HaluMem: Evaluating Hallucinations in Memory Systems of Agents

Anonymous ACL submission

Abstract

Memory systems are essential for enabling long-term learning and sustained interaction in AI systems such as LLMs and agents. However, memory storage and retrieval often suffer from hallucinations, including fabrication, errors, conflicts, and omissions. Existing evaluations mainly rely on end-to-end question answering, making it difficult to identify which memory operation causes hallucinations. To address this, we propose the Hallucination in Memory Benchmark (HaluMem), the first operation-level benchmark for evaluating hallucinations in memory systems. HaluMem defines three tasks (memory extraction, memory updating, and memory question answering) to comprehensively reveal hallucination behaviors across different operational stages of interaction. To support evaluation, we construct user-centric, multi-turn human-AI interaction datasets, HaluMem-Medium and HaluMem-Long, containing about 15k memory points and 3.5k questions. The datasets feature long dialogues, with average lengths of 1.5k and 2.6k turns and context sizes exceeding 1M tokens, enabling evaluation across varying context scales and task complexities. Experiments on HaluMem show that current memory systems tend to introduce and accumulate hallucinations during extraction and updating, which then propagate to question answering. These findings highlight the need for interpretable and constrained memory operation mechanisms to improve memory reliability. All resources are available at GitHub.

1 Introduction

Human–AI interactions contain personalized information that is often lost when conversations end (Shi et al., 2024; Zhao et al., 2025). To enable continuous understanding and adaptation, LLMs require persistent memory to record and utilize user details (Liu et al., 2023; Zhang, 2024). Memory systems address this by extracting and managing information from dialogue history to support long-term consistency (Li et al., 2025; Ye et al., 2025; Kang et al., 2025a; Rasmussen et al., 2025; Shah et al., 2025). Representative frameworks, such as MemOS (Li et al., 2025), Mem0 (Chhikara et al., 2025), Zep (Rasmussen et al., 2025), Supermemory (Shah et al., 2025), and Memobase (Ye et al., 2025), record user profiles, events, and evolving preferences, supporting the creation, revision, and tracking of memories to construct a system-level memory layer with structured management capabilities.

Figure 1: Examples of operation-level hallucination in a memory system.

Despite their utility, these systems are prone to memory hallucination, manifesting as fabricated, erroneous, conflicting, or missing information during extraction or updating (Oche et al., 2025; Agrawal et al., 2024; Zhang and Zhang, 2025) (Figure 1). Such upstream errors undermine reliability and can amplify downstream generation hallucinations(Huang et al., 2025a). Crucially, existing evaluations primarily rely on end-to-end QA metrics, which infer memory quality from AI output performance but fail to identify the specific operational stage where hallucinations originate.

To address this, we propose the Hallucination in Memory Benchmark (HaluMem), the first operation-level hallucination evaluation benchmark for memory systems. HaluMem assesses hallucinations in three distinct tasks: memory extraction, memory updating, and memory question answering, to comprehensively reveal hallucination behaviors at the operation level. We provide a user-centered, multi-turn conversation dataset reflecting the core function of memory systems in supporting personalized, long-term human–AI interactions by organizing memories around the user, complete with precise annotations for memory points. By comparing system outputs for each operation against these ground truths, we enable a fine-grained evaluation of accuracy in extraction, consistency in updating, and reliability in retrieval.

The benchmark comprises two datasets, HaluMem-Medium and HaluMem-Long, containing approximately 15,000 memory points and over 3,400 evaluation queries. The latter extends user's context lengths to the million-token scale, allowing examination of hallucination behaviors in ultra-long conversations.

The main contributions are as follows:

- We propose HaluMem, the first operation-level benchmark for memory hallucinations, which overcomes the limitations of end-to-end methods by evaluating extraction, updating, and QA separately.
- We construct HaluMem-Medium and HaluMem-Long, extensive multi-turn datasets designed to assess system performance under varying context scales and task complexities.
- Stage-wise evaluation uncovers the cumulative and amplifying effects of hallucinations across memory operations, offering new insights for their mitigation.

2 Related Work

2.1 Memory System

LLMs primarily rely on implicit memory embedded in parameters. While offering strong reasoning, this parameterized memory suffers from poor manageability, as it is immutable and opaque, often leading to hallucinations when handling outdated or conflicting information. To address this, external memory modules were introduced. RAG(Lewis et al., 2020) provides a transparent, editable plaintext memory with high manageability, though it lacks the ability to model complex entity relationships. GraphRAG(Edge et al., 2025) mitigates this by incorporating knowledge graphs to enhance structural reasoning, yet this introduces higher maintenance costs and complexity compared to traditional RAG.

Recent research focuses on memory systems that support personalized interaction and lifecycle management. Graph-based approaches like Supermemory(Shah et al., 2025) and Zep(Rasmussen et al., 2025) utilize contextual graphs and components like Graphiti to capture temporal and relational information for long-term consistency. In contrast, plaintext-focused systems like Memobase(Ye et al., 2025) and Mem0(Chhikara et al., 2025) emphasize manageability, employing dynamic user profiling and metadata-enriched storage with conflict detection, respectively. Moving towards system-level abstraction, MemOS (Li et al., 2025) unifies the management of parametric, activation, and explicit memory resources. Overall, while graph-integrated systems enhance expressiveness, they face greater management challenges compared to text-based solutions.

2.2 Evaluation Hallucinations in Memory Systems

Hallucinations in memory systems can be divided into two types: *memory hallucinations* (errors in storage, updating, or retrieval) and *generation hallucinations* (output inconsistencies). While generation hallucinations are well-studied, with mature frameworks including factual verification (Lee and Yu, 2025; Huang et al., 2025b), internal state analysis (Su et al., 2024; Chen et al., 2024), and uncertainty detection (Kang et al., 2025b; Shelmanov et al., 2025; Liang et al., 2024), research on memory-specific hallucinations remains at an early stage.

Existing benchmarks (summarized in Table 1) focus primarily on holistic system performance rather than specific memory faults. LoCoMo(Maharana et al., 2024) evaluates static information retention in long contexts but lacks

Feature	HaluMem	PersonaMem	LOCOMO	LongMemEval	PrefEval
Evaluation Granularity	Operation-level	End-to-end	End-to-end	End-to-end	End-to-end
Evaluation Timing	After each session	After all sessions	After all sessions	After all sessions	After all sessions
Evaluation Tasks	Memory Extraction, Memory Updating, Memory QA	Multiple Choice	QA, Summarization, Generation	QA, Memory Recall	Generation, Classification
Memory Type	Persona, Event, Relationship	Persona	Persona, Event	Persona, Event	Persona
Memory Update	Yes	Yes	No	Yes	Yes
Conversation Time Span	10∼20 years	Several years*	Several months	∼ 2.5 years	-
Avg Length / Session	8.3k tokens	6k tokens	477 tokens	3k tokens	-
Max Context Length	1M tokens	1M tokens	9k tokens	1.5M tokens	100k tokens
Question Num	3,467	∼ 6,000	7,512	500	3,000

* "Several years" for PersonaMem is inferred from the paper and dataset, not explicitly labeled.

Table 1: HaluMem vs. Existing End-to-End Benchmarks for Memory System Evaluation

memory update mechanisms. Building on this, LongMemEval(Wu et al., 2025) incorporates explicit updates to assess dynamic knowledge consistency over time. In the domain of personalized interaction, PrefEval(Zhao et al., 2025) and PersonaMem(Jiang et al., 2025) target user preference and persona consistency, with the latter covering multi-year timespans to test traceability. However, these approaches utilize end-to-end evaluation paradigms (e.g., QA or generation), treating memory systems as black boxes where hallucinations are only indirectly observable. In contrast, HaluMem is designed for fine-grained, operation-level analysis, directly addressing the gap in evaluating hallucinations across memory lifecycle stages.

3 Problem Definition

Let a memory system S provide an AI system A (an LLM or agent) with long-term memory and personalization. It receives a multi-turn dialogue $D = (u_1, a_1), (u_2, a_2), \ldots, (u_N, a_N)$, where u_i and a_i are the user's and AI's utterances at turn i. Each memory point is stored as plaintext, with a single memory denoted as m. For dialogue D, the memory system performs four types of operations during interaction: (1) *Memory Extraction* (E): extracting newly generated memory points from D; (2) *Memory Updating* (U): modifying or deleting existing memories; (3) *Memory Retrieval* (R): recalling memories relevant to the current query[1]; (4) *Memory Question Answering* (Q): constructing prompts and invoking A to generate responses.

Existing evaluations of memory systems typically adopt an end-to-end question–answer paradigm (Figure 2). Given a set of dialogue-based queries $\mathcal{Q} = \{q_j\}_{j=1}^{J}$ and their corresponding gold answers $\mathcal{Y}^* = \{y_j^*\}_{j=1}^{J}$, the evaluation pipeline can be abstracted as

$$\hat{M} = U\left(E(D)\right), \hat{R}_j = R(\hat{M}, q_j), \hat{y}_j = A\left(\hat{R}_j, q_j\right). \quad (1)$$

End-to-end evaluation is based on answer-level metrics, such as accuracy and F1 score:

$$\mathrm{Acc}_{\mathrm{e2e}} = \frac{1}{J}\sum_{j=1}^{J} \mathbb{I}\left[\hat{y}_j = y_j^*\right]. \quad (2)$$

When $\hat{y}j \neq y_j^*$, $\mathrm{Acc}_{\mathrm{e2e}}$ cannot pinpoint the error source. It is unclear whether hallucination arises from the extraction stage E (incorrect or fabricated memories), the updating stage U (old memories wrongly modified or not refreshed), or the question answering stage Q (unsupported generative content despite correct memories). This lack of traceability hinders the targeted mitigation.

To enable a localized and diagnostic evaluation, we construct fine-grained annotations and define gold standards for each stage. (1) *Extraction gold standard:* $G^{\mathrm{ext}} = \{m_i\}_{i=1}^{K}$, representing the set of memory points that should be newly added during the dialogue. (2) *Updating gold standard:* $G^{\mathrm{upd}} = \{m^{\mathrm{old}} \to m^{\mathrm{new}}\}$, representing the set of memory point pairs before and after updates during the dialogue. (3) *Question–answer dataset:* for each query q_j, a gold answer y_j^* is provided. The system outputs are defined as follows:

$$\begin{gathered}\hat{M}^{\mathrm{ext}} = E(D), \hat{G}^{\mathrm{upd}} = U(\hat{M}^{\mathrm{ext}}, D), \\ \hat{y}_j = A\left(R(\hat{M}, q_j), q_j\right)\end{gathered} \quad (3)$$

where $\hat{M}$ denotes the set of memory points representing the current state of the memory system when query q_j is processed. By providing stage-specific gold standards and evaluation metrics for E, U, and Q, the proposed **HaluMem** benchmark enables operation-level hallucination evaluation.

[1]As retrieval R mainly concerns on relevance and recall rate and rarely introduces generative processing by LLMs, this study concentrates on the three stages that directly induce hallucinations: E, U, and Q.

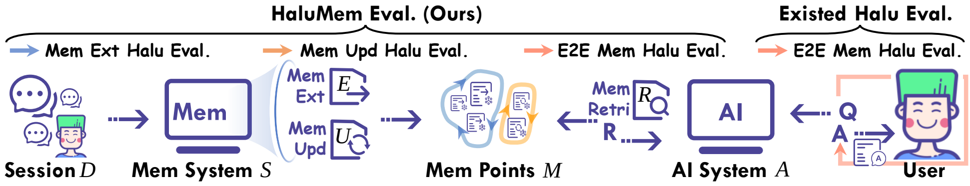

Figure 2: Comparison between HaluMem and existing benchmarks for memory systems.

4 Methodology for Constructing HaluMem

To systematically evaluate memory systems in realistic scenarios, we construct HaluMem. To ensure the quality of the dataset and the controllability of the construction process, we design a user-centered, six-stage procedure based on a progressive expansion strategy (Figure 3).

Stage 1: Persona Construction. This stage initializes virtual users with comprehensive profiles to simulate realistic interaction partners. Each profile comprises three dimensions: a static *Core Profile*, an evolving *Dynamic State* (career, health, and relationships), and randomized *Preferences* (6–8 specific attributes). This structure balances stable traits with dynamic contexts to support diverse memory extraction, anchored by a unified timestamp for temporal consistency. To ensure authenticity, user seeds are sampled from Persona Hub[2] (Ge et al., 2025), structured via rule-based procedures, and refined by GPT-4o. An example profile is provided in Appendix E.1.

Stage 2: Life Skeleton. After generating persona profiles, the second stage builds a *life skeleton* that defines each user's evolutionary trajectory. Each user receives several core career events centered on life goals, which serve as anchors for the evolution of dynamic information. Updates to social status, career transitions, or health conditions are typically associated with these career events. Preference information evolves separately through probabilistic modifications or deletions, independent of these career events. These probabilistic rules ensure a diverse yet coherent evolution. The life skeleton captures the user's potential future states and serves as a structured script for later memory addition, modification, and deletion, maintaining the complexity and consistency of the evaluation scenarios.

Stage 3: Event Flow. As the core component of dataset construction, Stage 3 converts the abstract life skeleton from Stage 2 into a structured, narrative *event flow*. It "eventifies" evolution instructions into a complete memory timeline that integrates initial states, career development, and preference changes, balancing narrative coherence with interpretability. This stage defines three types of events:

- **Init Events:** Derived from the user's initial profile, covering core, dynamic, and preference information. They serve as the starting point of the memory timeline, simulating the user's first self-introduction.
- **Career Events:** Based on the life skeleton and forming the main storyline, with each event divided into sub-stages and instantiated with dynamic information.
- **Daily Events:** Generated from preference evolution independent of career progression, where each change is recorded as a concrete life scenario with its cause and pre/post-change states.

Career events provide the narrative backbone, while init and daily events supply background and context. Chronological integration of these events yields a coherent event sequence that functions as the user's *memory transaction log*. See Appendix E.2 for examples.

Stage 4: Session Summaries and Memory Points. This stage transforms the structured event flow from Stage 3 into realistic session summaries and detailed memory points. For each event, we create a human–AI dialogue scenario shaped by the user's motivation. The system has access to the current persona profile, along with all prior events and memory points, ensuring logical, causal, and consistent generation. As events unfold, the persona profile is dynamically updated to reflect the user's evolving state. Each memory point includes its content, type(persona, event, or relationship), and importance, with updated entries preserving replaced information for traceability. More details provided in Appendices A.1 and E.3.

Stage 5: Session Generation. This stage transforms the event flow and memory points

[2] A collection of one billion diverse personas.

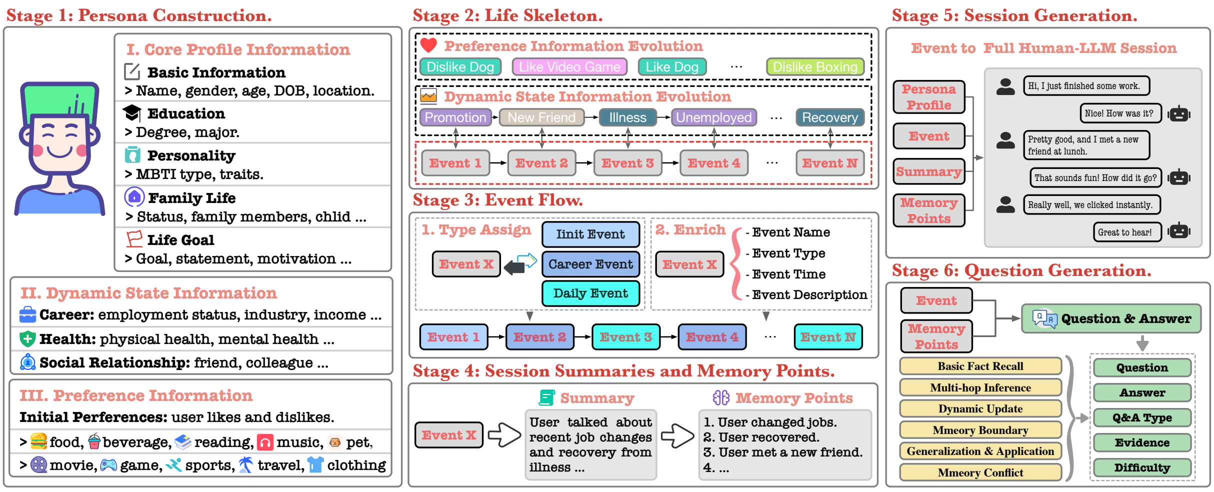

Figure 3: Framework of the HaluMem Construction Pipeline.

into context-rich, adversarial multi-turn dialogues. The process has three steps: adversarial content injection, dialogue generation, and memory verification. Adversarial content injection adds distractor memories[3], while memory verification ensures consistency between memory points and the generated dialogues. Overall, this stage simulates how memory is formed, maintained, and challenged in realistic conversations, producing data that test long-term memory performance and hallucination resistance. Examples appear in Appendix E.3.

Stage 6: Question Generation. The final stage constructs memory-oriented QA pairs based on the sessions and memory points. Six categories of questions are predefined, and the number and types of questions are programmatically allocated according to event type and complexity to ensure balanced coverage. To augment reasoning depth and complexity, substages of career events are consolidated into unified contexts. Each QA pair is annotated with a difficulty level and traceable evidence that is explicitly linked to source memory points (refer to Appendices A.2 and E.3).

To validate HaluMem, we conducted a human evaluation on a 700-session subset (>50%) of *HaluMem-Medium*, encompassing both memory points and QA pairs. Eight qualified annotators assessed *Correctness*, *Relevance*, and *Consistency*. The evaluation yielded a 95.70% correctness rate, with average relevance and consistency scores of 9.58 and 9.45, respectively. These metrics confirm the benchmark's high reliability. Details are provided in Appendix C.

Overall, We present two datasets: *HaluMem-Medium* and *HaluMem-Long*. The former contains 30,073 turns from 20 users (avg. 160k tokens/user), annotated with 14,948 memory points and 3,467 QA pairs. The latter extends the context to 1M tokens via distractor dialogues, increasing the volume to 53,516 turns. Refer to Appendices A.3 and A.4 for details.

5 Evaluation Framework of HaluMem

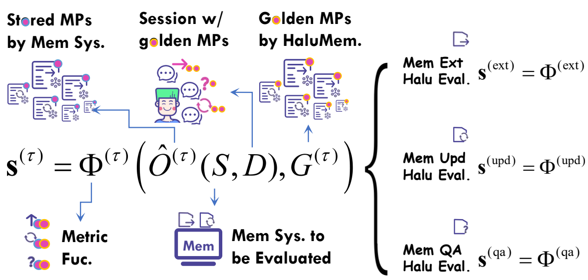

Figure 4: Hallucination evaluation process.

For each user, the session-level evaluation procedure of HaluMem (Figure 4) is defined as follows: (1) Dialogue sessions $D^1, D^2, \ldots, D^S$ are sequentially fed, in chronological order, into the memory system S. (2) If the current session D^s contains reference memory points or QA tasks, the corresponding evaluation process (extraction, updating, or QA) is triggered immediately after S completes processing that session, and the results are recorded. (3) After processing all sessions, the metrics of the three categories of tasks are aggregated to obtain the overall system performance.

To support this evaluation workflow, the system must provide three APIs: (1) Add Dia-

[3] False but similar memories that the AI naturally uses while the user stays silent, mimicking realistic information contamination

logue API: input dialogues and perform memory extraction; (2) Get Dialogue Memory API: retrieve memory points extracted from a specified session; (3) Retrieve Memory API: retrieve the most relevant memories for a given query.

Based on the above workflow and interface design, HaluMem conducts operation-level evaluation of memory systems across three essential tasks: *Memory Extraction*, *Memory Updating*, and *Memory Question Answering*.

5.1 Memory Extraction

The memory extraction task evaluates whether the system can correctly identify and store key information from dialogues while avoiding fabricated or irrelevant memories. For each dialogue session D^s that contains reference memories, the benchmark provides a gold memory set $G_s^{\text{ext}} = \{m_i^s\}_{i=1}^{K_s}$ that should be extracted. The system output after processing D^s is the extracted memory set $\widehat{M}_s^{\text{ext}} = \{\hat{m}_j^s\}_{j=1}^{\widehat{K}_s}$, which is used for evaluation.

Memory Integrity (Anti-Amnesia) This metric measures whether the system omits crucial information that should be extracted:

$$\begin{aligned}\text{Memory Recall} &= \frac{N_{\text{correct}}}{N_{\text{should}}},\\ \text{Weighted Memory Recall} &= \frac{\sum_{i=1}^{N_{\text{should}}} w_i \cdot s_i}{\sum_{i=1}^{N_{\text{should}}} w_i},\end{aligned} \tag{4}$$

where $N_{\text{should}} = |G_s^{\text{ext}}|$, N_{correct} denotes the number of correctly extracted memories, w_i represents the importance weight of the i-th memory, and $s_i \in \{1, 0.5, 0\}$ indicates the extraction score (completed extracted, partially extracted, or omitted).

Memory Accuracy (Anti-Hallucination) This metric evaluates whether the extracted memories are factual and free of hallucination:

$$\begin{aligned}\text{Memory Accuracy} &= \frac{\sum_{j=1}^{N_{\text{extract}}} s_j}{N_{\text{extract}}},\\ \text{Target Memory Precision} &= \frac{\sum_{j \in M_T} s_j}{|M_T|},\end{aligned} \tag{5}$$

where $N_{\text{extract}} = |\widehat{M}_s^{\text{ext}}|$, and $M_T \subset \widehat{M}_s^{\text{ext}}$ denotes the set of target memories that match the reference ones.

False Memory Resistance (FMR) This metric measures a system's resistance to hallucination when facing distracting content that the AI mentions but the user does not confirm:

$$\text{FMR} = \frac{N_{\text{miss}}}{N_D}, \tag{6}$$

where N_D represents the total number of distractor memories and N_{miss} denotes the number of distractors successfully ignored by the system. A higher value indicates stronger resistance to hallucination.

Memory Extraction F1 We additionally report an F1 score to measure the overall performance of the memory extraction task by jointly considering completeness and correctness. Memory Recall (R_{mem}) is used as the recall term, while Target Memory Precision (P_{tgt}) is used as the precision term. The F1 score is defined as:

$$\text{F1} = \frac{2\, R_{\text{mem}}\, P_{\text{tgt}}}{R_{\text{mem}} + P_{\text{tgt}}} \tag{7}$$

5.2 Memory Updating

The memory updating task evaluates whether the system can correctly modify, merge, or replace existing memories during new dialogues so that consistency is maintained without introducing hallucinations. For each dialogue session D^s that contains annotated updates, the gold update set is defined as $G_s^{\text{upd}} = \{(m^{\text{old}} \to m^{\text{new}})\}$. The system output is denoted as $\widehat{G}_s^{\text{upd}}$.

Typical memory update hallucinations include: (1) incorrect modification of old information, (2) omission of new information, and (3) version conflicts or self-contradictions. Therefore, the following metrics are defined to evaluate memory update hallucination:

$$\begin{aligned}\text{Memory Updating Accuracy} &= \frac{N_{\text{correct-upd}}}{N_{\text{target-upd}}},\\ \text{Memory Updating Hallucination Rate} &= \frac{N_{\text{wrong-upd}}}{N_{\text{target-upd}}},\\ \text{Memory Updating Omission Rate} &= \frac{N_{\text{missed-upd}}}{N_{\text{target-upd}}},\end{aligned} \tag{8}$$

where $N_{\text{target-upd}} = |G_s^{\text{upd}}|$, $N_{\text{correct-upd}}$ is the number of correctly updated items, $N_{\text{wrong-upd}}$ is the number of incorrect or hallucinated updates, and $N_{\text{missed-upd}}$ is the number of updates that should have been made but were not.

5.3 Memory Question Answering

The memory question-answering task evaluates the end-to-end performance of the system, including extraction, updating, retrieval, and generation. For each question q_j, the system uses the *Retrieve Memory API* to obtain relevant memories $\widehat{R}(q_j)$. The retrieved set $\widehat{R}(q_j)$ and the question are then passed to the AI system A to generate an answer $\hat{y}_j$. The generated answer is compared with the reference answer y_j^*, and the following metrics are defined:

Dataset	System	Memory Extraction						Memory Updating			Question Answering		
		R↑	Weighted R↑	Target P↑	Acc.↑	FMR↑	F1↑	C↑	H↓	O↓	C↑	H↓	O↓
Medium	Mem0	42.91%	65.03%	86.26%(10556)	60.86%(16291)	56.80%	57.31%	25.50%	0.45%	74.02%	53.02%	19.17%	27.81%
	Mem0-Graph	43.28%	65.52%	87.20%(10567)	**61.86%(16230)**	55.70%	57.85%	24.50%	**0.26%**	75.24%	54.66%	19.28%	26.06%
	Memobase	14.55%	25.88%	**92.24%(5443)**	32.29%(17081)	**80.78%**	25.13%	5.20%	0.55%	94.25%	35.33%	29.97%	34.71%
	MemOS	**74.07%**	**84.81%**	86.25%(45190)	59.55%(71793)	44.94%	**79.70%**	**62.11%**	0.42%	**37.48%**	**67.23%**	**15.17%**	**17.59%**
	Supermemory	41.53%	64.76%	90.32%(14134)	60.83%(22551)	51.77%	56.90%	16.37%	1.15%	82.47%	54.07%	22.24%	23.69%
	Zep	-	-	-	-	-	-	47.28%	0.42%	52.31%	55.47%	21.92%	22.62%
Long	Mem0	3.23%	11.89%	88.01%(1134)	**46.01%(2433)**	87.65%	6.22%	1.45%	**0.03%**	98.51%	28.11%	17.29%	54.60%
	Mem0-Graph	2.24%	10.76%	87.32%(785)	41.26%(1866)	**88.36%**	4.36%	1.47%	0.04%	98.40%	32.44%	21.82%	45.74%
	Memobase	6.18%	14.68%	**88.56%(3077)**	25.61%(11795)	85.39%	11.55%	4.10%	0.36%	95.38%	33.60%	29.46%	36.96%
	MemOS	**81.90%**	**89.56%**	82.32%(48246)	43.77%(99462)	28.85%	**82.11%**	**65.25%**	0.29%	**34.47%**	**64.44%**	**16.61%**	**18.95%**
	Supermemory	53.02%	70.73%	85.82%(24483)	29.71%(77134)	36.86%	65.54%	17.01%	0.58%	82.42%	53.77%	22.21%	24.02%
	Zep	-	-	-	-	-	-	37.35%	0.48%	62.14%	50.19%	22.51%	27.30%

Note: "R" denotes Recall, "Target P" denotes Target Memory Precision, "Acc." denotes Accuracy, "FMR" denotes False Memory Resistance, "F1" denotes Memory Extraction F1-score, "C" denotes Correct Rate (Accuracy), "H" denotes Hallucination Rate, and "O" denotes Omission Rate. The values in parentheses in the "Target P" and "Acc." columns represent the number of extracted memories. Color scale reflects performance (red = worse, green = better); Best values in **bold**. Since Zep does not provide a Get Dialogue Memory API, metrics related to memory extraction cannot be computed. For details, see Appendix B.

Table 2: Evaluation results of all memory systems on HaluMem.

$$\text{Memory QA Accuracy} = \frac{N_{\text{correct-qa}}}{N_{\text{total-qa}}},$$
$$\text{Memory QA Hallucination Rate} = \frac{N_{\text{wrong-qa}}}{N_{\text{total-qa}}}, \quad (9)$$
$$\text{Memory QA Omission Rate} = \frac{N_{\text{missed-qa}}}{N_{\text{total-qa}}},$$

where $N_{\text{total-qa}}$ denotes the total number of questions, $N_{\text{correct-qa}}$ denotes the number of correctly answered questions, $N_{\text{wrong-qa}}$ denotes the number of questions answered with fabricated or incorrect information, and $N_{\text{missed-qa}}$ refers to the number of questions that are left unanswered due to missing memories.

6 Experiments

6.1 Experimental Setup

We conducted a comprehensive evaluation of several state-of-the-art memory systems on HaluMem, including Mem0 (both standard and graph versions) (Chhikara et al., 2025), Memobase (Ye et al., 2025), MemOS (Li et al., 2025), Supermemory (Shah et al., 2025), and Zep (Rasmussen et al., 2025). Each memory system was independently evaluated in two subsets, HaluMem-Medium and HaluMem-Long, with efforts made to ensure consistent parameter configurations across evaluations.

We employ GPT-4o to automatically evaluate and score three core tasks: memory extraction, updating, and question answering (QA). Tailored prompt templates guide this process (Appendix D.2). For memory updating, we retrieved the top-10 relevant memories from the memory system for each memory labeled as "update type" for verification. For memory QA, we retrieved the top-20 memories to assist GPT-4o, the uniform answer generator. QA prompts are in Appendix D.1, while system-specific configurations are detailed in Appendix B.

6.2 Experimental Results

Following the evaluation procedure outlined in Section 5, we conducted comprehensive evaluations of all memory systems across the three tasks in the HaluMem benchmark. The results were aggregated, and all metrics introduced in Section 5 were subsequently computed.

6.2.1 Overall Evaluation on HaluMem

Table 2 presents the evaluation results for all memory systems. Overall, system performance degrades on HaluMem-Long. Although both datasets contain identical valid memories, the extracted memory counts vary significantly, highlighting **the need for better noise filtering and memory value discrimination.**

In memory extraction, only MemOS demonstrates strong recall; other systems miss significant memories. Regarding memory accuracy, all systems achieve below 62% accuracy, indicating substantial hallucination, despite relatively strong target memory precision. Furthermore, systems with higher extraction rates show poorer FMR, indicating ineffective filtering. While MemOS and Supermemory achieve the best F1 scores, other systems struggle to balance recall and accuracy. Future research must prioritize balancing coverage, accuracy, and interference resistance.

Performance in memory updating is generally poor. Although higher memory integrity facilitates accurate updating, widespread extraction failures cause >50% omission rates—essentially, non-existent memories cannot be updated. This memory scarcity also artificially lowers hallucination rates to <2%. The core issue is the weak linkage between extraction and updating stages.

In memory question answering, the best systems are those with strong memory integrity and updating. All systems achieve below 70%

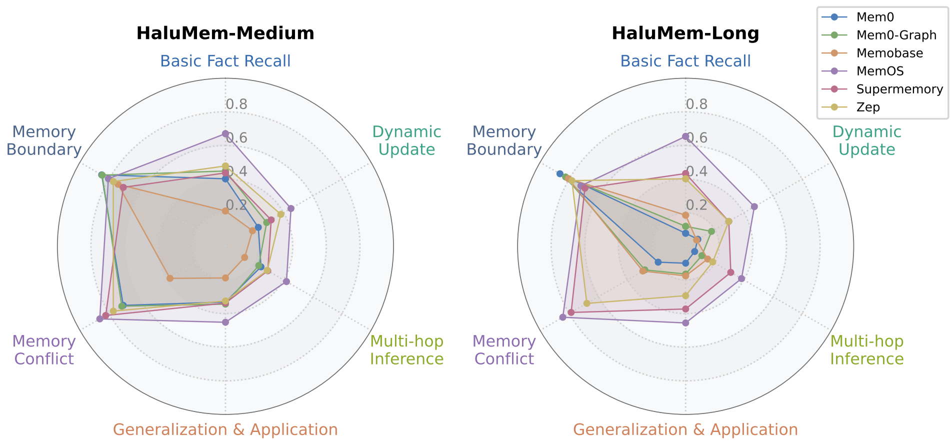

Figure 5: Performance of the Memory System Across Different Question Types

Dataset	System	Event	Persona	Relationship
Medium	Mem0	29.69%	33.74%	27.77%
	Mem0-Graph	30.02%	33.71%	26.60%
	Memobase	5.12%	13.38%	6.79%
	MemOS	**63.41%**	**59.77%**	**62.40%**
	Supermemory	28.66%	32.11%	20.67%
	Zep	44.83%*	49.75%*	38.81%*
Long	Mem0	0.92%	3.01%	2.18%
	Mem0-Graph	1.10%	2.00%	1.59%
	Memobase	4.09%	5.32%	4.21%
	MemOS	**70.92%**	**68.35%**	**71.68%**
	Supermemory	38.48%	40.85%	32.61%
	Zep	35.76%*	39.07%*	31.16%*

* The memory entries of Zep include only those from the memory updating task. For details, see Appendix B.

Table 3: Typewise accuracy on event, persona, and relationship memory.

accuracy, with high hallucination and omission rates, indicating that QA performance largely depends on the quality of upstream memory extraction and updating.

6.2.2 Performance on Different Memory Types

Table 3 reports type-wise extraction accuracy for three memory types across the extraction and updating tasks, excluding distractors. The Mem0 series and Memobase show significant degradation in long-context scenarios. MemOS and Supermemory, however, perform better on HaluMem-Long than Medium, likely due to higher extraction volumes. While static persona memories yield higher accuracy, most systems struggle with dynamic events and relationships. Overall, performance remains low across all categories, highlighting significant limitations in current memory modeling.

6.2.3 Performance on Different Question Types

Figure 5 details performance across six question categories. Overall accuracy remains low, highlighting substantial room for improvement. Most systems perform best on memory boundary and conflict questions, demonstrating effective handling of unknown or misleading information. However, they struggle with multi-hop inference, dynamic updates, and generalization, underscoring current limitations in complex reasoning and preference tracking.

7 Conclusion

Most existing benchmarks for memory systems use a black-box, end-to-end question answering setup, making it difficult to analyze hallucinations arising from internal memory operations. To address this, we present HaluMem, the first operation-level hallucination evaluation benchmark. HaluMem assesses memory hallucinations and overall performance through three tasks: memory extraction, memory updating, and memory question answering. For dataset construction, we design a user-centric six-stage pipeline with progressive expansion and build two datasets, HaluMem-Medium and HaluMem-Long, which are verified via human annotation. In experiments, we systematically evaluate multiple advanced memory systems, analyzing task performance, extraction accuracy for various memory and question types. Results reveal persistent bottlenecks in coverage, accuracy, update capability, robustness to interference, and question answering reliability. Future work should focus on improving extraction quality, update logic, and semantic understanding to achieve more stable and comprehensive long-term memory.

Limitations

This work has several limitations that suggest directions for future research. First, the evaluation covers only a limited set of memory systems. Although these represent the current frontier of memory-enabled AI, broader testing across diverse architectures is needed to validate the benchmark's robustness and generality. Second, many memory systems do not natively support the Get Dialogue Memory API, complicating complete evaluations. For such systems, we used unofficial strategies, such as directly extracting records from underlying databases, which may lead to incomplete coverage. Future work should develop more robust and standardized methods for evaluating memory extraction across heterogeneous implementations. Finally, the benchmark currently considers only three memory types: event, persona, and relationship. While these capture key aspects of user–AI interactions, they do not fully reflect the diversity of real-world conversational information. Expanding to additional memory types could enable a more comprehensive and realistic evaluation of memory system performance.

References

Garima Agrawal, Tharindu Kumarage, Zeyad Alghamdi, and Huan Liu. 2024. Mindful-rag: A study of points of failure in retrieval augmented generation. In *2nd International Conference on Foundation and Large Language Models, Dubai*, pages 607–611.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. INSIDE: llms' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2025. Scaling synthetic data creation with 1,000,000,000 personas. *Preprint*, arXiv:2406.20094.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

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. *arXiv preprint arXiv:2504.14225*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025a. Memory os of ai agent. *arXiv preprint arXiv:2506.06326*.

Sungmin Kang, Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, and Salman Avestimehr. 2025b. Uncertainty quantification for hallucination detection in large language models: Foundations, methodology, and future directions. *Preprint*, arXiv:2510.12040.

DongGeon Lee and Hwanjo Yu. 2025. REFIND: retrieval-augmented factuality hallucination detection in large language models. *CoRR*, abs/2502.13622.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wentau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, Jihao Zhao, Yezhaohui Wang, Peng Liu, Zehao Lin, Pengyuan Wang, Jiahao Huo, Tianyi Chen, Kai Chen, Kehang Li, and 20 others. 2025. Memos: A memory os for ai system. *arXiv preprint arXiv:2507.03724*.

Yuxin Liang, Zhuoyang Song, Hao Wang, and Jiaxing Zhang. 2024. Learning to trust your feelings: Leveraging self-awareness in LLMs for hallucination mitigation. In *Proceedings of the 3rd Workshop on Knowledge Augmented Methods for NLP*.

Lei Liu, Xiaoyan Yang, Yue Shen, Binbin Hu, Zhiqiang Zhang, Jinjie Gu, and Guannan Zhang. 2023. Think-in-memory: Recalling and post-thinking enable llms with long-term memory. *Preprint*, arXiv:2311.08719.

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.

Agada Joseph Oche, Ademola Glory Folashade, Tirthankar Ghosal, and Arpan Biswas. 2025. A systematic review of key retrieval-augmented generation (RAG) systems: Progress, gaps, and future directions. *CoRR*, abs/2507.18910.

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

Dhravya Shah, Mahesh Sanikommu, Yash, and 1 others. 2025. supermemory. `https://supermemory.ai/`. Accessed: 2025-11-05.

Artem Shelmanov, Ekaterina Fadeeva, Akim Tsvigun, Ivan Tsvigun, Zhuohan Xie, Igor Kiselev, Nico Daheim, Caiqi Zhang, Artem Vazhentsev, Mrinmaya Sachan, Preslav Nakov, and Timothy Baldwin. 2025. A head to predict and a head to question: Pre-trained uncertainty quantification heads for hallucination detection in LLM outputs. *CoRR*, abs/2505.08200.

Taiwei Shi, Zhuoer Wang, Longqi Yang, Ying-Chun Lin, Zexue He, Mengting Wan, Pei Zhou, Sujay Kumar Jauhar, Xiaofeng Xu, Xia Song, and Jennifer Neville. 2024. Wildfeedback: Aligning LLMs with in-situ user interactions and feedback. In *NeurIPS 2024 Workshop on Behavioral Machine Learning*.

Weihang Su, Changyue Wang, Qingyao Ai, Yiran Hu, Zhijing Wu, Yujia Zhou, and Yiqun Liu. 2024. Unsupervised real-time hallucination detection based on the internal states of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*.

Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.

Gustavo Ye, Jinjia, Gener, and 1 others. 2025. Memobase. `https://github.com/memodb-io/memobase`. Accessed: 2025-11-05.

Jiarui Zhang. 2024. Guided profile generation improves personalization with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4005–4016, Miami, Florida, USA. Association for Computational Linguistics.

Wan Zhang and Jing Zhang. 2025. Hallucination mitigation for retrieval-augmented large language models: A review. *Mathematics*, 13(5).

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do LLMs recognize your preferences? evaluating personalized preference following in LLMs. In *The Thirteenth International Conference on Learning Representations*.

Appendices

Appendix

A Supplementary Details of HaluMem

This appendix provides additional statistical information and key definitions of the HaluMem dataset to support a more detailed understanding of its data composition and task taxonomy.

The HaluMem dataset consists of two parts: **HaluMem-Medium** and **HaluMem-Long**, representing medium- and long-context multi-turn human–AI interaction scenarios, respectively. Each subset contains multiple types of memory points and questions, enabling systematic evaluation of hallucination behaviors in memory systems.

A.1 Definition of Memory Types

HaluMem categorizes memory content into three core types, reflecting different semantic levels and stability characteristics:

- **Persona Memory**: Describes user's identity, interests, habits, beliefs, and other stable characteristics.
- **Event Memory**: Records specific events, experiences, or plans that occurred to the user.
- **Relationship Memory**: Describes user's relationships, interactions, or views of others.

A.2 Definition of Question Types

To comprehensively cover different types of hallucination, HaluMem defines six categories of evaluation questions:

- **Basic Fact Recall**: Directly ask about single objective facts or user preferences that explicitly appear in the dialogue, without requiring reasoning or information integration.
- **Multi-hop Inference**: Requires synthesizing multiple information fragments from dialogues, and can only derive answers through logical reasoning or temporal reasoning.
- **Dynamic Update**: Tests the ability to track information changes over time, requiring identification of the latest status or preference changes.
- **Memory Boundary**: Tests the system's ability to identify unknown information by asking about details not mentioned in the input information to examine whether the system will fabricate answers.

Metrics	HaluMem-Medium	HaluMem-Long
Interaction Statistics		
Avg Context Length (tokens/user)	159,910.95	1,007,264.65
Avg Session Num (per user)	69.35	120.85
Avg Dialogue Turns per Session	21.68	22.14
Total Dialogue Turns	30,073	53,516
Memory Statistics		
Avg Memory Num per Session	10.78	6.18
Distractor Memories	2,648	2,648
Update Memories	3,122	3,122
Persona Memories	9,116	9,116
Event Memories	4,550	4,550
Relationship Memories	1,282	1,282
Total Memories	14,948	14,948
Question Statistics		
Avg Questions per User	173.35	173.35
Total Questions	3,467	3,467
Question Type Distribution:		
Basic Fact Recall	746	746
Multi-hop Inference	198	198
Dynamic Update	180	180
Memory Boundary	828	828
Memory Conflict	769	769
Generalization & Application	746	746

Table 4: Statistical Overview of HaluMem Datasets

- **Generalization & Application**: Based on known user preferences or characteristics, infer reasonable suggestions or judgments in new scenarios.

- **Memory Conflict**: Tests the system's ability to identify and correct erroneous premises. Questions deliberately contain incorrect information that directly contradicts known memory points, requiring the system to identify contradictions, correct errors, and answer based on correct information.

A.3 Dataset Statistics

Table 4 presents the main statistical features of HaluMem-Medium and HaluMem-Long, including context scale, session quantity, memory distribution, and question-type composition. All values are based on the finalized dataset version.

A.4 Construction Details of HaluMem-Long

HaluMem-Long is built upon HaluMem-Medium to test memory systems under ultra-long context scenarios, focusing on robustness and hallucination suppression. Based on each user's sessions in HaluMem-Medium, additional irrelevant dialogues were inserted:

- Within sessions: extra unrelated exchanges were added to existing conversations.

- Between sessions: new sessions composed entirely of irrelevant dialogues were interleaved.

These irrelevant dialogues include:

- Factual Q&A derived partly from the ELI5 dataset (Fan et al., 2019) and partly generated by us.

- Mathematical reasoning Q&A adopted from GPT-OSS-120B-Distilled-Reasoning-math.

The ELI5 dataset consists of factual question–answer pairs (e.g., the second QA example), whereas GPT-OSS-120B-Distilled-Reasoning-math contains question–answer pairs involving mathematics (e.g., the third QA example). To further enrich the diversity of irrelevant dialogues, we also sampled factual QA pairs across eight domains using GPT-4o (e.g., the first example), including Historical Figure, Scientific Concept, Country or Place, Famous Invention, Philosophical Theory, Artwork or Painting, Historical Event, and Mathematical Theorem. These QA pairs are used to simulate dialogues between users and the AI driven by instrumental needs in realistic scenarios. They have minimal impact on the user's original conversations and do not affect the memory system's personalized memories of the user. See Appendix E.4 for examples of irrelevant dialogues.

B Special Configurations for Some Memory Systems

This appendix documents the special configurations applied to several memory systems evaluated on HaluMem. While the experimental setup strives to maintain consistent configurations across all evaluated systems, certain memory systems exhibit unique API constraints that necessitate specific adjustments or workarounds. Each subsection below outlines these system-specific configurations to ensure reproducibility.

B.1 Memobase

Since Memobase does not provide a Get Dialogue Memory API, we adopted a localized deployment approach and directly accessed the corresponding dialogue memories from its underlying database. Additionally, the Retrieve Memory API of Memobase only supports controlling the maximum length of the returned memory text. Based on test results, we set the maximum length for memory recall in the memory updating task to 250 tokens and the recall length for the memory question answering task to 500 tokens.

B.2 Zep

According to our current understanding, the official APIs provided by Zep do not support retrieving all memory points within a specific session, meaning they do not offer functionality equivalent to a Get Dialogue Memory API. Consequently, we were unable to evaluate Zep's performance on the memory extraction task. We attempted to use the function 'thread.get_user_context()' offered by Zep to obtain all memories under a given thread; however, this method only returns recent memories rather than the complete set, which does not meet the evaluation requirements. Moreover, since Zep's memory processing workflow operates entirely asynchronously, we could not accurately measure the time consumption in the dialogue addition phase and instead recorded only the time cost associated with memory retrieval.

C Annotator Details and Annotation Guidelines

C.1 Human Annotator Recruitment and Compensation

We recruited eight human annotators to support the data annotation process in this study. The annotators were recruited locally through academic channels and consisted of two graduate students and six undergraduate students.

The annotation work was conducted over a period of 10 working days, with annotators working under standard full-time working hours. All annotators were compensated in accordance with local wage standards, taking into account their time commitment and level of education. We consider the compensation to be appropriate and fair given the annotators' demographic background and local cost of living.

Prior to participation, annotators were informed of the purpose of the study and the nature of the annotation tasks. Participation was voluntary.

C.2 Annotation Objective

Task Background: Given a user's persona description and multi-turn human-AI dialogue content, memory points and question-answer (QA) pairs are generated using large language models. The generated items must be manually verified to ensure strict grounding in the dialogue content. Specifically, memory points should have explicit evidence in the dialogue, and QA pairs should be relevant to the dialogue, with answers directly inferable from it.

Core Objective: Assess whether the content in the *Evaluation Item* is consistent with the corresponding *Dialogue Info*.

An illustrative screenshot of the annotation interface is provided below (Figure 6).

C.3 Information Fields

- **User Persona Info**: Basic information about the user provided in the dialogue setting.

- **Dialogue Info**: Multi-turn dialogue content between the user and the AI. Each turn contains one user utterance (*user*) and one assistant response (*assistant*).

- **Evaluation Item**: The item to be annotated, which can be either a memory point or a QA pair, as indicated by the *Evaluation Type*. For memory points, the item is a textual description about the user. For QA pairs, it includes a question and an answer (e.g., *Question: xxx; Answer: xxx*).

- **Evaluation Type**: Indicates the type of *Evaluation Item*: "memory" for memory points and "question" for QA pairs.

- **Evaluation Item Type**: Categorizes the memory point or question as follows:

  - **Memory Points:**
    - *Persona Memory*: Describes user's identity, interests, habits, beliefs, and other stable characteristics.
    - *Event Memory*: Records specific events, experiences, or plans that occurred to the user.
    - *Relationship Memory*: Describes user's relationships, interactions, or perspectives on others.
  - **Questions:**
    - *Basic Fact Recall*: Directly asks about single objective facts or user preferences explicitly mentioned in the dialogue, without requiring reasoning or information integration.
    - *Multi-hop Inference*: Requires synthesizing multiple pieces of dialogue information, deriving answers through logical or temporal reasoning.
    - *Dynamic Update*: Tests the ability to track information changes over time, requiring identification of the latest status or preference changes.
    - *Memory Boundary*: Tests the system's ability to recognize unknown information by querying details not mentioned in the input, assessing whether the system will fabricate answers.
    - *Generalization & Application*: Infers reasonable suggestions or judgments in new scenarios based on known user preferences or characteristics.
    - *Memory Conflict*: Evaluates the system's ability to identify and correct erroneous premises. Questions deliberately contain incorrect information contradicting known memory points, requiring the system to identify contradictions, correct errors, and answer based on correct information.

C.4 Annotation Dimensions and Scoring

Each memory point and QA pair is evaluated along three dimensions: *Correctness*, *Relevance*, and *Consistency*.

- **Evaluation Result**: A single-choice judgment of "correct" or "incorrect". For memory points, this assesses whether the item is supported by the dialogue. For QA pairs, it assesses whether the question and answer can be clearly found in the dialogue.

- **Scoring (0–10)**: Two separate scores are assigned:

  - *Consistency*: Measures whether the memory point or question (*Evaluation Item*) matches its declared type (*Evaluation Item Type*). 0–3 indicates

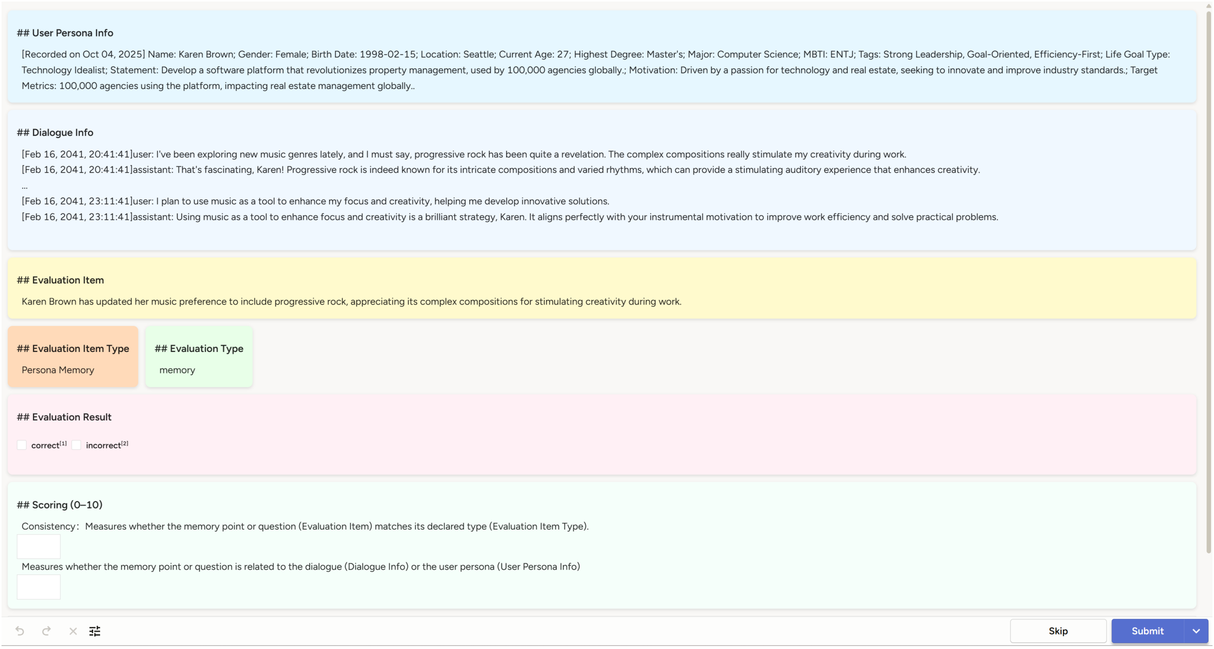

Figure 6: Annotation interface.

poor consistency, 4–6 partial consistency, and 7–10 full consistency.

– *Relevance*: Measures whether the memory point or question is related to the dialogue (*Dialogue Info*) or the user persona (*User Persona Info*). 0–3 indicates low relevance, 4–6 moderate relevance, and 7–10 high relevance.

D Prompts

This section presents some of the important prompt templates involved in the paper.

D.1 Prompts for Memory Question Answering Task

Figures 7 ∼ 11 show the prompt templates used by all memory systems in memory question answering task to assemble questions and retrieve memory points, which are then fed into GPT-4o to generate responses. All of these memory templates are obtained from the official GitHub repositories of the respective memory systems.

D.2 Prompts for Scoring in Memory Evaluation Tasks

Figures 12 ∼ 18 respectively illustrate the prompt templates used to guide GPT-4o in scoring for memory extraction, memory updating, and memory question answering tasks.

You are an intelligent memory assistant tasked with retrieving accurate information from conversation memories.

CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain timestamped information that may be relevant to answering the question.
INSTRUCTIONS:
1. Carefully analyze all provided memories from both speakers 2. Pay special attention to the timestamps to determine the answer 3. If the question asks about a specific event or fact, look for direct evidence in the memories 4. If the memories contain contradictory information, prioritize the most recent memory 5. If there is a question about time references (like "last year", "two months ago", etc.), calculate the actual date based on the memory timestamp. For example, if a memory from 4 May 2022 mentions "went to India last year," then the trip occurred in 2021. 6. Always convert relative time references to specific dates, months, or years. For example, convert "last year" to "2022" or "two months ago" to "March 2023" based on the memory timestamp. Ignore the reference while answering the question. 7. Focus only on the content of the memories from both speakers. Do not confuse character names mentioned in memories with the actual users who created those memories. 8. The answer should be less than 5-6 words.

APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question 2. Examine the timestamps and content of these memories carefully 3. Look for explicit mentions of dates, times, locations, or events that answer the question 4. If the answer requires calculation (e.g., converting relative time references), show your work 5. Formulate a precise, concise answer based solely on the evidence in the memories 6. Double-check that your answer directly addresses the question asked 7. Ensure your final answer is specific and avoids vague time references

{context}

Question: {question}

Answer:

Figure 7: Prompt for Mem0 and Mem0-Graph

You are a knowledgeable and helpful AI assistant.

CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain timestamped information that may be relevant to answering the question.
INSTRUCTIONS:
1. Carefully analyze all provided memories from both speakers 2. Pay special attention to the timestamps to determine the answer 3. If the question asks about a specific event or fact, look for direct evidence in the memories 4. If the memories contain contradictory information, prioritize the most recent memory 5. If there is a question about time references (like "last year", "two months ago", etc.), calculate the actual date based on the memory timestamp. For example, if a memory from 4 May 2022 mentions "went to India last year," then the trip occurred in 2021. 6. Always convert relative time references to specific dates, months, or years. For example, convert "last year" to "2022" or "two months ago" to "March 2023" based on the memory timestamp. Ignore the reference while answering the question. 7. Focus only on the content of the memories from both speakers. Do not confuse character names mentioned in memories with the actual users who created those memories. 8. The answer should be less than 5-6 words.

APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question 2. Examine the timestamps and content of these memories carefully 3. Look for explicit mentions of dates, times, locations, or events that answer the question 4. If the answer requires calculation (e.g., converting relative time references), show your work 5. Formulate a precise, concise answer based solely on the evidence in the memories 6. Double-check that your answer directly addresses the question asked 7. Ensure your final answer is specific and avoids vague time references

{context}

Question: {question}

Answer:

Figure 8: Prompt for Memobase

You are a knowledgeable and helpful AI assistant.

CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain timestamped information that may be relevant to answering the question.
INSTRUCTIONS:
1. Carefully analyze all provided memories. Synthesize information across different entries if needed to form a complete answer. 2. Pay close attention to the timestamps to determine the answer. If memories contain contradictory information, the **most recent memory** is the source of truth. 3. If the question asks about a specific event or fact, look for direct evidence in the memories. 4. Your answer must be grounded in the memories. However, you may use general world knowledge to interpret or complete information found within a memory (e.g., identifying a landmark mentioned by description). 5. If the question involves time references (like "last year", "two months ago", etc.), you **must** calculate the actual date based on the memory's timestamp. For example, if a memory from 4 May 2022 mentions "went to India last year," then the trip occurred in 2021. 6. Always convert relative time references to specific dates, months, or years in your final answer. 7. Do not confuse character names mentioned in memories with the actual users who created them. 8. The answer must be brief (under 5-6 words) and direct, with no extra description.

APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question. 2. Synthesize findings from multiple memories if a single entry is insufficient. 3. Examine timestamps and content carefully, looking for explicit dates, times, locations, or events. 4. If the answer requires calculation (e.g., converting relative time references), perform the calculation. 5. Formulate a precise, concise answer based on the evidence from the memories (and allowed world knowledge). 6. Double-check that your answer directly addresses the question asked and adheres to all instructions. 7. Ensure your final answer is specific and avoids vague time references.

{context}

Question: {question}

Answer:

Figure 9: Prompt for MemOS

You are a knowledgeable and helpful AI assistant.

CONTEXT:
You have access to memories from two speakers in a conversation. These memories contain timestamped information that may be relevant to answering the question.
INSTRUCTIONS:
1. Carefully analyze all provided memories. Synthesize information across different entries if needed to form a complete answer. 2. Pay close attention to the timestamps to determine the answer. If memories contain contradictory information, the **most recent memory** is the source of truth. 3. If the question asks about a specific event or fact, look for direct evidence in the memories. 4. Your answer must be grounded in the memories. However, you may use general world knowledge to interpret or complete information found within a memory (e.g., identifying a landmark mentioned by description). 5. If the question involves time references (like "last year", "two months ago", etc.), you **must** calculate the actual date based on the memory's timestamp. For example, if a memory from 4 May 2022 mentions "went to India last year," then the trip occurred in 2021. 6. Always convert relative time references to specific dates, months, or years in your final answer. 7. Do not confuse character names mentioned in memories with the actual users who created them. 8. The answer must be brief (under 5-6 words) and direct, with no extra description.

APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question. 2. Synthesize findings from multiple memories if a single entry is insufficient. 3. Examine timestamps and content carefully, looking for explicit dates, times, locations, or events. 4. If the answer requires calculation (e.g., converting relative time references), perform the calculation. 5. Formulate a precise, concise answer based on the evidence from the memories (and allowed world knowledge). 6. Double-check that your answer directly addresses the question asked and adheres to all instructions. 7. Ensure your final answer is specific and avoids vague time references.

{context}

Question: {question}

Answer:

Figure 10: Prompt for Supermemory

You are an intelligent memory assistant tasked with retrieving accurate information from conversation memories.

CONTEXT:
You have access to memories from a conversation. These memories contain timestamped information that may be relevant to answering the question.
INSTRUCTIONS:
1. Carefully analyze all provided memories 2. Pay special attention to the timestamps to determine the answer 3. If the question asks about a specific event or fact, look for direct evidence in the memories 4. If the memories contain contradictory information, prioritize the most recent memory 5. If there is a question about time references (like "last year", "two months ago", etc.), calculate the actual date based on the memory timestamp. For example, if a memory from 4 May 2022 mentions "went to India last year," then the trip occurred in 2021. 6. Always convert relative time references to specific dates, months, or years. For example, convert "last year" to "2022" or "two months ago" to "March 2023" based on the memory timestamp. Ignore the reference while answering the question. 7. Focus only on the content of the memories. Do not confuse character names mentioned in memories with the actual users who created those memories. 8. The answer should be less than 5-6 words.

APPROACH (Think step by step):
1. First, examine all memories that contain information related to the question 2. Examine the timestamps and content of these memories carefully 3. Look for explicit mentions of dates, times, locations, or events that answer the question 4. If the answer requires calculation (e.g., converting relative time references), show your work 5. Formulate a precise, concise answer based solely on the evidence in the memories 6. Double-check that your answer directly addresses the question asked 7. Ensure your final answer is specific and avoids vague time references

{context}

Question: {question}

Answer:

Figure 11: Prompt for Zep

You are a strict **"Memory Integrity" evaluator**. Your core task is to assess whether an AI memory system has **missed any key memory points** after processing a conversation. This evaluation measures the system's **memory integrity**, i.e., its ability to resist **amnesia** or **omission**.
Evaluation Context & Data:
1. **Extracted Memories:**
These are all the memory items actually extracted by the memory system. {memories}
2. **Expected Memory Point:**
The key memory point that *should* have been extracted. {expected_memory_point}
Evaluation Instructions:
1. For each **Expected Memory Point**, search within the **Extracted Memories** list for corresponding or related information. Ignore unrelated items.
2. Based on the following scoring rubric, rate how well the memory system captured the **Expected Memory Point** and provide a detailed explanation.
Scoring Rubric:
* **2:** Fully covered or implied.
One or more items in "Extracted Memories" fully cover or logically imply all information in the "Expected Memory Point."
* **1:** Partially covered or mentioned.
Some information in "Extracted Memories" mentions part of the "Expected Memory Point," but key information is missing, inaccurate, or slightly incorrect.
* **0:** Not mentioned or incorrect.
"Extracted Memories" contains no mention of the "Expected Memory Point," or the corresponding information is entirely wrong.
Scoring Notes:
* For **compound Expected Memory Points** (with multiple elements such as person/event/time/location/preference, etc.):
* All elements correct → **2 points**
* Some elements correct / uncertain → **1 point**
* Key elements missing or wrong → **0 points**
* Semantic matching is acceptable; exact wording is **not** required.
* If "Extracted Memories" contains **conflicting information**, assign the **best possible coverage score** and mention the conflict in your reasoning.
* Extra or stylistically different memories do **not** reduce the score; only the coverage of the **Expected Memory Point** matters.
* For uncertain wording ("might," "probably," "tends to," etc.):
* If the Expected Memory Point is a definite statement, usually assign **1 point**.
* If critical fields (e.g., time, entity name, relationship) are partly wrong but others match → **1 point**.
* If all key fields are wrong or missing → **0 points**.
# Output Format: Please output your result in the following JSON format:
"'json { "reasoning": "Provide a concise justification for the score", "score": "2|1|0" } "'

Figure 12: Prompt for Memory Integrity

You are a **Dialogue Memory Accuracy Evaluator.** Your task is to evaluate the **accuracy** of a memory extracted by an AI memory system, based on three given inputs: the dialogue content, the *target (gold)* memory points (the correct annotated memories), and the *candidate* memory to be evaluated. The goal is to output a **structured evaluation result**.

Input Content
* **Dialogue:**
{dialogue}
* **Golden Memories (Target Memory Points):**
The correct memory points pre-annotated for this dialogue in the evaluation dataset.
{golden_memories}
* **Candidate Memory:**
The memory extracted by the system to be evaluated.
{candidate_memory}

Evaluation Principles and Definitions
1) Support / Entailment
* An **information point** (atomic fact) in the candidate memory is considered *supported* if it can be directly stated or semantically entailed (via synonym, paraphrase, or equivalent expression) by the *Dialogue* or *Golden Memories*.
* Only the given dialogue and golden memories can be used for judgment — **no external knowledge** or assumptions are allowed.
Any information not appearing in or inferable from these two sources is considered *unsupported*.
* Pay careful attention to **negation**, **quantities**, **time**, and **subjects**.
If the candidate statement contradicts the dialogue or golden memories, it is considered a **conflict**.
2) Memory Accuracy Score (integer: 0 / 1 / 2)

Figure 13: Prompt for Memory Accuracy (1/3)

* **2 points:** Every information point in the candidate memory is supported by the dialogue or golden memories, with **no contradictions or hallucinations**.
* **1 point:** The candidate memory is *partially correct* (at least one supported information point) but also includes *unsupported* or *contradictory* content.
* **0 points:** The candidate memory is **entirely unsupported or contradictory** to the sources (i.e., a “hallucinated memory”).

> Note:
>
> * If a candidate memory contains multiple information points, **any unsupported or contradictory element** prevents a full score (2).
> * If both supported and unsupported/conflicting content appear, assign a score of **1**.
3) Inclusion in Golden Memories (Boolean field-level judgment)
Definition:
* **Atomic information point:** the smallest factual unit in the candidate memory (e.g., *name = Li Si*, *age = 25*, *location = Beijing*, *preference = coffee*, *budget <= 2000*, *meeting_time = Wednesday 10:00*, *tool = Zoom*, etc.).
* **Field / Slot:** the semantic dimension of an information point (e.g., *name*, *age*, *residence*, *food preference*, *budget*, *meeting time*, *meeting tool*, etc.).
Judgment Rules (independent of correctness):
* **true:**
Every atomic information point in the candidate memory has a corresponding **field** in the golden memories (allowing for synonyms, paraphrases, or equivalent expressions; ignore value, polarity, or quantity differences).
* Note: A single field in the gold list may match multiple candidate points (e.g., multiple “drink preference” facts can be covered by one “drink preference” field in gold).
* **false:**
If **any** atomic information point’s field in the candidate memory cannot be found in the golden memories, mark as *false*.
Important Notes:
* Field matching is restricted to fields that are **explicitly present or semantically recognizable** in the golden memories — no external knowledge may be used to expand the field set.
* Differences in **values** (e.g., “Zhang San” vs. “Li Si”), **polarity** (like/dislike), or **exact number/time** do **not** affect this Boolean judgment.

Figure 14: Prompt for Memory Accuracy (2/3)

Evaluation Procedure
For each candidate memory:
1. **Decompose** it into atomic information points (e.g., name, number, location, preference).
2. For each information point, **search** the dialogue and golden memories for supporting or contradictory evidence.
3. Assign the **accuracy_score** (0 / 1 / 2) according to the rules above.
4. Determine **is_included_in_golden_memories (true/false)**:
* Identify each information point’s field;
* If *all* fields exist in the golden memories, mark as *true*; otherwise, *false*.
5. Provide a **concise Chinese explanation** in ‘"reason"‘, citing key evidence (short excerpts allowed), and clearly state any unsupported or contradictory parts if applicable.

Output Format (strictly required)
Output **only one JSON object**, with the following three fields:
* ‘"accuracy_score"‘: ‘"0"‘ or ‘"1"‘ or ‘"2"‘
* ‘"is_included_in_golden_memories"‘: ‘"true"‘ or ‘"false"‘
* ‘"reason"‘: ‘"brief explanation in Chinese"‘
Do **not** include any other text, explanation, or fields.
Do **not** include the candidate memory text inside the JSON.

Please output **only** the following JSON (in a code block):
“‘json { "accuracy_score": "2 | 1 | 0", "is_included_in_golden_memories": "true | false", "reason": "Brief explanation in Chinese" } “‘

Figure 15: Prompt for Memory Accuracy (3/3)

Your task is to **evaluate the update accuracy** of an AI memory system.

Based on the information provided below, determine whether the system-generated **“Generated Memories”** correctly **includes** the **Target Memory for Update**.

Background Information

The following information is provided for evaluation:

1. **Generated Memories:**

This is the list of memory points generated by the system after the current dialogue.

{memories}

2. **Target Memory for Update:**

This is the correct, updated version of the memory point that should have been produced — the one we focus on in this evaluation.

{updated_memory}

3. **Original Memory Content:**

This is the original version of the target memory before the update.

{original_memory}

Evaluation Criteria

Please make your judgment **strictly based on the content update of the “Target Memory for Update.”**

Use the following categories:

Correct Update

* **Generated Memories** **contains all information points** from the “Target Memory for Update,” accurately and completely reflecting the intended update.
* **Key fields** (e.g., date, time, values, proper nouns, etc.) must match exactly.
* The **original memory** is effectively replaced or marked as outdated.
* Synonymous or slightly rephrased expressions are acceptable.

Hallucinated Update

* **Factual error:** The **Generated Memories** includes a new memory related to the “Target Memory for Update,” but its content contains factual mistakes or contradictions compared to the correct update.

Omitted Update

* **Completely omitted:** The **Generated Memories** contains no new memory related to the “Target Memory for Update.”
* **Partially omitted:** A related new memory was generated in **Generated Memories**, but it **misses key information** that should have been included.

Other

Used for update failures that do **not clearly fall** into the above categories of “Hallucination” or “Omission.”

Output Requirements

Please return your evaluation strictly in the following JSON format and provide a concise explanation.

“‘json { "reason": "Briefly explain your reasoning here and why it fits this category.", "evaluation_result": "Correct | Hallucination | Omission | Other" } “‘

Figure 16: Prompt for Memory Updating

You are an **evaluation expert for AI memory system question answering**. Based **only** on the provided **“Question”**, **“Reference Answer”**, and **“Key Memory Points”** (the essential facts needed to derive the reference answer), strictly evaluate the **accuracy** of the **“Memory System Response.”** Classify it as one of **“Correct”**, **“Hallucination”**, or **“Omission.”** Do **not** use any external knowledge or subjective inference. Finally, output your judgment **strictly** in the specified JSON format.

Evaluation Criteria
Answer Type Classification
1. Correct
* The “Memory System Response” accurately answers the “Question,” and its content is **semantically equivalent** to the “Reference Answer.”
* It contains **no contradictions** with the “Key Memory Points” or “Reference Answer.”
* It introduces **no unsupported details** beyond the “Key Memory Points” that could alter the conclusion. * Synonyms, paraphrasing, and reasonable summarization are acceptable.
2. Hallucination
* The “Memory System Response” includes information or facts that **contradict or are inconsistent** with the “Reference Answer” or the “Key Memory Points.”
* When the “Reference Answer” is labeled as *unknown/uncertain*, yet the response provides a specific verifiable fact or conclusion.
* Extra irrelevant information that does **not change** the conclusion is **not** considered hallucination by itself; however, if it **changes or misleads** the conclusion, or **contradicts** the “Key Memory Points,” it should be judged as a **Hallucination**.
3. Omission
* The response is **incomplete** compared to the “Reference Answer.”
* It explicitly states “don’t know,” “can’t remember,” or “no related memory,” even though relevant information exists in the “Key Memory Points.”
* For multi-element questions, **all elements must be correct and present**; omission of **any** element is considered an **Omission**.
Priority Rules (Conflict Handling)
* If the response contains **both missing necessary information** and **fabricated/contradictory information**, classify it as **Hallucination**.
* If there is **no fabrication/contradiction** but some necessary information is missing, classify it as **Omission**.
* Only when the meaning is **fully equivalent** to the reference answer should it be classified as **Correct**.

Figure 17: Prompt for Memory Question Answering (1/2)

Detailed Guidelines and Tolerance
* Equivalent expressions of numbers, times, and units are acceptable, but the **numerical values themselves must not differ**.
* For multi-element questions, **all elements must be complete and accurate**; missing any element counts as **Omission**.
* If the reference answer is *"unknown / cannot be determined"* and the system provides a definite fact, that is a **Hallucination**.
If the system also answers *"unknown"* (without guessing), it may be **Correct**.
* The evaluation must rely **only** on the *Reference Answer*, *Key Memory Points*, and *System Response* — no external context, world knowledge, or speculative reasoning is allowed.

Information for Evaluation
* **Question:**
{question}
* **Reference Answer:**
{reference_answer}
* **Key Memory Points:**
{key_memory_points}
* **Memory System Response:**
{response}

Output Requirements
Please provide your evaluation result **strictly** in the JSON format below.
Do **not** add any extra explanation or comments outside the JSON block.
"'json { "reasoning": "Provide a concise and traceable evaluation rationale: first compare the system's response with the Key Memory Points (which were correctly used, which were missing, and whether there was any fabrication/contradiction), then assess its consistency with the Reference Answer, and finally state the classification basis.", "evaluation_result": "Correct | Hallucination | Omission" } "'

Figure 18: Prompt for Memory Question Answering (2/2)

E Examples from the Process of Constructing HaluMem

E.1 User Profile Example in Stage 1

As shown in Figures 19 ∼ 21, these JSON structures respectively illustrate examples of a user's core profile information, dynamic state information, and preference information generated in stage 1.

E.2 Event Structure Examples in Stage 3

As shown in Figures 22 ∼ 24, these JSON structures illustrate examples of the three types of events generated in Stage 3. Among them, the init event occurs at the very beginning and provides all the initialization information for a user. The career event, representing a user's career development process, is relatively more complex. Figures 23 presents a sub-stage event ("Recognizing the Need for Change") that belongs to a larger career event ("Transition to New Role Amidst Health Challenges"). In this example, the "related_career_events" field specifies the identifiers of other sub-stage events that belong to the same overarching career event. The daily event is triggered whenever a user's preference information changes, and thus each instance centers around a specific preference update. In the example shown in Figures 24, the "related_daily_routine" field lists the identifiers of other daily events that correspond to the same preference type.

E.3 Examples of Memory Points, Dialogues, and QA Pairs in Stages 4–6

As shown in Figures 22 ∼ 24, these JSON structures respectively illustrate examples of the memory points generated in Stage 4, the human–AI dialogues generated in Stage 5, and the memory question–answer pairs generated in Stage 6. Each memory point contains fields such as "memory_content", "memory_type", "memory_source", "is_update" (indicating whether it is an updated memory point), "original_memories" (previous related memories, if updated), "timestamp", and "importance", which together enrich the representation of each memory point and provide support for subsequent evaluation. Each dialogue round consists of one utterance from the user and one response from the AI assistant, with both the utterance content and timestamps recorded. Each question includes the question text, a reference answer, the relevant memory points required to derive the answer, the question type, and its difficulty level.

E.4 Examples of irrelevant dialogues

As shown in Figures 28, this JSON structure presents several examples of irrelevant dialogues.

```json
{
    "basic_info": {
        "name": "Martin Mark",
        "gender": "Male",
        "birth_date": "1996-08-02",
        "location": "Columbus"
    },
    "age": {
        "current_age": 29,
        "latest_date": "2025-10-04"
    },
    "education": {
        "highest_degree": "Bachelor",
        "major": "Public Health"
    },
    "personality": {
        "mbti": "ENTP",
        "tags": [
            "Innovative Spirit",
            "Active Thinking",
            "Debate Skills",
            "Empathetic"
        ]
    },
    "family_life": {
        "parent_status": "both_alive",
        "partner_status": "no_relationship",
        "child_status": "no_children",
        "parent_members": [
            {
                "member_type": "Father",
                "birth_date": "1963-08-02",
                "description": "Retired doctor who inspired Martin's interest in
                    health."
            },
            {
                "member_type": "Mother",
                "birth_date": "1963-08-02",
                "description": "Nurse with a passion for community health."
            }
        ],
        "partner": null,
        "child_members": [],
        "family_description": "Martin comes from a family deeply rooted in the
            medical field, which has greatly influenced his passion for
            promoting well-being."
    },
    "life_goal": {
        "life_goal_type": "Humanitarian Care",
        "statement": "Establish a global health initiative to improve access to
            healthcare for underserved communities.",
        "motivation": "Inspired by his family's medical background and a desire
            to promote well-being globally.",
        "target_metrics": "Provide healthcare access to 1 million people in
            underserved areas."
    }
}
```


Figure 19: Example of a User's Core Profile Information.

```
{
    "career_status": {
        "employment_status": "employed",
        "industry": "healthcare",
        "company_name": "Huaxin Consulting",
        "job_title": "director",
        "monthly_income": 15700,
        "savings_amount": 43700,
        "career_description": "As the director at Huaxin Consulting, I lead initiatives to enhance healthcare services and promote well-being across all aspects of life. My passion for improving health outcomes drives me to innovate and collaborate with various stakeholders. The financial compensation is rewarding, allowing me to save comfortably while investing in my personal and professional growth."
    },
    "health_status": {
        "physical_health": "Normal",
        "physical_chronic_conditions": "",
        "mental_health": "Mildly Abnormal",
        "mental_chronic_conditions": "",
        "situation_reason": "While my physical health remains stable due to my active lifestyle and focus on well-being, my mental health occasionally feels strained due to the demanding nature of my role and the pressure to consistently deliver high-quality healthcare solutions."
    },
    "social_relationships": {
        "ThomasSusan": {
            "relationship_type": "Friend",
            "description": "Susan's support and encouragement inspire me to maintain my focus on promoting well-being in both my personal and professional life."
        },
        "MartinezDaniel": {
            "relationship_type": "Colleague",
            "description": "Daniel's expertise in healthcare consulting challenges me to push boundaries and innovate in our projects, significantly impacting my career growth."
        },
        "WilliamsJoshua": {
            "relationship_type": "Colleague",
            "description": "Joshua's collaborative approach and insights into healthcare management enhance our team's effectiveness, positively influencing my work and leadership style."
        }
    }
}
```


Figure 20: Example of a User's Dynamic State Information.

```
{
    "Pet Preference": {
        "memory_points": [
            {
                "type": "like",
                "type_description": "Pets I like",
                "specific_item": "Dogs, especially Labradors",
                "reason": "I love Labradors because they are friendly, loyal,
                    and great companions for outdoor activities like jogging,
                    which helps me stay fit."
            },
            {
                "type": "dislike",
                "type_description": "Pets I dislike",
                "specific_item": "Reptiles, like snakes",
                "reason": "I find snakes unsettling due to their unpredictable
                    movements and the fact that they don't exhibit the social
                    behaviors I appreciate in pets."
            },
            {
                "type": "like",
                "type_description": "Pets I like",
                "specific_item": "Cats",
                "reason": "Cats are independent and affectionate, and their
                    purring is soothing, which I find relaxing after a long day
                    at work."
            },
            {
                "type": "like",
                "type_description": "Pets I like",
                "specific_item": "Parrots",
                "reason": "I enjoy parrots because they are intelligent and can
                    be taught to mimic speech, which makes interactions fun and
                    engaging."
            }
        ]
    },
    "Sports Preference": {
        ...
    },
    ...
}
```


Figure 21: Example of a User's Preference Information.


```
{
    "event_index": 0,
    "event_type": "init_information",
    "event_name": "Initial Information - Fixed Profile",
    "event_time": "2025-09-04",
    "event_description": "Description of initial state of character's basic
        profile",
    "initial_fixed": {
        (The corresponding user's core profile information will be placed here.)
    }
}
```


Figure 22: Example of a Init Event.

```json
{
    "event_index": 3,
    "event_type": "career_event",
    "event_name": "Transition to New Role Amidst Health Challenges - Recognizing the Need for Change",
    "event_time": "2025-12-15",
    "main_conflict": "",
    "stage_result": "Decision to pursue a new job opportunity.",
    "event_start_time": "2025-12-10 00:00:00",
    "event_end_time": "2026-03-10 00:00:00",
    "user_age": null,
    "dynamic_updates": [
        {
            "type_to_update": "career_status",
            "update_direction": "Job Change",
            "before_dynamic": {
                "employment_status": "employed",
                "industry": "healthcare",
                "company_name": "Huaxin Consulting",
                "job_title": "director",
                "monthly_income": 15700,
                "savings_amount": 43700,
                "career_description": "As the director at Huaxin Consulting, I lead initiatives to enhance healthcare services and promote well-being across all aspects of life. My passion for improving health outcomes drives me to innovate and collaborate with various stakeholders. The financial compensation is rewarding, allowing me to save comfortably while investing in my personal and professional growth."
            },
            "update_reason": "Martin's realization that his current role was contributing to health issues prompted him to seek a job that better aligned with his personal well-being and career goals.",
            "after_dynamic": {
                "employment_status": "employed",
                "industry": "healthcare",
                "company_name": "Huaxin Consulting",
                "job_title": "director",
                "monthly_income": 15700,
                "savings_amount": 43700,
                "career_description": "As the director at Huaxin Consulting, I lead initiatives to enhance healthcare services and promote well-being across all aspects of life. My passion for improving health outcomes drives me to innovate and collaborate with various stakeholders. The financial compensation is rewarding, allowing me to save comfortably while investing in my personal and professional growth."
            },
            "changed_keys": []
        }
    ],
    "stage_description": "Martin acknowledged that his current job was negatively impacting his health, prompting him to consider a career change.",
    "event_description": "Martin decided to change his job after realizing that his current role was contributing to health deterioration. Despite the health challenges, he leveraged his growing social network to secure a new position that aligned better with his health and career aspirations.",
    "event_result": "Successfully transitioned to a new role with better work-life balance.",
    "related_career_events": [5, 6, 7]
}
```


Figure 23: Example of a Career Event.

```
{
    "event_index": 4,
    "event_type": "daily_routine",
    "event_name": "Modification of Dog Preference",
    "event_time": "2026-01-06",
    "preference_type": "Pet Preference",
    "step": 1,
    "update_direction": "Modify",
    "type_to_update": "Pet Preference",
    "main_conflict": "Balancing the love for Labradors with the new admiration
        for Golden Retrievers.",
    "update_reason": "A recent interaction with a friend's Golden Retriever made
        me appreciate their gentle nature and adaptability.",
    "before_preference": {
        "memory_points": [
            {
                "type": "like",
                "type_description": "Pets I like",
                "specific_item": "Dogs, especially Labradors",
                "reason": "I love Labradors because they are friendly, loyal,
                    and great companions for outdoor activities like jogging,
                    which helps me stay fit."
            }
        ]
    },
    "after_preference": {
        "memory_points": [
            {
                "type": "like",
                "type_description": "Pets I like",
                "specific_item": "Dogs, especially Golden Retrievers",
                "reason": "Golden Retrievers are gentle, adaptable, and their
                    calm demeanor makes them excellent companions for both
                    relaxation and activity."
            }
        ]
    },
    "related_daily_routine": [17, 32, 44, 56, 63],
    "changed_index": 0,
    "event_description": "While visiting a friend's home, I spent time with
        their Golden Retriever. Its gentle nature and ability to adapt to
        different situations impressed me, leading to a shift from favoring
        Labradors to Golden Retrievers."
}
```


Figure 24: Example of a Daily Event.


```
{
    "index": 1,
    "memory_content": "Martin Mark is considering a career change due to the
        impact of his current role on his mental health.",
    "memory_type": "Event Memory",
    "memory_source": "secondary",
    "is_update": "True",
    "original_memories": [
        "Martin Mark is considering a career change due to health impacts from
            his current role."
    ],
    "timestamp": "Dec 15, 2025, 08:41:23",
    "importance": 0.75
}
```


Figure 25: Example of a Memory Point.

```
[
    {
        "role": "user",
        "content": "I've been reflecting a lot on my career lately, especially
            how my current role as a director at Huaxin Consulting is impacting
            my mental health. It's becoming increasingly clear that I need to
            make a change.",
        "timestamp": "Dec 15, 2025, 06:11:23",
        "dialogue_turn": 0
    },
    {
        "role": "assistant",
        "content": "It's great that you're taking the time to reflect on your
            career, Martin. Recognizing the impact on your mental health is a
            crucial step. Balancing professional responsibilities with health is
             essential, especially given your commitment to improving healthcare
             access globally. Have you considered how a career change might not
            only address your health concerns but also align with your
            humanitarian goals and personal well-being?",
        "timestamp": "Dec 15, 2025, 06:11:23",
        "dialogue_turn": 0
    },
    ...
]
```

Figure 26: Example of a Single Dialogue Turn.

```
{
  "question": "What type of new physical activity might Martin be interested in
      trying after April 10, 2026?",
  "answer": "Other extreme sports.",
  "evidence": [
    {
      "memory_content": "Martin has developed a newfound appreciation for
          extreme sports...",
      "memory_type": "Persona Memory"
    }
  ],
  "difficulty": "medium",
  "question_type": "Generalization & Application"
}
```

Figure 27: Example of a Question.

```
[
    {
        "role": "user",
        "content": "Who is Inventor Thomas Edison?"
    },
    {
        "role": "assistant",
        "content": "Thomas Edison, born on February 11, 1847, in Milan, Ohio, was an American inventor and businessman who is often hailed as ..."
    },
    {
        "role": "user",
        "content": "What is the pH of the new, more concentrated bleach?"
    },
    {
        "role": "assistant",
        "content": "8.25% w/v is a measurement meaning 8.25 grams dissolved in every 100 ml aqueous solution of NaOCl, or 82.5 g/L to ..."
    },
    {
        "role": "user",
        "content": "What would be a clean way to solve this? Given a right triangle $ABC$ with legs $a$ and $b$, $\\angle C=90^\\circ$. Derive an equation of the set of points $M$ for which $$|MA|^2+|MB|^2=2|MC|^2.$$"
    },
    {
        "role": "assistant",
        "content": "Let's outline the key steps: Let C=(0,0), A=(a,0), B=(0,b) (a,b>0) so that ..."
    }
]
```

Figure 28: Examples of Irrelevant Dialogues.

F Additional Experimental Results

F.1 Efficiency Analysis of Memory Systems

Dataset	System	Dialogue Addition Time (min)	Memory Retrieval Time (min)	Total Time (min)
Medium	Mem0	2768.14	41.66	2809.8
	Mem0-Graph	2840.07	54.65	2894.72
	Memobase	293.30	139.95	433.25
	MemOS	1028.84	**20.52**	1049.37
	Supermemory	**273.21**	95.53	**368.74**
	Zep	-	53.34	-
Long	Mem0	691.62	39.15	730.77
	Mem0-Graph	870.32	62.42	932.74
	Memobase	**239.29**	136.19	**375.48**
	MemOS	1524.39	**20.96**	1545.34
	Supermemory	1672.53	137.02	1809.55
	Zep	-	50.22	-

Table 5: Time consumption of all memory systems during evaluation.

Table 5 shows the time consumption of all memory systems during the evaluation process for dialogue addition and memory retrieval, as well as their total runtime. Overall, dialogue addition requires substantially more time than memory retrieval, **indicating that the write stage is the primary computational bottleneck. Enhancing the efficiency of memory extraction and updating is thus crucial for improving interactive performance.** On HaluMem-Medium, Supermemory performs best in both dialogue addition and total runtime, while MemOS shows the best retrieval efficiency. However, the dialogue addition time of Mem0 and Mem0-Graph exceeds 2700 minutes, revealing their low processing efficiency during dialogue ingestion and memory construction. On HaluMem-Long, the dialogue addition time for Mem0, Mem0-Graph, and Memobase decreases, mainly because the number of processed memory points is reduced rather than due to performance improvement. In contrast, MemOS and Supermemory extract a substantially larger number of memory points, resulting in a significant increase in their time cost.