMo, LongMemEval-s, LongMemEval-m using Contriever,MPNet,MiniLM. We rank all sessions by their total turn counts and group them into four equal percentiles (0–25%, 25%–50%, 50%–75%, 75%–100%).

Table 11: Ablation Study. Contriever is the retrieval backbone, with `gpt-4o-mini-2024-07-18` as the generator. *R* and *G* represents our retrieval and generation mechanism, respectively.

Model	**4o-J**	**F1**	**BLEU4**	**ROUGE1**	**ROUGE2**	**ROUGEL**	**BERTScore**	**Avg. Tokens**
LoCoMo								
Baseline	40.33	15.76	2.77	16.08	7.75	15.10	84.70	2,348
+ *R*	43.42	19.62	3.85	19.86	10.43	18.68	85.33	2,685
+ *R* + *G*	**48.84**	**22.66**	**5.33**	**22.92**	**12.04**	**21.68**	**85.94**	1,403
LongMTBench+								
Baseline	63.54	31.82	7.51	33.42	18.91	26.53	87.01	12,045
+ *R*	62.15	36.77	11.31	38.77	21.69	29.70	87.95	12,105
+ *R* + *G*	**64.15**	**42.40**	**16.83**	**44.55**	**25.45**	**35.57**	**89.34**	6,174
LongMemEval-s								
Baseline	55.40	13.37	2.15	13.90	6.72	12.43	83.64	8,471
+ *R*	56.00	14.29	2.20	14.83	7.21	13.45	83.85	8,439
+ *R* + *G*	**57.20**	**21.06**	**4.89**	**21.80**	**10.77**	**20.16**	**85.54**	4,302
LongMemEval-m								
Baseline	42.80	11.88	1.66	12.63	5.51	11.11	83.31	8,274
+ *R*	46.00	12.86	1.92	13.54	6.30	12.15	83.60	8,357
+ *R* + *G*	**46.60**	**18.09**	**3.96**	**18.88**	**9.03**	**17.40**	**85.10**	4,261

C Prompts

The following query-aware pruning prompt selectively retains content relevant to the input query while preserving original token fidelity.

Question Answer

You are an intelligent dialog bot. You will be shown a "Fused Historical Event" which contains all the consolidated details relevant to your question. Filter the information to extract only the parts directly relevant to the Question. Preserve original tokens, do not paraphrase. Remove irrelevant turns, redundant info, and non-essential details.
Fused Historical Event: <fused_event>
Question Date: <question date>
Question: <question>
Answer:

For QA task, we follow (Xu et al., 2026) and prompt the model with the following template to produce a concise, history-based response.

Question Answer

You are an intelligent dialog bot. You will be shown History Dialogs. Please read, memorize, and understand the given Dialogs, then generate one concise, coherent and helpful response for the Question.
History Dialogs: <retrieved texts>
Question Date: <question date>
Question: <question>

For evaluation, we prompt GPT-4o with the template from (Zheng et al., 2023) to judge whether the model's output matches the correct answer, as detailed below.

GPT-4o-as-Judge

I will give you a question, a reference answer, and a response from a model. Please answer <yes> if the response contains the reference answer. Otherwise, answer <no>. If the response is equivalent to the correct answer or contains all the intermediate steps to get the reference answer, you should also answer <yes>. If the response only contains a subset of the information required by the answer, answer <no>.
User Question: <question>
Correct Answer: <answer>
Model Response: <response>
Is the model response correct? Answer <yes> or <no> only.

D More Discussion

D.1 Current Limitation

While our minimalist memory framework simplifies the retrieval process by effectively leveraging raw conversational history, it remains subject to the inherent limitations of non-parametric memory systems, specifically regarding privacy, interpretability, and security risks. Long-term agentic memory inevitably stores persistent and potentially sensitive user-specific content, including interaction histories and behavioral traces. Although retrieving such information is essential for generating personalized, high-quality responses, it inherently introduces risks of data leakage and excessive information retention. Furthermore, current systems are constrained by limited interpretability, lacking robust diagnostic

tools to precisely trace which specific memory entries were retrieved and how they directly influenced the large language model's generative process.

D.2 Future Direction

Future research will explore two directions. First, to transition from a passive retrieval framework to a proactive memory system, memory management should be integrated into the LLM's generation process via outcome-driven Reinforcement Learning (Yue et al., 2026; Lei et al., 2025a;b). Fine-tuning the agent (Yu et al., 2025; Wu et al., 2025) to autonomously emit structured memory operations (e.g., add, update, delete) during inference. Second, the integration of privacy-preserving mechanisms such as differential privacy or selective forgetting is essential to ensure data security. These advancements will be pivotal for maintaining user trust while enhancing the long-term utility of memory systems.

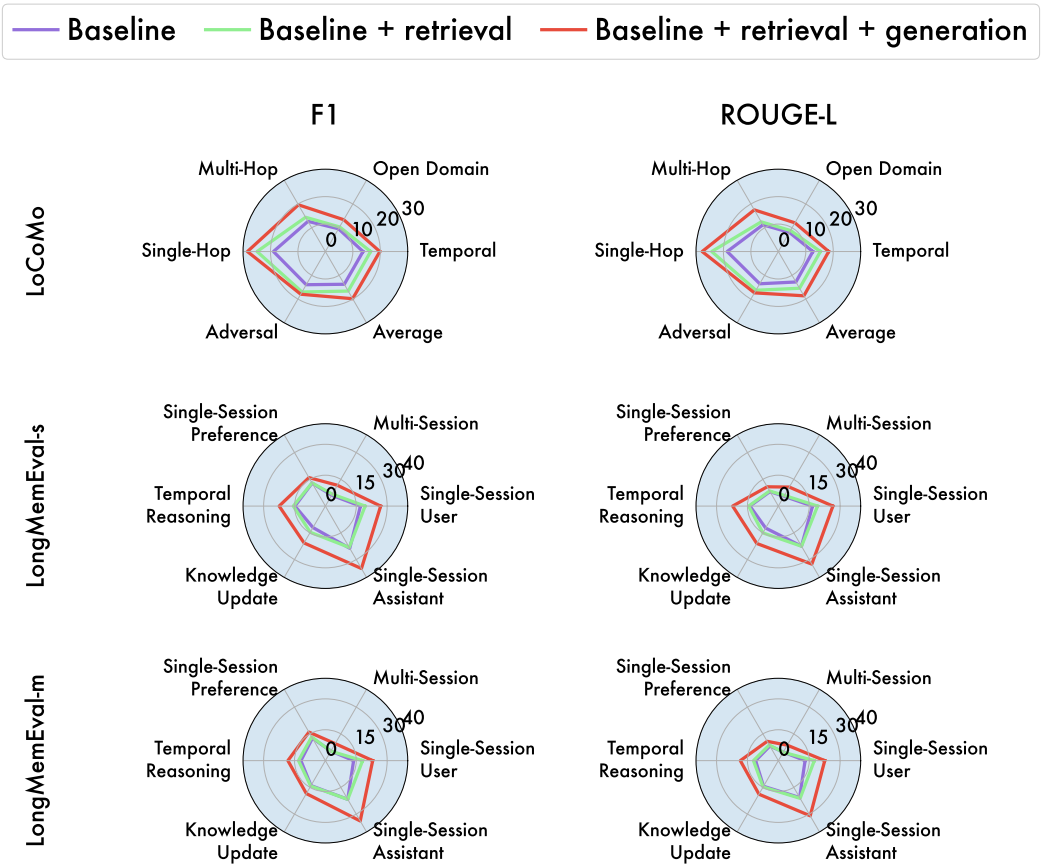


Figure 12: Ablation study on different query settings using *gpt-4o-mini-2024-07-18* and Contriever.

E LLM Disclosure

Following the conference guidelines regarding Generative AI models (GenAI, include large language model, a.k.a. LLM), we disclose that GenAI was utilized solely to improve sentence clarity and grammatical correctness. All aspects of conceptual development, experimental methodology, data analysis, and core manuscript content were independently produced by the authors without GenAI assistance.