

Evaluating Memory Capability in Continuous Lifelog Scenario

Jianjie Zheng^{1,2*}, Zhichen Liu^{1,2*}, Zhanyu Shen^{2,4}, Jingxiang Qu^{2,5}
Guanhua Chen^{1†}, Yile Wang⁴, Yang Xu¹, Yang Liu³, Sijie Cheng^{2,3†}

¹Southern University of Science and Technology, ²RayNeo.AI

³Tsinghua University, ⁴Shenzhen University, ⁵Shanghai Jiao Tong University

Abstract

Nowadays, wearable devices can continuously lifelog ambient conversations, creating substantial opportunities for memory systems. However, existing benchmarks primarily focus on online one-on-one chatting or human-AI interactions, thus neglecting the unique demands of real-world scenarios. Given the scarcity of public lifelogging audio datasets, we propose a hierarchical synthesis framework to curate **LIFEDIALBENCH**, a novel benchmark comprising two complementary subsets: **EgoMem**, built on real-world egocentric videos, and **LifeMem**, constructed using simulated virtual community. Crucially, to address the issue of temporal leakage in traditional offline settings, we propose an **Online Evaluation** protocol that strictly adheres to temporal causality, ensuring systems are evaluated in a realistic streaming fashion. Our experimental results reveal a counterintuitive finding: current sophisticated memory systems fail to outperform a simple RAG-based baseline. This highlights the detrimental impact of over-designed structures and lossy compression in current approaches, emphasizing the necessity of high-fidelity context preservation for lifelog scenarios.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across a wide range of tasks (OpenAI, 2022; OpenAI et al., 2024; Yang et al., 2025a), especially in the single-turn scenario with short-term conversational context. Subsequently, LLMs show superior reasoning ability as automatic agents to process a series of complex tasks in real world (Schick et al., 2023; Yang et al., 2023), meanwhile, place a higher requirement on the context length. To explore long-term memory capability of LLMs, one line of works (Chen et al.,

* Equal contribution. Work done during an internship at RayNeo AI.

† Corresponding authors.

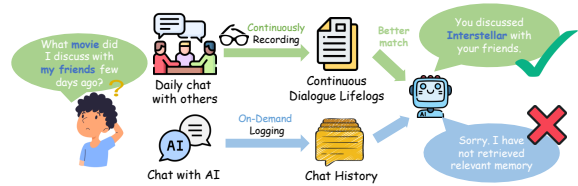


Figure 1: Comparison between (1) The microphone-always-on scenario, which continuously recording dialogue with others in daily life, and (2) Chatting with AI scenario, which on-demand logging to form the chat history.

2024; Grattafiori et al., 2024; Yang et al., 2025a) focus on probing the accuracy of locating evidence in extremely long-context passages, such as Needle In A Haystack (NAIH, 2025). However, the strategy of increasing context length indefinitely is not a solution to long-term memory, due to the exponential growth in inference costs and the ability of long-term memory utilization (Hsieh et al., 2024; Li et al., 2024; Liu et al., 2024). Consequently, the development of memory system has emerged. It requires LLMs to adaptively remember and retrieve relative evidence from massive information over extended periods.

Meanwhile, there exist various benchmarks primarily focus on dialogue scenarios, covering Person-AI interaction (Jiang et al., 2025; Wu et al., 2024) and dyadic dialogues (i.e., person-person conversations; Maharana et al. (2024)). However, the above-mentioned studies neglect a promising scenario as illustrated in Figure 1: *continuous dialogue lifelogs*. Nowadays, there emerges a series of commercial wearable devices with potential to achieve microphone always-on, such as smart glasses (e.g., Ray-Ban Meta, RayNeo V3/X3, Xiaomi AI Glasses), and recording machines (e.g., Plaud). Equipped with these wearable devices, users can continuously record the surrounding audio which fully filled with intensive dialogue content. Using automatic speech recognition (ASR), the audio stream is transcribed into text and stored

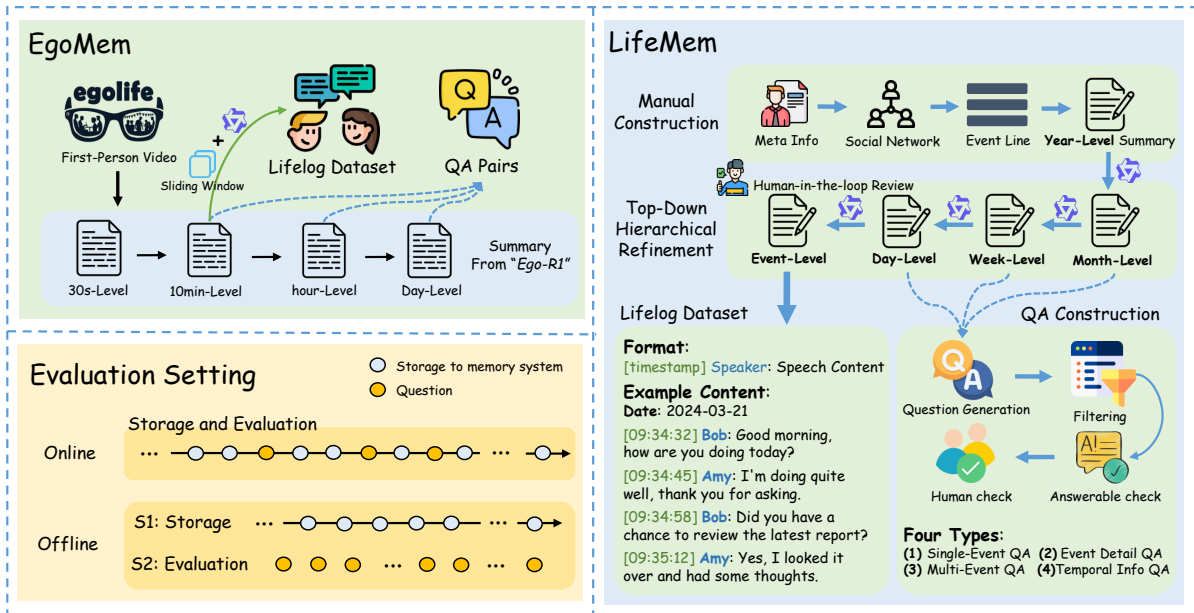


Figure 2: We introduce LIFEDIALBENCH, which consists of two subsets: EgoMem (top left), constructed from real-world egocentric videos (EgoLife), and LifeMem (right), a more comprehensive dataset built upon a Human-in-the-loop Hierarchical Life-Simulation Framework. Additionally, we propose a novel online evaluation method that assesses model performance incrementally during data storage, in contrast to conventional evaluation conducted only at the end of storage.

in a long-memory database after post-processing. Compared to prior passages and Person-AI interaction, continuous dialogue lifelogs have several unique characteristics: (1) The daily conversations integrate multi-person interactions, casual and temporal event threads, and simulated social networks. (2) Through round-the-clock recording, the lifelogs enables the AI assistant to accumulate an extensive understanding of users’ facts, perfectly embodying the highly promising usage scenario of a personalized assistant.

To systematically evaluate the long-term memory capacity of agents in continuous dialogue lifelogs, we introduce LIFEDIALBENCH, which comprises two complementary subsets as shown in Figure 2: **EgoMem** and **LifeMem**. Both subsets adopt a hierarchical life simulation framework to generate the dataset. EgoMem is constructed using a bottom-up (i.e., from second to week) summarization based on the real-life first-person video dataset EgoLife (Yang et al., 2025b), which records egocentric video from six individuals over a seven-day period. To extend the temporal span and ensure long-horizon coherence, we further use LLMs with a top-down (i.e., from year to day) elaboration to simulate a year-long personal lifelog rich in multi-party conversations, forming the LifeMem. For both subsets, we generate QA pairs from multi-

level event summaries, enabling systematic probing of memory retrieval across different temporal granularities. Notably, we first propose an **online evaluation** protocol that follows the linear flow of time with information updates and conflicts, offering a more realistic assessment of long-term memory in real-world conditions.

We evaluate four representative memory systems on LIFEDIALBENCH. Unexpectedly, a simple RAG baseline consistently outperforms specialized methods, highlighting the advantage of raw text preservation over the lossy compression inherent in current designs. Furthermore, we identify temporal retrieval as a universal bottleneck across all methods. Crucially, our analysis validates the necessity of online evaluation, demonstrating that traditional offline settings distort assessment by permitting temporal leakage and misjudging dynamic memory updates. These findings collectively underscore the dual importance of context fidelity and strict temporal linearity in next-generation lifelog systems.¹

2 Related Work

Memory Systems. Modern long-term memory systems for large language models (LLMs) typi-

¹We release our code and data at <https://github.com/RayNeo-AI-2025/LifeDialBench>.

Table 1: Comparison of memory benchmarks. The key properties are summarized including the type of scenario (**Scenario**), the temporal coverage (**Time Span**), number of sessions (**#Sessions**), whether continuous recording is supported (**Cont. Rec.**), whether the queries contain explicit timestamp for simulation (**TS**), whether support for online evaluation (**Online**).

| Benchmark | Scenario | Time Span | #Sessions | Cont. Rec. | TS | Online |
|---------------------------------|---------------|------------|-----------|------------|----|--------|
| LoCoMo (Maharana et al., 2024) | Person-Person | Few months | 1k | ✗ | ✗ | ✗ |
| MemoryBank (Zhong et al., 2024) | Person-AI | 10 days | 300 | ✗ | ✓ | ✗ |
| LongMemEval (Wu et al., 2024) | Person-AI | N/A | 50k | ✗ | ✓ | ✗ |
| MemBench (Tan et al., 2025) | Person-AI | N/A | 65k | ✗ | ✗ | ✗ |
| EgoMem (Ours) | Multi-Person | 7 days | 1.7k | ✓ | ✓ | ✓ |
| LifeMem (Ours) | Multi-Person | 1 year | 3.8k | ✓ | ✓ | ✓ |

cally adopt a modular architecture to handle extended contexts. These systems generally consist of a memory manager for database maintenance, a summary agent for information compression (Xu et al., 2025), and a retriever for context-aware fetching. Various implementations have been explored to optimize this framework. One prevailing approach involves using a Summary Agent to condense historical interactions into concise insights before storage in vector databases (Wang et al., 2025; Xu et al., 2025; Chhikara et al., 2025). To better capture structural dependencies, recent works such as Chhikara et al. (2025), Gutiérrez et al. (2025), and Rasmussen et al. (2025) have transitioned from flat vector embeddings to graph-based representations. Furthermore, inspired by traditional computing, works like Packer et al. (2023) and Li et al. (2025) treat LLM context as a tiered memory hierarchy, managing data through mechanisms analogous to operating systems. A systematic categorization of these architectural paradigms is provided in Figure 7 in the Appendix.

Benchmarks for Long-Term Memory. The rapid evolution of memory systems has spurred the development of specialized evaluation benchmarks, as summarized in Table 1. Early efforts like MemoryBank (Zhong et al., 2024) utilized manually constructed QA pairs for basic retrieval testing. Subsequent benchmarks increased complexity by focusing on session-based dynamics; for instance, LoCoMo (Maharana et al., 2024) evaluates person-person dialogues across multiple dimensions, while PersonaMem (Jiang et al., 2025) and LongMemEval (Wu et al., 2024) scale the context up to 1.5M tokens for person-AI interactions. They have made significant progress for chatbot-like memory system. However, remaining a gap be-

tween real world scenarios—the scenarios of multi-person communication, and the situation where the memory system is continuously activated, such as an always-on personal agent. In this paper, we propose a benchmark contains multi-person dialogue, which is rolling from day to night, and continuous for year-long. This benchmark behaves closer to the real world compared to previous works.

3 Benchmarking Lifelog Memory

To explore the long-term memory capacity of agents in continuous dialogue lifelogs, we specifically construct a benchmark named **LIFEDIALBENCH** with two complementary subsets for egocentric memory (Cheng et al., 2024), as illustrated in Figure 1. The first subset, named **EgoMem**, is constructed based on the existing Ego-Life dataset (Yang et al., 2025b) which contains daily video recording across seven days. Moreover, to mimic the continuous dialogue lifelogs in real-life with more time span and scene diversity, we further adopt data synthesis to construct a year-long subset, named **LifeMem**. Both subsets are constructed with a hierarchical life simulation framework, where we use bottom-up and top-down manners due to different data sources. Specifically, all dialogue and summary content are generated using Qwen3-235B-Instruct².

More details will be introduced in this section.

3.1 Design Principles

Constructing a rigorous benchmark for lifelog memory evaluation requires confronting challenges that are qualitatively distinct from those in standard dialogue settings. We identify three core principles that govern the design of LIFEDIALBENCH.

²Corresponds to Qwen3-235B-A22B-Instruct-2507

Temporal Causality. In realistic lifelogging, an agent must respond to a query using *only* information available up to that moment—future context is physically inaccessible. Traditional offline evaluation violates this constraint by granting agents access to the complete dataset prior to answering any query, introducing *temporal leakage* that systematically overestimates real-world performance. This motivates a causally-constrained **Online Evaluation Protocol**, formally defined in §3.6, which enforces a streaming interaction flow strictly aligned with the physical arrow of time.

Compositional Query Complexity. Lifelog queries extend well beyond simple fact retrieval. They demand (i) *Temporal Grounding*—localizing *when* an event occurred relative to the query moment—and (ii) *Long-Horizon Multi-Hop Reasoning*—synthesizing disjoint events dispersed across days or months to form a coherent answer. Benchmarks built on isolated dialogue sessions fail to stress-test these capabilities. Accordingly, LIFEDIALBENCH incorporates four structured question types (*single_event*, *multi_event*, *time_query*, *event_detail*) to systematically probe both dimensions, as detailed in §3.4.

Ecological Validity of the Lifelog Stream. Authentic lifelogging differs from mere session concatenation along two critical axes. First, the data stream is *multi-party*: interactions span diverse social contexts—family members, colleagues, and strangers—rather than a fixed dyadic pair, introducing speaker heterogeneity absent from prior benchmarks. Second, the timeline is *semantically coherent*: topics evolve, recur, and interleave organically across time rather than resetting at session boundaries. Together, these properties demand data that preserves genuine chronological continuity and multi-speaker dynamics, which directly guides the hierarchical construction methodology described as follows.

3.2 EgoMem

Data Source. We construct EgoMem based on the Ego-R1 corpus (Tian et al., 2025), which provides multi-scale textual summaries of the real-world egocentric dataset EgoLife (Yang et al., 2025b). Initially, we attempted to transcribe EgoLife’s raw audio via Automatic Speech Recognition (ASR). However, the inherent sparsity of continuous conversation and severe ASR degradation in noisy daily environments made it impractical

to extract coherent transcripts necessary for constructing high-quality Q&A pairs. To overcome this, we propose a hybrid proxy approach: we leverage Ego-R1’s structured 10-minute summaries as factual anchors to prompt an LLM to synthesize lifelog-style dialogues. This strategy produces *summary-grounded simulated dialogues* that preserve the authentic physical event sequences of the real world while yielding contextually rich conversational streams suitable for rigorous memory evaluation.

Data Curation. To maintain narrative coherence across extended temporal horizons, we employ a *sliding-window generation strategy* at a 10-minute granularity. Rather than generating isolated dialogue segments, the LLM is conditioned on a dynamic historical context window (the preceding 60 minutes) alongside the target summary. This empirically determined window balances speaker consistency, relational continuity, and topic flow while mitigating hallucinations and repetitive patterns.

Quality Assurance & Conversational Grounding. To ensure our synthesized data reflects the natural dynamics of daily communication rather than overly formalized written prose, we implement a human-LLM collaborative review pipeline (Yuksekgonul et al., 2024). Beyond standard *Factual Consistency* checks against the source summaries, we specifically optimize for *Conversational Naturalness*. Human annotators utilize LLM-flagged feedback to iteratively revise the text, ensuring the dialogues capture casual interactions, appropriate tone, and implicit contexts typical of real-world continuous recordings. Comprehensive details regarding ASR limitations, granularity selection, and the full review workflow are provided in Appendix D.

3.3 LifeMem

While EgoMem provides grounded real-life recordings, it is constrained by a seven-day window and limited social diversity. To overcome these barriers and enable the study of long-term memory at scale, we develop a **Human-in-the-loop (HITL) Hierarchical Life-Simulation Framework** to synthesize LifeMem—a year-long continuous dialogue lifelog. Diverging from conventional end-to-end synthesis, our framework treats lifelog synthesis as a principled, multi-stage trajectory expansion process

where humans steer the narrative consistency and logical grounding.

Hierarchical Synthesis Framework. The core of LifeMem is a top-down refinement strategy that decomposes a person’s entire year into granular, dialogue-centric records across four stages:

1) Identity and Social Seeding: We first manually define a virtual agent’s persona (e.g., age, occupation, personality) and a comprehensive social network encompassing family, colleagues, and friends. Human curators serve as the system architects, validating the realism of these relationships to ensure a stable social foundation for long-term interactions. This social network serves as the basis for generating multi-party conversations that reflect realistic interpersonal dynamics.

2) Multi-dimensional Event Trajectories: Rather than generating random experiences, we construct eleven distinct *event lines* (e.g., professional growth, health, household management) across key life dimensions. These event lines serve as structured narrative threads that capture diverse and realistic life dynamics, establishing a solid basis for generating lifelog data that maintains semantic richness and long-horizon coherence. We then align all event lines into a *year-level summary*, which serves as the backbone of the year, ensuring that an event mentioned in January (e.g., a project kickoff) has logical echoes in subsequent months.

3) Iterative Refinement (Allocation & Enrichment): The framework progressively expands high-level annual summaries into monthly, weekly, and daily scales through a top-down refinement strategy. This involves two sub-steps: (i) *Allocation*: LLMs distribute annual narrative into monthly and weekly placeholders. (ii) *Enrichment*: Each placeholder is expanded into detailed narratives while preserving higher-level context. To further align the simulated lifelog with real-world temporal structures, we incorporate external calendar signals such as statutory holidays, weekends, and workdays. Notably, directly generating lifelogs from daily experience often leads to coarse or repetitive descriptions, as LLMs struggle to capture the fine-grained variations that naturally arise within a day; our two-stage refinement effectively mitigates this issue.

4) Dialogue Grounding: Daily experiences are further segmented into *event-level narratives*, each annotated with temporal boundaries, locations, and participants. Finally, we synthesize continuous dialogue streams conditioned on these fine-grained

event contexts, the virtual user’s background, and the social relationship network, ensuring natural conversational flow and long-horizon coherence.

Human-in-the-Loop Calibration. Crucially, our framework operates as a transparent and controllable pipeline; it integrates human intervention at every hierarchical transition to mitigate the “stochastic parrot” effect and temporal drift. As detailed in the calibrated workflow (Appendix B):

- **Consistency Auditing:** After each refinement stage (e.g., Year-to-Month), human annotators audit the generated trajectories. They identify and correct temporal contradictions (e.g., a character being in two places at once) or identity inconsistencies.
- **Manual Trajectory Pruning:** Annotators prune repetitive or mundane event patterns that frequently emerge in vanilla LLM outputs, ensuring the lifelog maintains high information density and narrative diversity.
- **Quality Gatekeeping:** Data only advances to the next level of granularity (e.g., Week-to-Day) once it passes a human-verified quality check. This prevents error propagation, ensuring that the final dialogues are grounded in a logically sound and human-validated annual history.

By combining automated hierarchical expansion with rigorous human steering, LifeMem achieves a balance of scalability and high-fidelity realism, providing a robust foundation for benchmarking memory systems in privacy-preserving, always-on scenarios.

3.4 Question-Answering Pairs Curation

Task Formats. LIFEDIALBENCH supports two complementary question-answering formats: *Multiple-Choice Questions (MCQ)* and *Open-Ended QA*. MCQ strictly assesses precise retrieval capacity, while Open-Ended QA—where reference answers are derived from the correct MCQ options—evaluates the agent’s generative synthesis capability via LLM-as-a-Judge (Zheng et al., 2023).

Question Types. To comprehensively evaluate memory retrieval and reasoning (Maharana et al., 2024), we design four distinct question types (consistent with labels in Figure 2): (i) *QT1: Single-Event QA*: Retrieving core event content based on ambiguous queries. (ii) *QT2: Event Detail QA*:

Table 2: The number of generated questions, the number of questions remained after filtering and human annotation, and the model accuracy in the final answerable verification.

| LifeMem | Total | Filtered | Keep Rate | Model Acc. |
|---------|-------|----------|-----------|------------|
| Daily | 1464 | 1430 | 97.68% | 99.23% |
| Weekly | 248 | 241 | 97.18% | 98.76% |
| Monthly | 48 | 46 | 95.83% | 100% |
| All | 1760 | 1717 | 97.56% | 99.18% |

Pinpointing specific snippet-level event attributes. (iii) *QT3: Multi-Event QA*: Retrieving and integrating information across multiple disconnected events over time. (iv) *QT4: Temporal Info QA*: A lifelog-specific type requiring the exact timestamp or temporal relation of a given event.

Question-Answering Construction. We prompt Qwen3-235B-Thinking³ (Yang et al., 2025a) to synthesize QA pairs spanning multiple temporal granularities. To ensure rigorous evaluation, we implement a strict quality assurance pipeline.

1) Causality Preservation: All distractors are generated exclusively using events prior to the query timestamp, preventing agents from trivially eliminating “future” options during online streaming.

2) Leakage Filtering: A secondary Qwen3-32B rewrites questions to eliminate explicit timestamp leakages, enforcing semantic understanding over temporal shortcuts.

3) Answerability Check: We prompt qwen3-max⁴ to answer the generated questions. To mitigate random guessing (a 25% chance in MCQ), the evaluator must output explicit evidence rationales before selecting an option. Unanswerable questions are discarded, and correct ones undergo human random sampling to mitigate LLM self-preference bias.

Table 2 details the filtering statistics and near-perfect evaluator accuracy for LifeMem, confirming the genuine answerability of the curated queries. Ultimately, the benchmark retains 1,774 questions for EgoMem and 1,717 for LifeMem. Detailed prompts are in Appendix G.3.

3.5 Dataset Statistics

Figure 3 illustrates the distributional characteristics of LifeMem across four dimensions: event types, social roles, locations, and monthly dialogue

counts. The dataset covers both routine and higher-level life pursuits, spanning intimate, professional, and casual social interactions across diverse geographical settings. Crucially, monthly event counts remain stable throughout the year with no significant seasonal bias, which prevents evaluation results from being skewed toward time-specific patterns—a critical property for long-horizon memory benchmarking.

For QA pairs, Figure 6 in appendix H.2 shows the proportion of each question type after filtering. The four types—event content recall, event detail retrieval, multi-hop event reasoning, and temporal information QA—are distributed as 25.3/25.0/25.0/24.6 in LifeMem and 25.1/25.1/24.9/24.9 in EgoMem, ensuring a balanced evaluation of memory systems across multifaceted retrieval and reasoning capabilities.

3.6 Online Evaluation Protocol

To rigorously assess memory systems in a realistic “always-on” setting, we propose a strict **Online Evaluation Protocol**. Unlike traditional offline paradigms (Maharana et al., 2024; Wu et al., 2024) that batch-process queries after indexing the entire dataset—thereby risking temporal leakage—our protocol enforces a linear, streaming interaction flow that strictly adheres to the physical arrow of time.

Formally, we model the lifelog as a time-ordered stream of data chunks $\mathcal{D} = \{d_1, d_2, \dots, d_T\}$, where each d_t corresponds to a discrete time window (e.g., a dialogue session or a fixed-length segment). A set of queries \mathcal{Q}_t is associated with each time step t , representing questions that become answerable only after observing d_t . As illustrated in Figure 2, the evaluation proceeds as a recursive sequential process:

1) Streaming Ingestion: At step t , the memory system receives and processes the new data chunk d_t . The system updates its internal memory state from \mathcal{M}_{t-1} to \mathcal{M}_t based on its specific retention and consolidation mechanisms (e.g., indexing, summarization, or graph updates).

2) State Freezing: Before revealing any future data d_{t+1} , the memory state \mathcal{M}_t is effectively “frozen.” This ensures a read-only evaluation snapshot where the system strictly cannot access future information, thereby eliminating look-ahead bias.

3) In-Stream Assessment: The system is tasked to answer all queries $q \in \mathcal{Q}_t$ using strictly the information available in the frozen state \mathcal{M}_t . The per-

³Corresponds to Qwen3-235B-A22B-Thinking-2507

⁴Corresponds to qwen3-max-2025-09-23

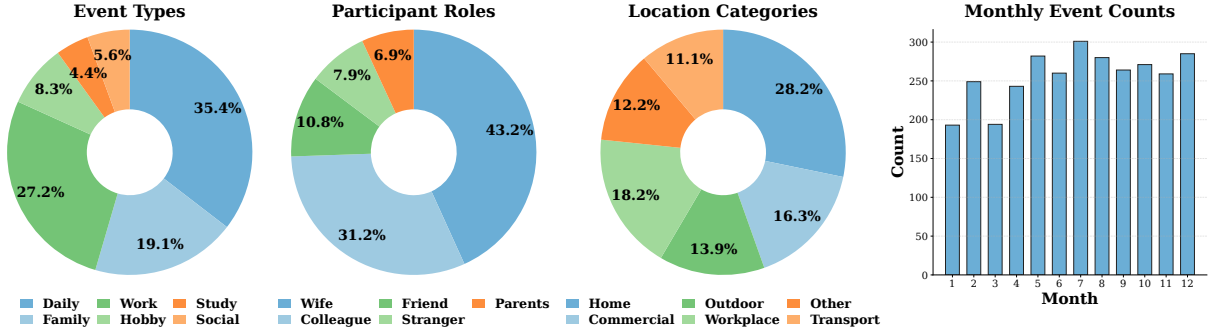


Figure 3: Distributional statistics of the LifeMem dataset. The plots summarize event types, social roles, locations, and monthly dialogue counts, showing that the dataset is balanced and closely aligned with real-world lifelog patterns.

formance is recorded for this specific timestamp.

4) Causal Progression: Only after all queries in Q_t are resolved does the system advance to time $t + 1$, repeating the cycle.

This protocol guarantees **Temporal Causality**: the answer a_t to a query $q \in Q_t$ depends solely on the available history $\mathcal{H}_t = \{d_1, \dots, d_t\}$. The necessity of this online mode stems from two fundamental flaws in traditional offline evaluation for lifelog scenarios:

- **Future Context Contamination:** In a causal online setting, the response to query q_t strictly relies on the current memory: $a_t^{\text{real}} = A(q_t | \mathcal{M}_t)$. Conversely, offline evaluation grants the system access to the complete history \mathcal{M}_T (where T is the end of the stream), yielding $a_t^{\text{offline}} = A(q_t | \mathcal{M}_T)$. Whenever $t < T$, future data (d_{t+1}, \dots, d_T) can illicitly influence the response through advanced indexing or global summarization, creating an uncontrolled confound and overestimating the system’s promptness.
- **Irreversible Memory Modification:** In many memory systems (especially those with overwrite or consolidation mechanisms), memories evolve dynamically: $\mathcal{M}_t = \text{Update}(\mathcal{M}_{t-1}, d_t)$. In the online mode, d_t is fresh and explicitly represented ($d_t \subseteq \mathcal{M}_t$). However, in offline mode, if d_t is compressed or purged by subsequent updates ($\exists \tau \in (t, T]$ causing the information in d_t to be lost in \mathcal{M}_T), the system will fail to answer q_t retrospectively. Thus, offline metrics fail to capture the system’s ability to provide real-time responses before information decay occurs, severely misestimating a model’s real-world dynamic retrieval capacity.

4 Experiments

4.1 Experimental Setup

To assess the performance of current mainstream memory systems in the context of continuous dialogue lifelogs and to derive insights, we select four representative memory systems for evaluation on LIFEDIALBENCH: (1) **RAG** (Lewis et al., 2021): A straightforward retrieval-augmented generation (RAG) baseline that directly stores and retrieves text chunks; (2) **A-Mem** (Xu et al., 2025): An enhanced variant of RAG that augments the retrieval process with additional semantic signals — such as context, tags, keywords, and links — to improve the representation and utilization of stored text; (3) **MemOS** (Li et al., 2025): A memory system that does not retain raw input, but instead abstracts each input segment into a single structured representation—including summaries, titles, and semantic tags; and (4) **Mem0** (Chhikara et al., 2025): A memory paradigm that similarly discards raw input, but extracts and stores multiple concise factual statements per input segment. More information about the implementation details of these methods can be found in Appendix I.

Experiments on EgoMem and LifeMem are conducted in an online setting using GPT-4o-mini⁵ and qwen-plus⁶

To further investigate the necessity of online evaluation, we also include Qwen3-8B in additional experiments. For embedding needs, we employ Qwen3-Embedding-8B.

We adopt two question-answering formats: multiple-choice and open-ended. For multiple-choice questions, correctness is determined by ex-

⁵Corresponds to GPT-4o-mini

⁶Corresponds to qwen-plus-2025-12-01

Table 3: Main results of memory systems’ performance on LIFEDIALBENCH in the online evaluation setting. The method with the best overall performance are **bold**, the second are underlined.

| Models | Method | Open-Ended | | | | | Multiple-Choices | | | | |
|----------------|--------|------------|-------|-------|-------|--------------|------------------|-------|-------|-------|--------------|
| | | QT1 | QT2 | QT3 | QT4 | Overall | QT1 | QT2 | QT3 | QT4 | Overall |
| EgoMem | | | | | | | | | | | |
| gpt-4o-mini | RAG | 39.11 | 57.91 | 10.91 | 26.57 | 34.07 | 73.79 | 86.25 | 60.69 | 57.20 | 69.86 |
| | A-Mem | 36.29 | 53.33 | 8.73 | 30.18 | <u>32.48</u> | 70.56 | 86.66 | 52.83 | 55.40 | <u>66.77</u> |
| | Mem0 | 15.72 | 17.91 | 3.49 | 16.21 | 13.41 | 56.85 | 53.75 | 44.54 | 37.83 | 48.56 |
| | MemOS | 26.61 | 33.75 | 6.55 | 11.71 | 20.02 | 69.35 | 72.91 | 66.37 | 38.73 | 62.30 |
| qwen-plus | RAG | 37.50 | 59.16 | 13.10 | 31.08 | <u>35.56</u> | 70.16 | 85.41 | 60.26 | 56.30 | 68.37 |
| | A-Mem | 43.54 | 58.75 | 14.41 | 33.33 | 37.91 | 71.37 | 86.66 | 55.89 | 55.40 | <u>67.73</u> |
| | Mem0 | 12.09 | 16.66 | 6.11 | 13.96 | 12.24 | 45.96 | 57.91 | 36.68 | 35.13 | 44.19 |
| | MemOS | 33.46 | 44.58 | 17.46 | 35.58 | 32.90 | 64.91 | 76.25 | 58.51 | 55.40 | 64.00 |
| LifeMem | | | | | | | | | | | |
| gpt-4o-mini | RAG | 36.86 | 73.25 | 20.41 | 23.12 | 38.46 | 74.42 | 92.09 | 69.93 | 56.76 | 73.63 |
| | A-Mem | 33.63 | 74.44 | 19.71 | 21.72 | <u>37.40</u> | 72.34 | 90.42 | 69.69 | 54.19 | <u>71.98</u> |
| | Mem0 | 7.36 | 21.47 | 5.39 | 8.64 | 10.72 | 67.96 | 69.90 | 68.06 | 36.67 | 60.97 |
| | MemOS | 23.21 | 56.87 | 11.39 | 13.95 | 26.20 | 74.32 | 80.83 | 69.19 | 31.67 | 64.00 |
| qwen-plus | RAG | 39.40 | 74.44 | 33.32 | 37.14 | <u>46.18</u> | 77.18 | 92.12 | 72.30 | 59.34 | 75.16 |
| | A-Mem | 43.78 | 78.97 | 38.25 | 36.67 | 49.54 | 75.34 | 92.12 | 67.60 | 63.31 | <u>74.51</u> |
| | Mem0 | 16.59 | 36.97 | 11.96 | 15.18 | 20.20 | 72.91 | 72.79 | 59.85 | 48.36 | 63.46 |
| | MemOS | 38.86 | 70.68 | 26.59 | 23.40 | 39.88 | 74.54 | 86.36 | 67.95 | 52.95 | 70.45 |

act matching. For open-ended questions, the reference answer is constructed from the correct option of the corresponding multiple-choice question. We then use Qwen3-32B as an evaluator to judge whether the generated response is semantically equivalent to the reference.

4.2 Main Results

To assess the performance of existing memory systems in the context of continuous dialogue lifelogs, we evaluate the four representative systems mentioned in §4.1 on LIFEDIALBENCH. Table 3 presents the full results. In the following, we analyze three aspects: (1) the importance of storing raw text, (2) the impact of compression level on performance, and (3) the difference in results between different question settings.

Importance of Raw Text Contrary to the prevailing trend favoring complex, structured memory architectures, our results reveal a counterintuitive finding in the lifelog domain: simple raw-text preservation (RAG, A-Mem) significantly outperforms sophisticated summarization-based paradigms (Mem0, MemOS).

Specifically, RAG and A-Mem consistently achieve the highest performance across both sub-

sets and backbone models (GPT-4o-mini and Qwen-Plus). While A-Mem represents an “enhanced” paradigm that augments raw text with auxiliary metadata (brief summaries and associations), it does not yield statistically significant gains over the vanilla RAG baseline. This suggests that for continuous dialogue streams, the fidelity of the original context is the dominant factor for retrieval success, rendering added structural complexity largely redundant.

Impact of Compression Level MemOS and Mem0 represent two memory approaches that compress raw text to varying extents. While MemOS applies lightweight summarization to the original dialogue, Mem0 performs fact-level compression, resulting in compression ratios of 62% and 35% in terms of token count, respectively. Our results show that MemOS, despite underperforming compared to the uncompressed baselines RAG and A-Mem, still substantially outperforms Mem0. This indicates that compressing raw text not only degrades performance, but also that the degree of compression correlates with the extent of performance loss.

Impact of Question Type By comparing system performance on four question categories, we find that Event Detail Retrieval (QT2) is the rel-

atively easiest task, followed by Event Content Recall (QT1), while Multi-hop Event Reasoning (QT3) and Temporal Grounding (QT4) prove to be the most challenging. The relative ease of QT2 is intuitive, as it merely requires retrieving specific details from the original text. QT1, by contrast, demands that the agent infer the occurred event from dialogue content, making it comparatively more difficult. The low performance on QT3 and QT4 suggests that multi-hop reasoning and temporal grounding remain the most formidable challenges in the context of continuous dialogue lifelogs.

Impact of QA Format Comparing the Open-Ended and Multiple-Choice settings reveals that questions are considerably more difficult in the Open-Ended format. Notably, while QT3 outperforms QT4 in the Multiple-Choice setting, this trend reverses in the Open-Ended setting. This suggests that without the guidance of candidate options, models struggle to aggregate and reason over disjoint evidence from the continuous stream, indicating that independent reasoning imposes higher demands on contextual fidelity.

5 Analysis of Online Evaluation

To quantify the "future context contamination" defined in section 3.6, we analyzed the correlation between the retrieval of future memories and answer correctness using RAG and A-Mem on LifeMem. We computed the AUROC scores for this relationship, obtaining **0.64** and **0.68**, respectively. These non-random values (> 0.5) demonstrate a dependency: the presence of future information materially distorts model performance. Whether this leakage acts as a "cheat" or "noise", it creates an uncontrolled confound that renders offline metrics unreliable for assessing real-world utility.

Mem0 exemplifies a memory system with overwrite-based updates. To demonstrate the impact of "irreversible memory modification". We identified a set of questions that Mem0 answers correctly under the Online setting but fails in the offline setting on LifeMem. Even after restricting the offline retriever to memories created before the query timestamp and increasing the retrieval budget (top- k) from 20 to 100, the accuracy on this subset remains only **34.91%**—far below 100%. This strongly suggests that the required information was likely overwritten during the offline memory construction process.

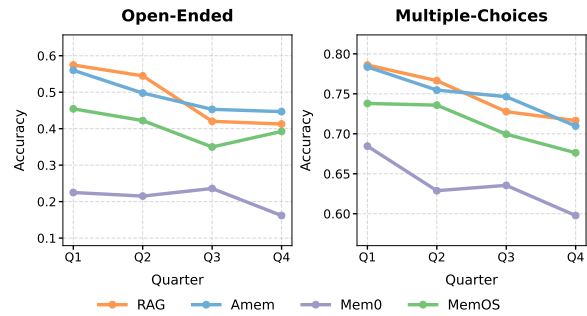


Figure 4: Comparison of accuracy decay rates across memory systems. RAG exhibits the steepest decline, whereas abstraction-based methods (A-Mem, MemOS) maintain better stability over long-term timelines despite lower absolute accuracy.

Accuracy Decay over Time. This is a characteristic feature of online evaluation. Figure 4 illustrates the quarterly accuracy of Qwen-Plus on LifeMem across different quarters. All systems exhibit a declining trend, as the memory pool continuously expands under the online setting, increasing the difficulty of retrieval. This phenomenon aligns with real-world scenarios.

6 Conclusion

In this paper, we bridge the gap between isolated dialogue sessions and continuous real-world existence by introducing **LIFEDIALBENCH**, a comprehensive benchmark comprising the real-world EgoMem and the synthesized year-long LifeMem. Crucially, we identify that traditional offline evaluation fundamentally violates the temporal causality of lifelogs, prompting the proposal of a rigorous Online Evaluation protocol.

Our extensive experiments yield a counter-intuitive yet critical finding: despite the growing complexity of agentic memory architectures (e.g., graph-based or summary-based systems), simple raw-text retrieval (RAG) consistently outperforms structured memory paradigms. Our analysis reveals that current abstraction mechanisms introduce "lossy compression," stripping away vital contextual details required for precise grounding in continuous streams. Furthermore, we demonstrate that offline metrics suffer from severe future context contamination, rendering them unreliable. By establishing these insights, **LIFEDIALBENCH** serves not only as a testbed but as a call to action for the community to rethink memory design—shifting focus from aggressive abstraction back to high-fidelity context preservation.

Ethics Statement

Wearable devices with always-on microphones are changing the balance between personal memory tools and unwanted monitoring. While continuous lifelogging offers new possibilities for personalized AI, it brings serious ethical questions around privacy, consent, and data protection.

Consent and Bystander Privacy. A primary concern of always-on lifelogging is the inadvertent capture of non-consenting third parties. Unlike voluntary human-AI interactions, continuous logs inherently record speech from colleagues, family, or strangers, raising significant legal and ethical issues under frameworks such as the GDPR and various wiretapping laws. LIFEDIALBENCH is constructed to sidestep these concerns: **EgoMem** draws from the publicly released and ethically vetted EgoLife dataset, while **LifeMem** employs human-in-the-loop simulation within a virtual community. For real-world deployment, we advocate for robust consent mechanisms, including visible recording indicators and granular opt-out affordances for bystanders.

Data Security and Sensitive Information. The temporal depth and granularity inherent in longitudinal lifelogging records, which encompass sensitive information from healthcare, personal finance, and social relationships, significantly amplifies the potential for adversarial behavioral profiling. To prevent unauthorized access to a user’s complete social graph or psychological state, a privacy-by-design architecture rooted in the edge-intelligence paradigm offers an effective solution. Raw audio capture can be processed entirely on-device within secure enclaves, with transcripts and vector embeddings confined to encrypted, sandboxed local storage. Furthermore, enforcing user-configurable data retention policies (e.g., rolling deletion of raw data within 24 hours) minimizes the persistent attack surface and renders centralized large-scale data exfiltration technically infeasible.

Limitations

We acknowledge three primary limitations in LIFEDIALBENCH.

First, the year-long LifeMem relies on simulation and lacks the stochasticity of raw ASR noise. However, we argue this abstraction is justifiable as it aligns with modern pipelines where ASR outputs typically undergo upstream refinement. Bench-

marking on high-fidelity text allows us to strictly isolate memory reasoning bottlenecks from perceptual errors, establishing a critical upper-bound baseline.

Second, our benchmark is currently restricted to the textual modality. The exclusion of visual cues limits the evaluation of tasks requiring visual grounding (e.g., locating objects), which we leave for future multimodal extensions.

Finally, the computational intensity of Online Evaluation limited our experiments to representative backbone models. This precluded an exhaustive exploration of emerging ultra-long-context models in no-retrieval settings, which remains an open avenue for future research.

References

- Yukang Chen, Shengju Qian, Haotian Tang, Xin Lai, Zhijian Liu, Song Han, and Jiaya Jia. 2024. [Longlora: Efficient fine-tuning of long-context large language models](#). *Preprint*, arXiv:2309.12307.
- Sijie Cheng, Kechen Fang, Yangyang Yu, Sicheng Zhou, Bohao Li, Ye Tian, Tingguang Li, Lei Han, and Yang Liu. 2024. Videogthink: Assessing egocentric video understanding capabilities for embodied ai. *arXiv preprint arXiv:2410.11623*.
- Prateek Chhikara, Dev Khant, Saket Aryan, Taranjeet Singh, and Deshraj Yadav. 2025. [Mem0: Building production-ready ai agents with scalable long-term memory](#). *Preprint*, arXiv:2504.19413.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, and Others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2025. Hipporag: neurobiologically inspired long-term memory for large language models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems, NIPS '24*, Red Hook, NY, USA. Curran Associates Inc.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekes, Fei Jia, and Boris Ginsburg. 2024. [RULER: What’s the real context size of your long-context language models?](#) In *First Conference on Language Modeling*.
- Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J. Taylor, and Dan Roth. 2025. [Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale](#). *Preprint*, arXiv:2504.14225.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). *Preprint*, arXiv:2005.11401.
- Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. [LooGLE: Can long-context language models understand long contexts?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333, Bangkok, Thailand. Association for Computational Linguistics.
- Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, and 1 others. 2025. [Memos: A memory os for ai system](#). *arXiv preprint arXiv:2507.03724*.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. [Evaluating very long-term conversational memory of LLM agents](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.
- Timur Mudarisov, Mikhail Burtsev, Tatiana Petrova, and Radu State. 2025. [Limitations of normalization in attention mechanism](#). *Preprint*, arXiv:2508.17821.
- NAIH. 2025. [GitHub - gkamradt/LLMTest_NeedleInAHaystack: Doing simple retrieval from LLM models at various context lengths to measure accuracy — github.com. https://github.com/gkamradt/LLMTest_NeedleInAHaystack](#). [Accessed 18-09-2025].
- OpenAI. 2022. [Chatgpt](#). Accessed: September 15, 2024.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, and Others. 2024. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. [Memgpt: Towards llms as operating systems](#). *CoRR*, abs/2310.08560.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. [Zep: a temporal knowledge graph architecture for agent memory](#). *arXiv preprint arXiv:2501.13956*.
- Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. [Toolformer: Language models can teach themselves to use tools](#). *Advances in Neural Information Processing Systems*, 36:68539–68551.
- Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. [Membench: Towards more comprehensive evaluation on the memory of llm-based agents](#). *Preprint*, arXiv:2506.21605.
- Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. [Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning](#). *arXiv preprint arXiv:2506.13654*.
- Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2025. [Scm: Enhancing large language model with self-controlled memory framework](#). *Preprint*, arXiv:2304.13343.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024. [Longmemeval: Benchmarking chat assistants on long-term interactive memory](#). In *The Thirteenth International Conference on Learning Representations*.
- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. [A-mem: Agentic memory for llm agents](#). *arXiv preprint arXiv:2502.12110*.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, and Others. 2025a. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, Bei Ouyang, Zhengyu Lin, Marco Cominelli, Zhonggang Cai, Bo Li, Yuanhan Zhang, Peiyuan Zhang, Fangzhou Hong, Joerg Widmer, and 3 others. 2025b. [Egolife: Towards egocentric life assistant](#). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28885–28900.
- Rui Yang, Lin Song, Yanwei Li, Sijie Zhao, Yixiao Ge, Xiu Li, and Ying Shan. 2023. [Gpt4tools: Teaching large language model to use tools via self-instruction](#). *Advances in Neural Information Processing Systems*, 36:71995–72007.
- Mert Yuksekogonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. [Textgrad: Automatic "differentiation" via text](#). *Preprint*, arXiv:2406.07496.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. [Judging](#)

llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [Memorybank: Enhancing large language models with long-term memory](#). In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2024, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press.

A LLM Usage

In the preparation of this paper, large language models (LLMs) were used solely as auxiliary tools. Specifically, we employed LLMs for grammar correction and text polishing, as well as to support dataset generation and assist in the manual review of data quality.

B Details of human-in-the-loop review

Overall Procedure and LLM-Assisted Inspection. The overall procedure begins with *monthly-level summary*, where annotators perform comprehensive reading, inspection, and revision. This is followed by *weekly-level summary*, which involves several checks: **(2.1) Consistency between parent and child summaries** verifies that weekly content does not contradict monthly content (e.g., ensuring events are not mistakenly placed before a meeting); **(2.2) Factual correctness** checks for obvious factual errors (e.g., accurately reflecting the initials on a ring); **(2.3) Repetition checking** uses an LLM to extract event descriptions, retrieves the five most similar events via similarity search, and inspects them to prevent unreasonable repetition (e.g., the protagonist reading the same book chapter and having identical reflections in different months); and **(2.4) Random sampling**, where 20 revised summaries are randomly selected for additional verification. The *day-level and event-level summaries* follow the same checking procedure as the weekly-level summaries. Given the substantial volume of content at the weekly, day, and event levels, we employ an LLM (specifically, Qwen3-235B-Instruct) to conduct a first-pass review for detecting potentially problematic segments, after which human annotators perform full manual inspection and correction.

Issue Detection and Final Quality. The proportions of issues detected by the LLM during the initial review are as follows: for parent-child consistency, weekly (11%), day (14%), and event (16%); for factual correctness, weekly (13%), day (17%), and event (19%); and for repetition checking, weekly (7%), day (10%), and event (12%). Following revisions based on these checks and subsequent human review, the final quality is confirmed through random sampling, which yields a pass rate of 100% for the weekly, day, and event-level summaries. All review work is conducted by several data annotators within our team.

C Annotator Details

All human annotators are professional data annotation employees of RayNeo AI, recruited internally from the company’s dedicated annotation team. Compensation was provided in accordance with the company’s standard salary structure, commensurate with their professional roles.

D Details of EgoMem Construction

In this section, we provide a comprehensive breakdown of the methodological decisions behind the construction of the EgoMem benchmark, specifically addressing the limitations of direct audio transcription, the rationale for our sliding-window generation strategy, and the human-in-the-loop review process designed to ensure conversational naturalness.

D.1 Limitations of Direct ASR Transcription

Our initial objective for EgoMem was to utilize Automatic Speech Recognition (ASR) to directly transcribe the raw audio captured by the Meta Aria smart glasses in the EgoLife dataset (Yang et al., 2025b). However, empirical investigations revealed several critical bottlenecks that rendered this approach impractical for constructing a robust long-term memory benchmark:

- **Data Sparsity:** Although EgoLife contains over 300 hours of recording (collected from six participants over seven days), the actual density of meaningful, continuous conversations is remarkably sparse. Direct transcription resulted in vast segments of silence or fragmented ambient noise, which is insufficient for evaluating complex memory retrieval and reasoning over long contexts.
- **ASR Degradation in the Wild:** Real-world egocentric audio is highly unstructured. It is plagued by severe background noise (e.g., wind, traffic, appliances), overlapping speech from multiple speakers,

and varying distances from the microphone. State-of-the-art ASR systems struggled to produce coherent transcripts under these conditions, introducing severe perceptual errors that would confound the evaluation of an agent’s memory reasoning capabilities (i.e., failing to answer a query due to ASR garbling rather than memory retrieval failure).

- **Incompatibility of Existing Q&A Pairs:** While the original EgoLife dataset provides its own human-annotated Q&A pairs, these queries are inherently *video-grounded* (e.g., asking about the visual state of an object or spatial relations). Because our benchmark specifically targets text-centric conversational memory architectures, these visual queries are not directly applicable.

Consequently, we pivoted to a hybrid proxy approach: using the high-fidelity, VLM-generated 10-minute visual summaries from Ego-R1 (Tian et al., 2025) as factual anchors, and prompting an LLM to synthesize the corresponding conversational interactions. This approach isolates memory reasoning bottlenecks from perceptual (ASR/Vision) errors while preserving the true physical event timeline.

D.2 Justification for Granularity and Context Window

A core challenge in generating summary-grounded simulated dialogues is selecting the appropriate temporal granularity. Ego-R1 provides summaries at multiple scales (30-second, 10-minute, 1-hour, 1-day, 1-week).

- **Granularity Selection:** We deliberately selected the 10-minute summaries. Finer-grained 30-second summaries are too fragmented, leading to disjointed micro-interactions that lack conversational flow. Conversely, coarser summaries (1-hour or 1-day) abstract away critical episodic details, resulting in superficial dialogues. We observed that asking an LLM to generate continuous dialogue directly from an hour-long summary often leads to severe information loss, a phenomenon consistent with the limited effective attention span of current LLMs when processing dense event descriptions (Mudarisov et al., 2025).
- **Sliding-Window Strategy:** To bridge discrete 10-minute segments into a continuous daily lifelog without losing temporal coherence, we designed a sliding-window generation strategy. When generating the dialogue for segment T , the LLM is conditioned on the textual summaries of segments $T - 6$ to T (representing a 60-minute historical context). This 60-minute window was empirically found to be the optimal sweet spot: it provides sufficient context to maintain speaker consistency, pronoun resolution, and ongoing topic flow, while preventing the LLM from entering hallucination loops or repeating distant historical interactions.

D.3 Iterative Review and Conversational Grounding

To prevent the generated dialogues from reading like overly formalized written prose—a common artifact of standard LLM generation—we implemented a rigorous human-LLM collaborative review pipeline inspired by *text-grad* (Yuksekgonul et al., 2024). The explicit goal of this pipeline was to optimize for **Conversational Naturalness**.

1. **LLM-Assisted Flagging:** We deployed a critique LLM instructed to evaluate the generated dialogues. Instead of optimizing for grammatical perfection, the critique model flagged segments that sounded “too formal,” “structurally rigid,” or “unnatural for close acquaintances.”
2. **Human Revision for Naturalness:** Human annotators reviewed the flagged segments and revised them to mimic real-world continuous recordings. Key adjustments included:
 - *Casual Phrasing and Tone:* Replacing highly structured, essay-like sentences with relaxed, colloquial expressions typical of daily roommate or family interactions.
 - *Implicit Contexts:* Ensuring that characters do not overly explain situations that close acquaintances would implicitly understand (e.g., referring to “the project” rather than “the software engineering project we discussed yesterday”). This forces the evaluated memory systems to rely on long-term context resolution rather than immediate explicit clues.

- *Dynamic Flow*: Smoothing out the transitions between the sliding windows to ensure topic shifts felt spontaneous rather than programmatic.

3. **Final Factual Alignment**: The revised, naturalized dialogues were checked one final time against the Ego-R1 summaries to guarantee that no physical events or critical factual details were altered during the conversational grounding process.

E Sensitivity Analysis

E.1 Impact of Backbone Capability

We investigate the influence of the underlying model’s capacity on memory performance by conducting experiments with Qwen3-4B, Qwen3-8B, and Qwen-Plus. As illustrated in Figure 5, the results demonstrate that the overall effectiveness of the memory system is positively correlated with the inherent capabilities of the base model. Notably, the transition from the 4B to the 8B variant yields only marginal performance gains and occasionally results in slight performance degradation. In contrast, upgrading to the qwen-plus model leads to a substantial improvement across evaluated metrics. These findings suggest that the deployment of high-capacity models is essential for maximizing the utility of agentic memory systems in complex tasks.

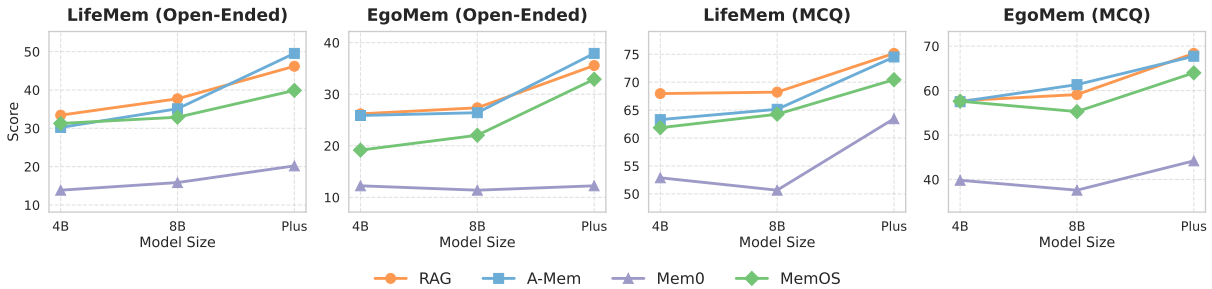


Figure 5: Comparison on the performances of memory systems using different backbone LLMs.

E.2 Sensitivity to Retrieval Top-K.

To investigate the influence of retrieval breadth on system efficacy, we conduct an ablation study on LifeMem leveraging Qwen3-8B under varying top- k configurations ($k \in \{10, 20, 40\}$). As illustrated in Table 4, performance generally scales positively with the retrieval depth k . Although A-Mem exhibits a minor performance regression in the Open-Ended task when k increases from 20 to 40, the results across both models and tasks consistently demonstrate that larger k values yield superior outcomes. This prevailing trend underscores that retrieval recall remains a primary bottleneck for memory-augmented agents. Furthermore, the findings suggest substantial potential for enhancing downstream performance by optimizing the retrieval modules of contemporary memory systems to more effectively manage broader context.

Table 4: Performance sensitivity analysis across varying top- k retrieval configurations on LifeMem.

| Method | Task | Top- k | | |
|--------|------------|----------|-------|-------|
| | | 10 | 20 | 40 |
| A-Mem | Open-Ended | 29.06 | 35.10 | 34.89 |
| | MCQ | 61.90 | 65.15 | 68.46 |
| MemOS | Open-Ended | 27.07 | 32.89 | 33.43 |
| | MCQ | 59.10 | 64.26 | 67.79 |

F Benchmark Data Samples

F.1 Examples of dataset

In this section, we provide illustrative snippets from the LifeMem Dataset. The EgoMem Dataset adopts the same formatting and structural schema.

Table 5: Jeremy and Jane at Home Organizing Old Items (2024-01-01)

| Time | Speaker | Utterance |
|------------|---------|---|
| [08:10:15] | Jane | All done eating. I'll go clear the bowls first, and then shall we tidy up the cabinet in the living room? |
| [08:10:22] | Jeremy | Okay, I'll help you clear up. No point letting them pile up. |
| [08:10:30] | Jane | Yeah, and give the tablecloth a good shake while you're at it, there are some breadcrumbs. |
| [08:10:38] | Jeremy | Alright, you go change into some clothes you don't mind getting dirty. I'll be over as soon as I finish here. |
| [08:11:05] | Jane | Hey, the bottom drawer of the cabinet is stuck. Can you give it a pull? |
| [08:11:10] | Jeremy | Let me see... Push it in a bit, then give it a sharp tug – There, it's open. |
| [08:11:18] | Jane | Wow, how did this box get so dusty? I think it's the old photo albums, right? |
| [08:11:24] | Jeremy | Should be. That was before we switched to a digital camera, all these were developed from film. |
| [08:11:30] | Jane | This one... was our first trip to Hangzhou? You were wearing that blue checkered shirt that year. |
| [08:11:36] | Jeremy | Haha, yes, taken at the entrance of Lingyin Temple. You insisted that monk was peeking at us while we took the picture. |
| [08:11:42] | Jane | He was looking! And you started laughing, the photo turned out all blurry. |
| [08:11:50] | Jeremy | Check the back, I think there are some from that Yunnan trip too? |
| [08:12:00] | Jane | Yes, here they are. At the gate of Dali Old Town, you had to wear your sunglasses crooked, trying to look all artsy. |
| [08:12:06] | Jeremy | That was called 'creating a vibe'. Look how happy you're laughing in this one. |
| [08:12:12] | Jane | Hmm... My hair wasn't gray back then. |
| [08:12:20] | Jeremy | It wasn't that long ago, was it? Seven or eight years? |
| [08:12:25] | Jane | Almost. Time really flies. Oh, how did this USB drive box get here too? |
| [08:12:32] | Jeremy | Used that for storing photos ages ago. I think it's labeled "2016 Family Photos". |
| [08:12:38] | Jane | Can we still read it? Should we find a computer and try? |
| [08:12:42] | Jeremy | I'll try it later on my study computer. The port should still be compatible. |
| [08:12:50] | Jane | No rush, let's sort these albums first. The old ones go on this side, the newer ones over here. |

Continued on next page...

Table 5 – continued from previous page

| Time | Speaker | Utterance |
|-------------|----------------|---|
| [08:13:00] | Jeremy | This yellow-edged one was from your mom, right? She said we should pick only the best ones to develop and keep. |
| [08:13:06] | Jane | Yes, she kept saying back then that when we got old, we could look through them together. |
| [08:13:12] | Jeremy | She was right. Isn't it nice looking through them now? |
| [08:13:20] | Jane | Mmm... This box also has a group photo from Weizhou's wedding. |
| [08:13:26] | Jeremy | Oh, look at him in the suit with a bow tie, like he stepped right out of a period drama set in the Republic of China era. |
| [08:13:32] | Jane | You're one to talk! Your tie was crooked, and he had to retie it for you. |
| [08:13:38] | Jeremy | Haha, you remember everything. We gotta keep this photo to tease him with next time we see him. |
| [08:13:45] | Jane | Don't overdo it. He's "Boss Zhang" now, you know. |
| [08:13:50] | Jeremy | To me, he'll always be that goofball who fell into the flowerbed playing basketball. |
| [08:14:00] | Jane | Oh, this one is of your dad fixing his bike in the yard... |
| [08:14:06] | Jeremy | Yeah, that old Phoenix brand bike. The chain kept falling off, he'd spend the whole afternoon fixing it. |
| [08:14:12] | Jane | He was so handy. All your repair skills, you learned from him. |
| [08:14:18] | Jeremy | Yeah... This is a really good photo. The light on his face, so peaceful. |
| [08:14:25] | Jane | Let's not throw these old things away. Let's find a box and store them properly. |
| [08:14:30] | Jeremy | Okay, I'll go get a storage bin from the storage room later. |

Table 6: Jeremy with Family Watching Spring Festival Gala (2024-02-10)

| Time | Speaker | Utterance |
|-------------|----------------|--|
| [16:00:12] | Jeremy | Mom, Jane, the Spring Festival Gala replay has started. The tea is freshly brewed, have it while it's hot. |
| [16:00:18] | Mother | Oh, this tea aroma is so comforting. Hangzhou Longjing really is something else. |
| [16:00:25] | Jane | Mmm, so light and refreshing. One sip and I feel completely relaxed. |
| [16:03:40] | Mother | These hosts look the same as always, wearing red dresses every year, smiling like flowers. |
| [16:05:10] | Jeremy | Mom, don't just look at what they're wearing. There's a cross-talk performance later, you love those. |
| [16:08:33] | Jane | I recognize this skit actor. He was hilarious last time playing that delivery guy. |
| [16:12:15] | Mother | Oh my, this kid acts so well, the way he talks is exactly like Auntie Wang next door back in my hometown. |

Continued on next page...

Table 6 – continued from previous page

| Time | Speaker | Utterance |
|-------------|----------------|--|
| [16:18:44] | Jeremy | Here, Mom, let me top up your tea. Careful, don't spill. |
| [16:19:01] | Jane | Did Dad used to love watching the Gala too? I remember you saying he always liked memorizing the punchlines. |
| [16:19:10] | Mother | Oh yes, your father-in-law would even take notes in a little notebook, saying he'd tell the students when school started. |
| [16:25:20] | Jeremy | This cross-talk is okay, but not as good as last year's. |
| [16:27:05] | Jane | Don't be so picky. Just being able to sit and watch it together as a family is nice enough. |
| [16:30:18] | Mother | Oh, speaking of cross-talk, it just reminded me—when Mingyuan was little, he went to pick bayberries on the hill behind the village. He fell out of the tree but insisted he didn't! |
| [16:30:30] | Jeremy | Mom, not this story again... |
| [16:30:33] | Jane | Huh? Tell me, tell me! I haven't heard this one! |
| [16:30:38] | Mother | That day, he insisted the sweetest bayberries were on the highest branch. Well, his hand slipped, and he landed right on his backside. Came back still stubbornly saying "I didn't cry," but his face was all swollen. Saying that with one side of his face puffed up, he looked like a little steamed bun. |
| [16:31:05] | Jane | Huh? Stung by a bee? Did you just say a bee? |
| [16:31:08] | Mother | Oh yes, right, it was a bee! I got mixed up—that was another time! Picking wild strawberries, there was a beehive in the grass, "buzz" and it stung him right on the face! |
| [16:31:18] | Jeremy | I really didn't cry, it's just... the tears came out on their own. |
| [16:31:22] | Jane | Hahaha, stop it! "Tears came out on their own"? What's that if not crying? |
| [16:31:27] | Mother | Exactly! He was so swollen even your dad couldn't recognize him, still insisting "I didn't cry." I put a cold towel on his face, and he's sniffing, saying "It's just a little itchy." |
| [16:31:40] | Jane | That's adorable! I have to write this down—(sound of typing on phone) Title it "Future Parenting Material". |
| [16:31:48] | Jeremy | Hey, don't write that down. What kind of positive example is that... |
| [16:31:52] | Mother | Why not? Stubborn kid, full of spirit! Kids these days don't have that kind of grit anymore. |
| [16:32:10] | Jane | When we... if we have kids in the future, I'll tell them this story. I'll add a subtitle: "On the Art of Graceful Stubbornness". |
| [16:32:18] | Jeremy | Don't you two gang up on me... |
| [16:32:25] | Mother | This isn't ganging up, it's family memories! Come on, Mingyuan, pour some more tea, let's keep watching. |
| [16:35:40] | Jane | This dance is so beautiful, the backdrop looks like an ink wash painting. |
| [16:36:15] | Mother | Yes, the costumes are lovely too, the colors are elegant, not too flashy. |

Continued on next page...

Table 6 – continued from previous page

| Time | Speaker | Utterance |
|-------------|----------------|---|
| [16:40:30] | Jeremy | The special effects here are used quite cleverly, they sync up well with the performers' movements. |
| [16:42:10] | Jane | See, isn't this what you called "cross-boundary integration"? |
| [16:42:15] | Jeremy | Heh, I guessed the start, but I didn't expect the effects to be this smooth. |
| [17:00:20] | Mother | This song is sung so beautifully, warms your heart listening to it. |
| [17:05:35] | Jane | This skit is starting to get interesting. This dad acts exactly like the department head at my clinic. |
| [17:10:12] | Jeremy | Shh—the accompaniment is coming up, I really like this melody. |
| [17:30:45] | Mother | Oh my, it's almost six o'clock. Shouldn't we start preparing dinner? |
| [17:31:00] | Jeremy | No rush. I've got some chicken soup with Chinese yam simmering, just need to heat it up, and there are dumplings too. |
| [17:31:10] | Jane | I'll set the table and pan-fry the leek dumplings we made yesterday. |
| [17:31:18] | Mother | Good, I'll help you with the tea. Time just flies when you're drinking this tea. |
| [19:02:10] | Jane | The song and dance numbers on the Gala are one after another, it's making me sleepy. |
| [19:02:25] | Mother | Yes, when I was young I could stay up until midnight, but now I feel like closing my eyes past nine. |
| [19:03:05] | Jeremy | How about we take a break? We can get up again closer to midnight? |
| [19:03:12] | Jane | Okay, I'll go charge my phone first, and I need to organize my notes. |
| [19:03:20] | Mother | I'll just stay put here. You two go ahead, I'll just listen to the Gala. |

Table 7: Jeremy in Emergency Project Post-Mortem Meeting (2024-06-03)

| Time | Speaker | Utterance |
|-------------|----------------|--|
| [10:15:00] | Jeremy | Is everyone here? Let's get started. As you all saw, last night's incident had a significant impact. We need to quickly piece together the timeline and identify the root causes. |
| [10:15:45] | Mike | Yes, we came straight from the morning stand-up. Wei and the Ops reps are here too. |
| [10:15:55] | Jeremy | Good. Let me briefly recap the timeline. Last night at 21:47, our monitoring platform started receiving a flood of 503 errors, concentrated on the user login and permission verification APIs. Frontend service response times spiked from an average of 80 milliseconds to over two seconds, lasting roughly twenty minutes. |
| [10:17:10] | Alex | On the backend side, we didn't receive alerts until 21:49, two minutes after the problem started. Furthermore, the initial alerts were scattered; no one realized it was a systemic issue initially. |

Continued on next page...

Table 7 – continued from previous page

| Time | Speaker | Utterance |
|-------------|----------------|--|
| [10:17:45] | Wei | The test environment monitoring didn't trigger because we hadn't simulated failure states for that authentication component. It appears a vulnerability in the third-party SDK was triggered by a scanning tool, causing it to crash outright, which then cascaded to our authorization service. |
| [10:18:35] | Other | Correct. Checking the logs confirms it's the CVE-2024-3187 mentioned in their urgent patch bulletin – a high-severity privilege escalation vulnerability. When their service restarted, our persistent connections were all severed, and we lacked reconnection safeguards. |
| [10:19:40] | Jeremy | So, fundamentally, it wasn't our code at fault. But the core issue is that our monitoring didn't flag the anomaly immediately. From 21:47 to 21:58 – a full 11 minutes – there was no clear, high-severity “service meltdown” alert. |
| [10:20:35] | Mike | That's unacceptable. Users couldn't access the app, and we were in the dark? |
| [10:20:50] | Jeremy | Exactly. Reviewing the Grafana dashboards, while we had heartbeat metrics, we lacked aggregated alerting for them. Also, the alert rules are too fragmented; a sea of red dots ended up masking the critical issue. |
| [10:21:45] | Wei | I checked the logs last night. The first call was to Alex at 21:55, reporting login timeouts. That's when we first suspected a common problem, but the command chain was unclear – no one took clear ownership of the emergency response. |
| [10:22:30] | Alex | I was initially checking logs, thought it might be a database issue, and even had the DBA team investigate. It took time to realize the upstream auth service was the root cause. |
| [10:23:15] | Ops Rep | We were also reactive. By the time we noticed the abnormal traffic drop and intervened, the golden window for mitigation had passed. |
| [10:24:00] | Jeremy | Therefore, while the trigger was a third-party component failure, this incident exposed our own weaknesses: insufficient monitoring sensitivity and a lack of a formalized emergency response process. |
| [10:24:50] | Mike | Agreed. The responsibility for the cause isn't ours, but our response was too slow. This has to change. |
| [10:25:15] | Jeremy | I propose we focus on two key areas moving forward. First, integrate health checks for external dependencies into our core monitoring. Heartbeat, version status, abnormal reconnection states – all need real-time, prominent alerting. |
| [10:26:20] | Wei | We can integrate that with our existing component health dashboard. Wasn't that already in progress? |
| [10:26:40] | Jeremy | Yes, this fits perfectly. Second, I've been thinking since last night: we need to prioritize implementing a robust canary release and automated rollback mechanism. If we could have automatically detected the spike in abnormal call rates and rolled back to the previous stable version, we could have halved the outage duration. |

Continued on next page...

Table 7 – continued from previous page

| Time | Speaker | Utterance |
|-------------|----------------|--|
| [10:27:55] | Alex | Automated rollback? Isn't that a bit aggressive? What about false positives? |
| [10:28:20] | Jeremy | Not a full, automatic rollback for all traffic. We can start with a canary release for a small percentage of users, say 1%, while closely monitoring key metrics – error rate, latency, authentication failure rate. If these exceed thresholds, automatically route traffic back to the old version and trigger alerts. |
| [10:29:30] | Ops Rep | We support this approach. We can configure the traffic switching using K8s; we've tested similar setups in our test environment before. |
| [10:30:10] | Wei | Then our release process needs updating too. The current manual tagging and manual image push is prone to missed steps. |
| [10:30:45] | Jeremy | Exactly. I want to implement a pre-release checklist, similar to the one we drafted earlier. Items like dependency scans, permission verification, rollback plan confirmation – all must be checked off before deployment. |
| [10:31:40] | Mike | I agree with this direction. Especially regarding external dependencies, we must confirm there are no known vulnerabilities and have a degradation plan before any future deployment. |
| [10:32:25] | Jeremy | I'll take the lead on drafting an improvement plan covering monitoring enhancements, the release process, and the emergency response mechanism. Target is to have a first draft by the end of this week. |
| [10:33:15] | Mike | Okay. You coordinate. Wei, you support with testing validation. Ops team, please provide a feasibility report for the automated traffic switching. |
| [10:33:55] | Ops Rep | Understood. We can schedule a technical alignment meeting this afternoon. |
| [10:34:25] | Jeremy | Good. Additionally, I suggest we conduct a failure drill next week, simulating a third-party service outage, to test if our current response procedures can handle it. |
| [10:35:15] | Wei | Agreed. I'll design the scenario, maybe add some complications like alerts being incorrectly marked as low priority. |
| [10:36:00] | Alex | I'll prepare an emergency procedure document then, clarifying roles and responsibilities – who does what under which circumstances – to prevent the lack of leadership we saw. |
| [10:36:50] | Jeremy | Alright, let's proceed on that basis. We'll schedule follow-up meetings for the details. Let's wrap up this post-mortem for now? |

F.2 Question Types and Examples

To provide a clearer understanding of the evaluation tasks, we first present the formal definitions of the four distinct question types designed in LIFEDIALBENCH, followed by concrete examples from the dataset.

- **QT1: Event Content Recall.** This type encompasses questions that demand the retrieval of core event content, and it falls under the broader category of Event Recall.
- **QT2: Event Detail Retrieval.** Questions of this type require precise retrieval of specific event details, and they are classified under Detail Retrieval.
- **QT3: Multi-hop Event Reasoning.** These questions involve both retrieving and reasoning across multiple events, and they belong to the Temporal Reasoning category.
- **QT4: Temporal Grounding.** As a lifelog-specific subcategory of Detail Retrieval, this unique type requires accurately pinpointing the exact timestamp of a particular event to generate a valid answer.

The following examples demonstrate these question types, featuring the query, associated timestamp, and candidate options.

Question Examples

Single Event: What was the main topic of discussion between Jeremy and Jane during the organization of old items? [query_timestamp=2024-01-03]

- (A) Memories of their 2018 trip to Dali and Lijiang in Yunnan;
- (B) Preliminary planning for the Spring Festival holiday;
- (C) Optimization solutions for household clutter management;
- (D) Discussion on edge computing communication protocols.

Event Detail: What specific item did Jane mention when recalling the Yunnan trip? [query_timestamp=2024-01-01]

- (A) A tie-dyed scarf;
- (B) A bicycle;
- (C) A hat;
- (D) A pair of shoes.

Multi Event: During which activities did Jeremy and Jane discuss topics related to children's health? [query_timestamp=2024-01-01]

- (A) During breakfast and balcony reading;
- (B) While organizing old items and watching a movie;
- (C) During grocery shopping and dinner preparation;
- (D) During lunch and while debugging the projector.

Temporal Info: What was the specific time when Jeremy and Jane began immersing themselves in the photos from their Yunnan trip? [query_timestamp=2024-01-03]

- (A) 9:00 AM;
- (B) 10:30 AM;
- (C) 11:00 AM;
- (D) 1:00 PM.

G Prompts

G.1 Judge Prompt used for Open-Ended Format

To evaluate the semantic accuracy of the Open-Ended generation, we employ an LLM-based judge. This judge compares the model's response against the ground-truth answer (derived from the correct option) to determine semantic equivalence. The specific prompt used is detailed below:

Judge Prompt for Open-Ended QA

You are given a question, its ground-truth answer, and a model response. Determine if the model response is semantically equivalent or meaningfully similar to the ground-truth answer. Consider the following as acceptable variations:

- Different wording but same core meaning
- Partial answers that contain the key information
- Answers with additional relevant context
- Answers that rephrase the same idea
- Minor factual details may differ if the main point is correct

Be lenient in your judgment - if the response captures the essence of the correct answer, consider it correct.

Question: {question}

Ground-truth answer: {reference}

Model response: {candidate}

G.2 Lifelog Generation Prompt

Here is the prompt we use for transforming 10-minutes summarization to lifelog.

Prompt Template

You are required to transform the target first-person narrative into lifelog-style conversation records. **Lifelog** refers to authentic daily spoken conversations captured by portable recording devices. Your task is not storytelling but converting the given narrative into natural dialogues that sound like real speech.

Character name
{character_name}

Previous Narratives (context for coherence):
{previous_narratives}

Target First-person Narrative:
{first_person_narrative}

Time range in target narrative:
{time_range}

Conversation Generation Requirements

Core Conversion Principles:

1. **Narrative-to-Lifelog Transformation:** Convert the target first-person narrative into lifelog dialogues, ensuring all important details from the narrative are preserved in the conversations.
2. **Continuity and Non-redundancy:** Previous narratives are provided to maintain timeline consistency, character relationships, and avoid repeating the same details unnecessarily.
3. **Authenticity:** The dialogues must sound natural, spontaneous, and spoken in real daily English, avoiding formal or literary expressions.

Format Specifications:

- Strictly use the format: [yyyy-mm-dd, HH:MM:SS] Character: Speech content

Content Requirements:

1. **Detail Preservation:** Every concrete detail in the target narrative (actions, observations, emotions, objects, times, etc.) must appear in the dialogues.
2. **Logical Flow:** Keep the event flow consistent with both the target narrative and previous lifelogs.
 - Ensure continuity of relationships between characters.
 - Keep the timeline reasonable and coherent.
3. **Boundary Control:** Do not introduce cross-day planning, greetings, farewells, or artificial summaries. End conversations naturally when the described event ends.

****Output Format:****

- Only output lifelog dialogues in English, without explanations, notes, or extra text.

Example Format

[2025-09-17, 09:23:11] Speaker A: Actual spoken words

[2025-09-17, 09:23:15] Speaker B: Dialogue continues

Now please generate lifelog conversations according to the above requirements.

The following prompts were employed in the Top-Down Hierarchical Life Simulation Framework. Year-level summaries are progressively allocated and enriched at the month level to generate detailed monthly summaries, while the prompts for the "month-to-week" and "week-to-day" stages have been slightly adjusted.

Prompt Template to Allocate

You are a professional lifelog analyst. Based on the provided annual experience summary, restructure and expand the content by month to generate detailed, coherent, and realistic monthly life records.

{holidays}

{important_days}

Annual Experience Summary:

{year_summary}

Requirements

- Each monthly record must clearly describe the time, location, people involved, process, and outcomes of events.
- While strictly reconstructing the annual experiences by month, you may expand each month's record.
- Your expansions must be realistic; ensure the content is substantial and natural, and avoid fabricated dramatic plots or supernatural elements.
- After reconstruction and expansion, each month's record must cover major events, work, exercise, entertainment, family communication, and social activities.
- If a specific time point for an event is clearly stated in the annual summary, you must not change it; if it is not specified, assign a reasonable time.

Output Format

Output strictly as a standard JSON array, and output only the JSON array without any explanations or comments. Each item in the JSON array should have the following structure:

```
[ {{ "Month": "{year} January", "Monthly Record": "..."}}, {{ "Month": "{year} February", "Monthly Record": "..."}}, ... ]
```

Prompt Template to Enrich

You are a professional lifelog analyst. Below are this person's monthly records for the target month and the adjacent months. Please enrich the current record for the target month to make the description more comprehensive.

{prev_months}

Existing monthly record for {month}:

{month_data}

Date information for {month}:

{month_dates_info}

Requirements:

- Each monthly record must clearly describe the time, location, people involved, process, and outcomes of events.
- Unless the current month's record already contains such mentions, do not add any cross-month plans during enrichment; for example, do not schedule April activities in the March record.

- The enriched content must be realistic; ensure the content is substantial and natural. Avoid fabricated dramatic plots or supernatural elements.
- The enriched record should cover all facets of life, including but not limited to major events, work, exercise, entertainment, family communication, and social activities.
- The enriched content must cover the entire month—early, mid, and late—and distribute events as evenly as possible. If the original record provides specific dates/times, you must keep them.
- The enriched content must remain temporally consistent with the records of the previous and following months, ensuring coherence without contradictions.
- Note that workdays are typically Monday through Friday, rest days are Saturday and Sunday, and public holidays are rest days. Arrange work and life content accordingly.

Output Format

Output strictly as standard JSON, and output only the JSON without any explanations or comments. The JSON fields are:
 { "Month": "{month}", "Monthly Record": "..."}

G.3 Question Generation Prompt

Prompt Template to Generate Daily-level Questions

Prompt for Event Extractor Evaluation Data Generation

You need to generate evaluation data for an event extractor. The event extractor will extract useful information from users' life records and store it in a database.

Now you will be provided with a user's daily experiences, and you need to generate four questions based on the content, with four options (A, B, C, D) for each question (one correct answer and three distractor options). These questions and options will be used to evaluate the extraction performance of the event extractor.

Daily Events (date)

{all_day_events}

Question Requirements

- Generate 4 question-answer pairs, which should ask about the following four aspects respectively:
 - The content of a specific event
 - A specific detail of a specific event
 - The content of multiple events
 - The specific time when a specific event occurred
- Question Guidelines:
 - Frame questions about events that involve interactions with others and can generate dialogue data; do not frame questions about events that cannot generate dialogue data.
 - The events targeted by the questions must be unique enough and must not be daily routine events.

Output Requirements

- You need to output a JSON list, where each JSON element contains the following fields:
 - 'question': The content of the question
 - 'options': A list containing four options, formatted as ["A. Option content", "B. Option content", "C. Option content", "D. Option content"]
 - 'answer': The option letter of the correct answer, e.g., 'A'
- Do not output any content other than the JSON list of question-answer pairs

H Additional Results

H.1 Offline Evaluation Results

We present the performance of various memory systems in the **offline evaluation setting** (Table 8). In this configuration, agents process the complete dialogue lifelog before answering queries. These results provide a benchmark for the models' fundamental memory capacity, offering a performance upper bound by decoupling the memory task from the requirements of real-time, causal streaming.

H.2 Question Types Distribution

The distribution of four question types is in Figure 6. We ensure that both subsets would have a balance question-type proportion.

Table 8: Results on Offline Evaluation Setting

| Models | Method | Open-Ended | | | | | Multiple-Choices | | | | |
|----------------|--------|------------|-------|-------|-------|---------|------------------|-------|-------|-------|---------|
| | | QT1 | QT2 | QT3 | QT4 | Overall | QT1 | QT2 | QT3 | QT4 | Overall |
| EgoMem | | | | | | | | | | | |
| gpt-4o-mini | RAG | 37.09 | 51.25 | 9.60 | 23.87 | 30.88 | 68.14 | 83.75 | 56.33 | 49.09 | 64.74 |
| | A-Mem | 33.06 | 51.25 | 9.17 | 27.47 | 30.56 | 68.54 | 81.25 | 55.02 | 54.05 | 64.74 |
| | Mem0 | 12.09 | 16.25 | 2.62 | 14.41 | 11.39 | 54.03 | 51.35 | 46.28 | 32.88 | 46.11 |
| | MemOS | 26.20 | 32.50 | 9.60 | 9.90 | 19.91 | 66.93 | 72.08 | 68.55 | 32.88 | 60.59 |
| qwen-plus | RAG | 38.70 | 56.25 | 11.79 | 32.43 | 35.14 | 66.12 | 82.91 | 55.89 | 51.80 | 64.53 |
| | A-Mem | 38.30 | 52.91 | 13.10 | 29.72 | 33.86 | 67.33 | 80.41 | 51.09 | 55.40 | 63.89 |
| | Mem0 | 12.90 | 17.08 | 3.93 | 10.36 | 11.18 | 40.72 | 51.25 | 34.93 | 33.33 | 40.25 |
| | MemOS | 29.83 | 42.08 | 13.53 | 31.53 | 29.39 | 62.50 | 79.58 | 61.13 | 52.25 | 64.11 |
| LifeMem | | | | | | | | | | | |
| gpt-4o-mini | RAG | 32.41 | 71.80 | 19.99 | 20.18 | 36.12 | 73.79 | 91.70 | 69.07 | 52.67 | 72.13 |
| | A-Mem | 28.27 | 65.17 | 17.20 | 19.71 | 32.61 | 71.95 | 88.85 | 68.36 | 46.16 | 69.14 |
| | Mem0 | 5.97 | 21.56 | 4.19 | 5.80 | 9.36 | 66.89 | 62.56 | 67.67 | 31.09 | 57.37 |
| | MemOS | 23.21 | 55.17 | 11.26 | 13.79 | 25.85 | 73.56 | 78.61 | 66.89 | 30.80 | 62.47 |
| qwen-plus | RAG | 35.17 | 68.24 | 25.81 | 31.32 | 40.22 | 74.02 | 90.52 | 69.76 | 57.20 | 72.80 |
| | A-Mem | 34.70 | 68.71 | 29.99 | 29.70 | 40.86 | 72.64 | 87.44 | 62.09 | 54.88 | 69.19 |
| | Mem0 | 11.49 | 30.33 | 8.83 | 9.97 | 15.16 | 68.50 | 68.72 | 58.37 | 40.23 | 58.94 |
| | MemOS | 36.36 | 69.31 | 25.45 | 20.45 | 37.89 | 73.18 | 83.86 | 65.68 | 56.59 | 69.82 |

I Memory Systems

I.1 Description of Memory Systems

We define the agent’s characters in a memory system as Figure 7. A memory system often constructed by a summary agent, a memory manager, a retrieve agent, and a chat agent. The actual role would be different corresponding to the system design, as some role could be merged to one (e.g., memory manager and retrieve agent). The chat agent could be a part of the memory system, while there are also some systems exclude it.

I.2 Implementation Details

General Settings To ensure a fair comparison across all baselines, we maintain a unified configuration for retrieval and processing granularity. Specifically, we set the retrieval depth **top- k** to **20** by default. Furthermore, we adopt the **session (event)** as the fundamental processing unit for memory ingestion and storage.

RAG The simple RAG baseline includes a chat-agent and an embedding model to save and retrieve the relevant text. Therefore, there is no summary agent, no LLM memory manager, and no LLM retriever inside the system. It directly embeds and retrieves the lifelog text chunks into a vector database.

A-Mem, Mem0 and MemOS We follow the official code of these memory systems’ GitHub repositories for evaluation. The prompts inside these systems are specifically refined to fit the requirement for our benchmark evaluation.

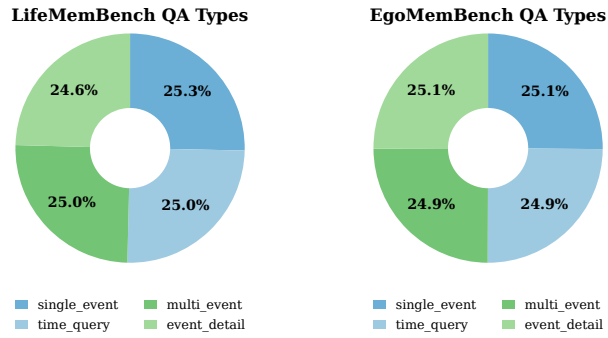


Figure 6: The Distribution of QA types.

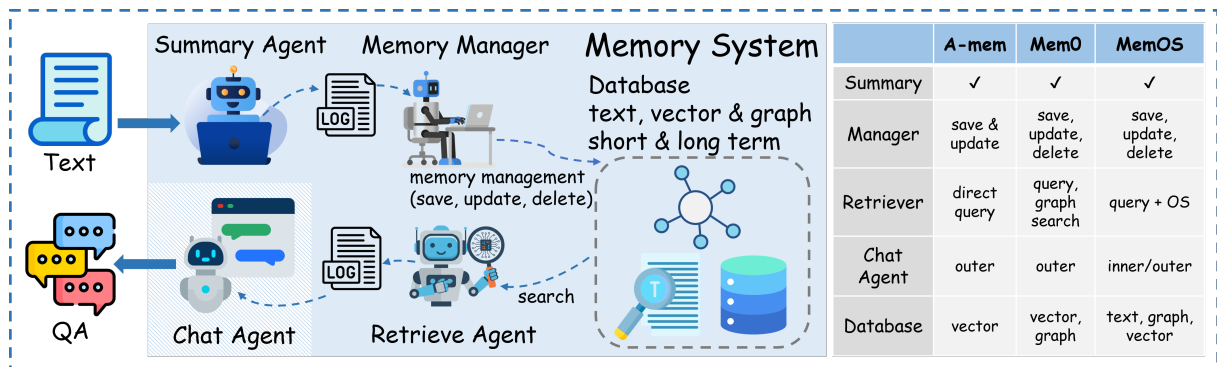


Figure 7: Definition of the structure of a memory system, and a comparison table of current memory system approaches under this structure.