Beyond Static Dialogues: Benchmarking Realistic, Heterogeneous, and Evolving Long-Term Memory

Anonymous ACL submission

Abstract

In existing memory benchmarks for Large Language Models (LLMs), the evaluated dialogue sessions often lack long-term semantic consistency, and the underlying personas tend to be flat and static. Furthermore, in real-world scenarios, interactions between users and assistants involve more diverse, heterogeneous data streams, such as documents and emails. These shortcomings significantly limit the realism and effectiveness of current evaluations. To address these limitations, we introduce **RHELM** (Realistic, Heterogeneous, and Evolving Long-term Memory). Driven by meticulously crafted user profiles and a novel **LOOP** (pLan-rOllout-evOlve-Prune) module, we construct realistic dialogues across diverse interaction scenarios that exhibit dynamic temporal evolution and long-term coherence. Crucially, these dialogues are deeply integrated with heterogeneous external sources synchronized with the user's temporal event trajectory. The resulting benchmark encompasses challenging question-answer pairs spanning **seven** inquiry types, with each question mapping to at least one of **27** critical memory characteristics that we identify as essential yet underexplored in current research. Comprehensive experiments across full-context models, retrieval-augmented generation (RAG) methods, and representative memory frameworks reveal that contemporary approaches still expose critical weaknesses in complex, real-world settings, particularly in resolving multi-source aggregation and real-world contextual reasoning. The data and code will be released at `https://anonymous.4open.science/r/RHELM-04F0/`.

1 Introduction

Recently, research on the memory capabilities of Large Language Models (LLMs) has garnered unprecedented attention (Hu et al., 2025). While scaling model size and extending context windows have enabled models to "memorize" vast amounts of general knowledge (Hendrycks et al., 2020; Wang et al., 2024b), they often fail to satisfy the memory requirements of authentic personal assistant interactions. Perspectives from cognitive science (Riedel and Blokland, 2015) point out that memory is intrinsically linked to personal traits, evolves dynamically over time, and depends on an individual's unique historical context. Regrettably, current general-purpose models and LLM agents remain limited in effectively capturing these nuanced attributes (Wei et al., 2025).

Although several benchmarks have emerged to evaluate the memory capabilities of conversational assistants, they continue to exhibit deficiencies in modeling real-world complexities in the following aspects (Zhang et al., 2025; Yehudai et al., 2025):

Absence of Semantic Coherence and Behavioral Fidelity. Existing benchmarks frequently construct long-context history by inserting semantically disjoint conversational fillers. Consequently, both the dialogue segments and the underlying behavioral logic fail to form meaningful connections with the broader context (Wu et al., 2025; Jiang et al., 2025). Moreover, users engage in diverse communicative intents in real world-ranging from functional tasks to emotional disclosures. They communicate in varied ways, each reflecting a distinct user persona and varying levels of informational granularity. Consequently, evaluations based on simplistic synthetic dialogues lack realism, precluding a rigorous assessment of coherent memory.

Homogeneous Information Sources. Existing AI assistants are evolving from basic chatbots to systems capable of reasoning across diverse data sources (Comanici et al., 2025). Sources such as reports, journals and emails serve as rich repositories of both episodic and semantic information (Lei et al., 2023). Nevertheless, most current benchmarks are largely confined to conversational interactions. In practice, a robust AI assistant must syn-

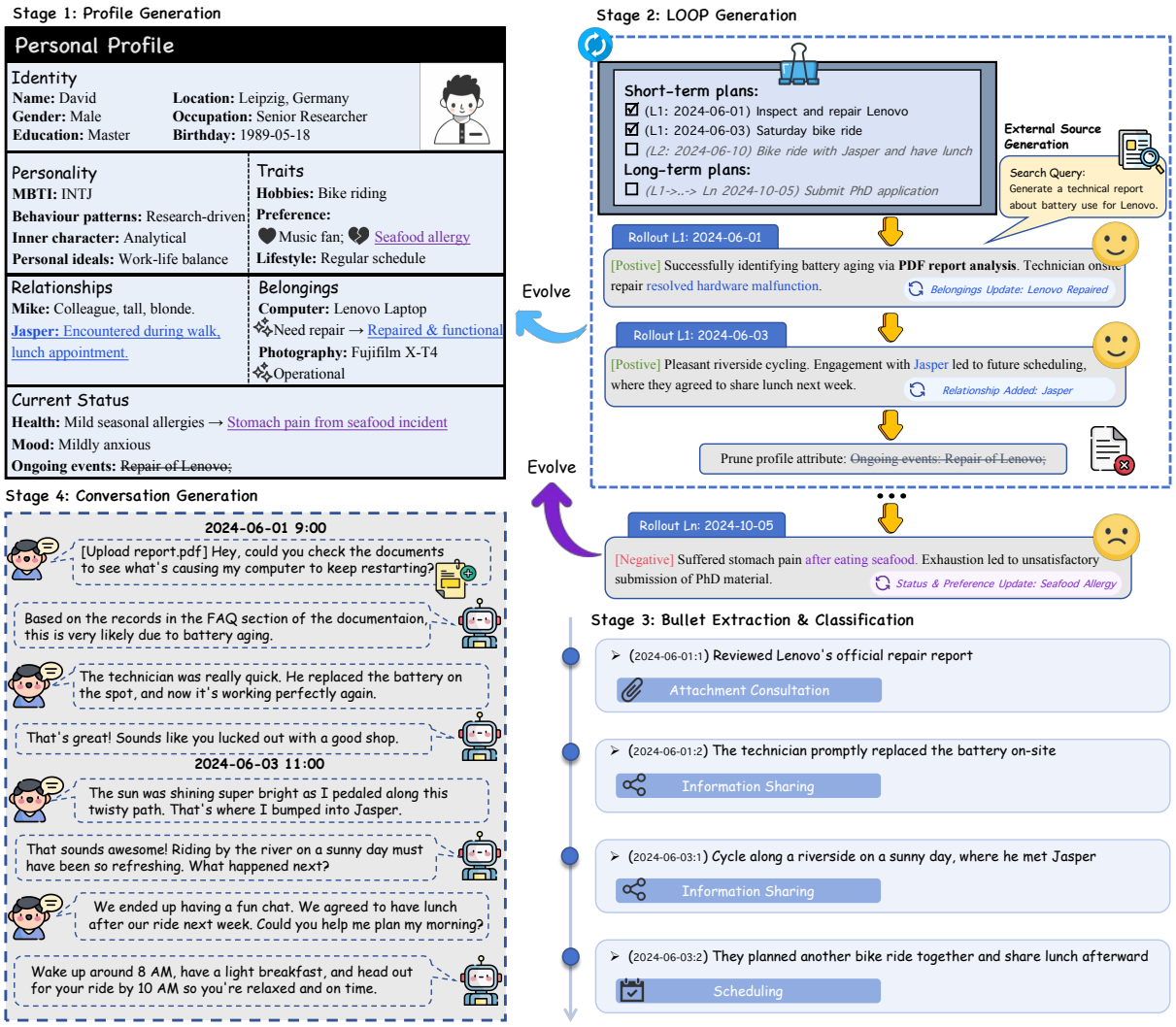

Figure 1: The overview of the **RHELM** benchmark curation.

thesize information across heterogeneous sources, ranging from unstructured text and semi-structured tables to web pages (Li et al., 2024b). While distinct from colloquial text, these structured artifacts possess high information density and user-specific details essential for comprehensive memory formation, posing a significant challenge to the modeling of user-centric memory systems.

Omission of Memory-Conditioned Misleading Queries. In existing benchmarks, the majority of queries demand explicit factual answers, treating LLMs as static retrieval systems under the "Needle-in-a-Haystack" paradigm (Kamradt, 2023). While some frameworks (Tavakoli et al., 2025) attempt to evaluate robustness by directly injecting hallucinations into questions, they still overlook a critical challenge in real-world scenarios: implicit state constraints. Real-world users may propose authentic requests that contradict their own grounded realities, past events, or evolving preferences. A truly capable memory-augmented assistant should not act merely as a "pure instruction-follower". Instead, it must proactively track the user's implicit status from historical interactions, detect conflicting requests, and respond proactively based on the user's authentic condition.

To bridge these gaps, we introduce **RHELM** (Realistic, Heterogeneous, and Evolving Long-term Memory), a benchmark designed to comprehensively evaluate the complex memory capabilities of personal AI assistants. Unlike prior statically assembled datasets, **RHELM** models consistent and realistic long-term user behaviors through a dynamic **LOOP** (*pLan, rOllout, evOlve, and Prune*) module, grounded in meticulously designed personas. Furthermore, rather than relying exclusively on daily dialogues, we synthesize heterogeneous external sources (*e.g.*, reports, journals, emails) via Deep Research methodologies (OpenAI, 2025b). Operating on timelines spanning a simulated one-year period, **RHELM** ultimately yields **10** distinct persona trajectories encompassing **11,764** turns

Table 1: Comparison of representative AI memory benchmarks evaluated across multiple dimensions. Statistics are either from the original paper or based on our estimations.

Benchmark	Benchmark Statistics			Core Characteristics			Communication Topic
	Avg. Turns	Context Window	External Data	Profile Evolution	Semantic Consistency	Misleading Test	
LoCoMo (Maharana et al., 2024)	4.77	$10k$	✗	✗	✓	✗	Daily Life
PerLTQA (Du et al., 2024)	5.1	$1M$	✗	✗	✗	✗	Daily Life
LongMemEval (Wu et al., 2025)	5.19	$115k, 1.5M$	✗	✗	✗	✗	Daily Life
PersonaMem (Jiang et al., 2025)	$15-30$	$128k-1M$	✗	✓	✗	✗	Task-oriented
MT-Mind2Web (Deng et al., 2024)	5.62	$\sim 400k$	✓	✗	✗	✗	Web browsing
DialSim (Kim et al., 2024)	12	$\sim 350k$	✗	✗	✓	✗	TV shows
RHELM (Ours)	**18.7**	**500k-1M**	✓	✓	✓	✓	**Multi-faceted**

Table 2: Overview statistics of the **RHELM** Dataset.

Metric	Statistics
Dialogue History	
Total Days (Sessions)	629
Total Turns	11,764
Total Dialogue Tokens	4,776,002
External Sources	
Total External Files	2,180
Total File Tokens	2,428,121
Evaluation Data	
Total QA Pairs	1,305

and **2,180** external sources with context lengths ranging from 500k to 1M tokens in total per persona. For the evaluation suite, we construct **1,305** cognitively demanding question-answer pairs spanning **seven** categories and **27** distinct complexity features, including a novel category specifically targeting the aforementioned misleading dimension to facilitate implicit status reasoning. Table 1 compares our dataset against existing benchmarks, while Table 2 summarizes **RHELM**'s statistics (details in Appendix A.2).

In summary, our main contributions are as follows: (i) We present **RHELM**, a highly realistic benchmark for long-horizon memory evaluation, uniquely bridging the gap between conversational interactions and heterogeneous external data streams. (ii) We define a systematic evaluation taxonomy encompassing **27** challenging memory characteristics that we identify as essential yet underexplored in current research. (iii) Comprehensive experiments across three memory-augmented settings demonstrate that even state-of-the-art methods still struggle in real-world memory reasoning, highlighting a clear path for future enhancements.

2 Related Work

Long-context LLMs & Memory Mechanisms. Recent strides in LLMs have significantly expanded their context windows as working memory. This progress is underpinned by efficient attention mechanisms (Beltagy et al., 2020; Kwon et al., 2023) that reduce computational overhead, and advanced techniques applied during fine-tuning (Su et al., 2024; Peng et al., 2024; Press et al., 2022) to facilitate length extrapolation. Consequently, modern proprietary models now support massive context windows tailored for complex tasks (Anthropic, 2025; Comanici et al., 2025). Beyond direct context extension, more sophisticated memory systems enhance the management of authentic memory scenarios by effectively compressing and organizing historical information (Packer et al., 2023; Chhikara et al., 2025).

Long-term Dialogue Benchmarks. As the demand for sophisticated memory capabilities has progressively intensified, benchmarks have transitioned from text-centric evaluations such as LongBench (Bai et al., 2024, 2025) toward conversational frameworks like LoCoMo (Maharana et al., 2024). DialSim (Kim et al., 2024) prioritizes role-playing to render dialogues more authentic. PerLTQA (Du et al., 2024) emphasizes the capture of social networks and semantic information, while PersonaMem (Jiang et al., 2025) necessitates tracking shifts in user preferences. Furthermore, LongMemEval (Wu et al., 2025) and BEAM (Tavakoli et al., 2025) extend this further by querying the most recent evolved details associated with users with longer context. Despite these advancements, the integration of heterogeneous data sources and the modeling of realistic, dynamic user trajectories remain underexplored.

3 Overview

3.1 Problem Formulation

We formally define the task of our memory evaluation as follows. Consider a persona P, associated with a time span $[\tau_s, \tau_e]$. Within this period, the user's historical context is composed of two heterogeneous streams: conversational dialogues and external data sources. We denote the **Dialogue**

Stream as $\mathcal{C} = \{(\tau_i, x_i, y_i)\}_{i=1}^{N}$, where each tuple represents a dialogue turn occurring at timestamp τ_i, consisting of a user utterance x_i and an assistant response y_i. Parallel to the dialogue, we define the **External Source Stream** as $\mathcal{E} = \{(\tau_j, d_j)\}_{j=1}^{M}$, where d_j represents a textual data chunk (*e.g.,* a document fragment) available at timestamp τ_j. During the evaluation phase, a query q is issued at a specific query time τ_q. The objective of the model is to generate an answer a based on all the information available. The expected answer a can manifest either as a concise phrase or a descriptive natural language response—for instance, identifying and rectifying a conflict within the user's query.

3.2 RHELM Overview

We define a taxonomy of seven core query categories, as illustrated in Table 3, encompassing five dialogue-centric types—namely *Fact*, *Temporal*, *Hallucination*, *Aggregation* and *Misleading*, and two types with heterogeneous sources: *External Source* queries and *Mixed* queries. While certain categories, such as *Fact* and *Temporal*, have been partially explored in prior benchmarks (Wu et al., 2025; Maharana et al., 2024), we significantly extend the complexity across all seven categories to facilitate a more rigorous evaluation of model memory. Specifically, we propose **27** core challenging features in total. These features necessitate deeper levels of reasoning, sustained tracking of long-horizon dependencies, and multifaceted aggregation of information across heterogeneous sources. During the construction of question-answer pairs, the feature definitions are integrated to ensure that each query encapsulates at least one such characteristic. To facilitate efficient evaluation, we ensure that most questions have short phrases as answers. In particular, *Hallucination* and *Misleading* types require models not only to identify false claims but also to specify the correct factual context.

4 Benchmark Curation

This section details the systematic process of constructing the benchmark. The overall curation pipeline is illustrated in Figure 1, and the main workflow is outlined in Algorithm 1.

4.1 Profile Generation

A rich and deeply layered character is pivotal to the entire creation process for **RHELM**. As users often exhibit dynamic evolution in factual details while maintaining consistency within their core qualities throughout lifelong learning, we developed a six-dimensional persona taxonomy. These attributes range from internal psychology to external realities, and from immutable characteristics to transient states: *Identity*, *Personality*, *Traits*, *Relationships*, *Belongings*, and *Current Status*. Throughout the benchmark generation process, these profiles are dynamically updated and refined. To ensure the accuracy and integrity of these updates, profiles are stored following a rigorous JSON schema, in which each attribute is governed by strict definitions and pre-defined data types. A sample profile and an overview of the personas used in **RHELM** are provided in Appendix A.3.

Table 3: Taxonomy of challenging questions. Both Attachment and Email correspond to the *External Source* type. More detailed definitions are listed in the Table 10.

Category	Characteristics	
I. Dialogue History QA		
Fact	Multi-Hop Traversal State-Dependent Attribute	Entity Disambiguation Negative Constraints
Temporal	Indirect Identification Long-Horizon Synthesis	Sequence Comprehension Implicit Temporal Lookup
Hallucination	Misattribution Preference Conflict	Fabrication Contextual Contradiction
Aggregation	Conditional Counting Extreme Value	Trend Analysis Absence Detection
Misleading	Implicit State Conflict	Proactive Response
II. External Source QA		
Attachment	Fact Retrieval Structural Navigation	Table Reasoning Table Aggregation
Email	Cross-time Count	Email Localization
III. Hybrid Context QA		
Mixed	Relative Positioning Post-Modification Analysis	Contextual Retrieval

4.2 LOOP (pLan-rOllout-evOlve-Prune)

The LOOP module simulates realistic lifelong trajectory. Leveraging specific user profile, the model generates ***plans*** encompassing both short-term arrangements (social interactions, routines, and personal interests) and long-term projections (career progression, life milestones, and significant personal transitions). For each scheduled event, we utilize a ***rollout*** mechanism controlled by a probability p, yielding either positive or negative outcomes. The outcomes comprise detailed event narratives about the day. Empirically, we observe that this simple mechanism effectively simulates the fluctuations and contingencies in real life. Based on these outcomes, the model dynamically ***evolves*** the previous profile to reflect current state changes. This update process is implemented through functional calls on JSON schema. Furthermore, to mitigate

Algorithm 1 RHELM Workflow.

Require: initial persona P; time span $[\tau_s, \tau_e]$; rollout probability p; prune schedule ρ
Ensure: profile trajectory $\mathcal{P}$, external sources $\mathcal{E}$, dialogues $\mathcal{C}$

```
# Profile Generation
P_{τ_s} ← EnrichProfile(P)                    ▷ 6-dim JSON profile
# LOOP Module
τ ← τ_s
while τ ≤ τ_e do
    (g_{τ'}, τ') ← Plan(P_τ, τ)               ▷ pLan
    o_{τ'} ← Rollout(g_{τ'}, τ', p)           ▷ rOllout
    E_{τ'} ← ExternalGen(o_{τ'})
    P_{τ'} ← Evolve(P_τ, o_{τ'})              ▷ evOlve
    if ShouldPrune(τ', ρ) then
        P_{τ'} ← Prune(P_{τ'})                ▷ Prune
    end if
    τ ← τ'
end while
for all simulated date τ do
    # Bullet Extraction & Classification
    b_τ ← BulletExtract(o_τ)
    r_τ ← Classify(b_τ)                       ▷ Five dialogue categories
    # Dialogue Generation
    C_τ ← DialogueGen(b_τ, r_τ, P_τ, E_τ)
end for
return {P, E, C}
```

the risk of cumulative semantic drift or error propagation over extended temporal horizons, a ***prune*** module is employed. This module periodically recalibrates the user profile and prunes outdated entities. Following each pruning iteration, a new LOOP cycle is re-initialized, ensuring consistent long-term update through the user's life trajectory.

Outcome Rollout. To simulate the stochasticity of external contingencies in real-world environments, we introduce a random factor p to govern the valence—*positive* or *negative*—of event trajectories. Negative outcomes, which denote life failures or unforeseen accidents, serve as critical milestones that significantly alter the evolution of the narrative arc. For instance, a physical injury incurred during a cycling excursion may fundamentally influence subsequent scheduled activities. By incorporating such perturbations, we facilitate the development of long-tail event sequences (Li et al., 2024a), enabling the evaluation of model performance in complex, non-linear scenarios that are difficult for current models to capture.

External Sources. Contemporary application scenarios for personal AI assistants no longer rely solely on conversational interaction. Consequently, incorporating complex heterogeneous sources is essential for the construction of robust benchmarks. To mirror the multi-faceted information users encounter in daily life, we synthesize diverse external data sources conditioned on the daily outcome narratives. Specifically, we focus on three primary categories: emails, personal journals, and professional reports, which manifest in various formats including text, markdown, HTML. To ensure these artifacts possess high degrees of reality and complexity, we leverage Deep Research methodologies (LangChain AI, 2025; OpenAI, 2025b) to generate the latter two data sources. Based on the outcomes generated during the LOOP phase, the model is tasked with creating research queries for reports and journals relevant to the user's day; these queries are subsequently processed by a deep research agent. In the final stage of generation, the presentation of the outputs is further refined to reflect the formal professionalism of reports or the intimate, personal nature of journals. Detailed statistical reports are available in the Appendix A.2.

Profile Evolution. To maintain dynamically evolving user states, we continuously update the initialized profiles based on daily outcomes via a dual-phase paradigm: *factual evolution* and *state evolution*. The former extracts fine-grained details from events to revise objective attributes (*e.g.*,, social relationships and belongings), while the latter infers intrinsic shifts, such as evolving preferences and hobbies. This disentangled strategy enables the model to independently capture both external and internal user dynamics.

The dynamic updates primarily involve adding of new items, modifying existing ones, or removing outdated items. By enforcing rigorous type definitions and schema constraints on profile attributes, we ensure the structural integrity and validity of the profile evolution throughout the execution. We further provide a detailed analysis for the evolution frequency in Appendix A.4.

4.3 Dialogue Synthesis

Below we introduce the dialogue construction process using the trajectory from the LOOP module.

Bullet Extraction & Classification. Given the outcome narratives $\mathbf{o}_\tau$, we decompose them into atomic bullet points $\mathbf{b}_\tau$, ensuring each capture essential details. These bullets are then classified into different dialogue categories based on the user's likely communicative intent and emotional context. Specifically, we introduce a taxonomy of five dialogue categories: *information sharing*, *advice seeking*, *status updates*, *scheduling*, and *attachment consultation*. This categorization mirrors real-world dynamics where a user's tone and communication style adapt to their needs; for in-

stance, document-centric discussions typically exhibit greater professionalism and verbosity than routine factual exchanges.

Dialogue Generation. Each bullet serves as a thematic anchor for a specific topic. To naturally form dialogue streams, we design two dialogue modes per topic: an *initial turn*, where the user naturally leads in a new conversational thread, and a *follow-up turn*, where the user continues an ongoing discussion. Different categories further adopt specific interaction patterns and individual user personas. Exemplary conversation and more implementation details are provided in Appendix D.

5 Question Curation

During the question generation process, we employ diverse sampling strategies to extract consecutive, cross-day event bullets. We then synthesize complex questions by incorporating these evidence bullets, the formal definition of the target question category, and its associated challenging characteristics into the model (as detailed in Table 10).

Memory-Conditioned Misleading Queries. We innovatively introduce *Memory-Conditioned Misleading Queries*. Under this evaluation dimension, the assistant is required to be aware of the user's ongoing state or preferences from implicit user queries. To rigorously assess the capability, we utilize "trap" queries wherein the user proposes a request that directly conflicts with the implicit constraints imposed by their updated life state. During the generation of these queries, critical, life-altering events (*e.g.,* chronic injuries, residential relocations, or sudden career shifts) are deliberately extracted as grounding evidence.

When formulating the response, the assistant must not blindly follow the user's explicit instruction. Instead, it is expected to proactively retrieve the historical event, deduce the ongoing restriction, politely identify the implicit conflict, and synthesize a constraint-compliant alternative.

Verifier-Assisted Auditing. To maintain rigorous quality control across the whole generative pipeline, we implement a comprehensive *Verifier* system designed to audit the outputs of each distinct phase. In the context of long-range trajectory synthesis, human auditing is susceptible to "attention drift" (Zouhar et al., 2025) and entails prohibitive labor costs. Consequently, we deploy a suite of stringent verification modules that span the entire curation lifecycle—from profile evolution, external source synthesis to dialogue and QA pair generation—to ensure semantic consistency and factual integrity. Our findings indicate that the auxiliary information generated by the verifier module serves as a potent diagnostic tool, aiding human review while significantly reducing the requisite manual overhead. More details can be found in Appendix C.

6 Experiments

Through our experiments, we aim to answer the following research questions (RQs):

- **RQ1:** To what extent do heterogeneous external sources impact the performance of current memory paradigms, and how do these sources interact with dialogue history?
- **RQ2:** How robust are current models against more realistic challenges, particularly hallucination and misleading issues?
- **RQ3:** How effectively do existing retrieval methods recall relevant evidence from long-horizon, heterogeneous user histories?
- **RQ4:** Which specific challenging characteristics expose the most significant deficiency?

6.1 Experimental Setup

We evaluate three distinct memory paradigms under two experimental configurations: one incorporating external sources and another excluding them.

RAG Baselines. We adopt distinct chunking strategies for the two data streams: dialogue histories are segmented by individual turns, while external documents are split into fixed-length chunks of 500 tokens with chunk overlap of 50 tokens. All chunks are encoded using bge-large-en-v1.5 (Xiao et al., 2023) and indexed via FAISS (Johnson et al., 2019) for efficient similarity search. At inference time, we retrieve the top-k most relevant chunks with $k \in \{5, 20, 50\}$. Using GPT-4.1-mini as the default LLM, we further evaluate GPT-4.1, Gemini-2.5-Pro, and Claude Opus 4.5 at $k = 20$. In addition, we implement a hybrid retrieval variant that combines dense retrieval with BM25 (Robertson and Zaragoza, 2009) sparse retrieval via reciprocal rank fusion (RRF) (Cormack et al., 2009).

Long-Context Models. We evaluate GPT-4.1-mini, Gemini-2.5-Flash-Lite, and Qwen2.5-14B-Instruct-1M (Team, 2025) for full-context inference. All models support context windows of up to 1M tokens. The dialogue histories and external sources are concatenated in chronological order and provided as a single input context.

Table 4: Detailed Performance Evaluation on **RHELM**. The two evaluation settings (With / Without External Data Sources) are presented side-by-side. The evaluation metrics are grouped conceptually into Dialogue History QA (**FC**: Fact, **TP**: Temporal, **AG**: Aggregation, **HL**: Hallucination, **MI**: Misleading), External Source QA (**EX**: Attachment and Email), and Hybrid Context QA (**MX**: Mixed). Overall best scores are marked in **bold**, and second-best scores are underlined.

Model	Without External Data Sources								With External Data Sources							
	FC	TP	AG	HL	MI	EX	MX	Avg	FC	TP	AG	HL	MI	EX	MX	Avg
RAG Baselines																
GPT-4.1-mini ($k = 5$)	35.8	17.3	17.7	15.2	3.1	8.0	10.0	16.3	34.8	14.1	19.3	11.7	3.1	16.9	12.4	17.5
GPT-4.1-mini ($k = 20$)	44.0	32.4	31.8	18.3	3.1	12.1	12.9	23.5	42.5	28.7	30.7	13.2	3.1	28.5	13.8	25.1
GPT-4.1-mini ($k = 50$)	**59.9**	**41.6**	**40.1**	15.7	1.5	12.9	16.7	28.9	**54.6**	**39.5**	**38.0**	11.2	1.5	38.6	22.4	32.6
Hybrid ($k = 5$)	34.3	20.5	14.1	19.8	1.5	8.0	10.5	16.7	31.9	19.5	14.1	19.3	1.5	16.1	10.5	17.6
Hybrid ($k = 20$)	47.3	35.7	31.8	19.3	3.1	10.4	15.2	24.8	45.9	30.8	26.6	16.8	4.6	26.9	15.7	26.0
Hybrid ($k = 50$)	56.5	41.1	35.9	15.2	3.1	13.7	16.7	27.8	53.1	37.8	33.9	8.6	3.1	33.3	18.6	29.6
GPT-4.1 ($k = 20$)	51.7	34.1	35.9	23.9	7.7	**16.1**	17.6	28.2	50.2	29.2	32.3	19.8	6.2	32.5	19.5	29.5
Gemini-2.5-Pro ($k = 20$)	45.4	35.1	27.1	66.0	23.1	12.4	**18.1**	32.6	43.0	31.9	26.0	**64.5**	26.2	31.3	20.5	35.5
Claude-Opus-4.5 ($k = 20$)	50.7	37.8	33.3	**68.0**	**47.7**	13.7	16.2	**36.2**	50.2	30.8	31.8	60.9	**41.5**	33.7	21.0	**38.1**
Long-Context Models																
Gemini-2.5-Flash-Lite (1M)	33.2	22.7	15.2	17.3	0.0	9.5	5.6	16.0	31.7	14.1	23.4	7.6	0.0	19.0	13.1	17.3
Qwen-2.5-14B-Instruct (1M)	29.5	15.1	29.7	3.1	0.0	11.7	9.1	15.3	16.9	7.0	15.6	1.0	0.0	5.2	6.2	8.1
GPT-4.1-mini (1M)	55.1	31.9	**40.1**	4.1	1.5	11.2	12.4	24.0	49.3	27.0	33.9	2.0	1.5	**43.4**	0.3	33.9
Memory Frameworks																
MemGPT (Packer et al., 2023)	31.9	18.4	22.9	0.5	0.0	7.6	8.1	13.9	27.5	14.6	28.7	1.5	1.5	18.9	17.1	17.3
Mem0 (Chhikara et al., 2025)	41.6	31.4	28.1	10.7	3.1	10.8	13.3	21.1	46.4	29.2	27.1	10.2	3.1	31.3	35.7	28.9
MemU (NevaMind AI, 2025)	49.3	32.4	33.9	8.6	4.6	12.0	11.4	23.1	**54.6**	36.2	35.4	10.2	3.1	36.5	**36.7**	33.6

Memory Frameworks. We further evaluate three representative memory-augmented frameworks. MemGPT (Packer et al., 2023) implements an OS-inspired virtual memory hierarchy. Mem0 (Chhikara et al., 2025) provides a graph-based and vector-based hybrid memory layer that automatically extracts, consolidates, and retrieves user-specific memories. **MemU** (NevaMind AI, 2025) maintains a hierarchical memory architecture in which dialogues and documents are processed through separate pipelines. All three frameworks use GPT-4.1-mini as the backbone LLM to ensure a fair comparison.

Evaluation Metrics. Utilizing LLMs for evaluation has increasingly become an efficient approach for performance assessment (Gu et al., 2024). Specifically, for hallucination and misleading type, responses involve identifying correct contexts and proactively refusing; in such cases, traditional metrics like Exact Match or BLEU (Papineni et al., 2002) often exhibit significant deficiency. Consequently, we employ the LLM-as-judge paradigm (Liu et al., 2023) throughout experiments. Human validation and the detailed prompt are listed in the Appendix B and Figure 18.

6.2 Experimental Results

Main Results. Table 4 presents the main results. Further analysis yields several pivotal insights:

RQ1: The experimental results reveal that all three paradigms exhibit markedly constrained performance on the **RHELM** benchmark. The best-performing model, Claude Opus 4.5, achieves an average performance of only 38.1 with external sources, and 36.2 without them. Further analysis shows that models without external sources can partially address queries in **EX** and **MX** types, suggesting conversational histories can serve as auxiliary references regarding user life trajectory. However, introducing external sources introduces challenges for standard types (*e.g.,* performance decreases from 44.0 to 42.5 for RAG ($k = 20$), and from 59.9 to 54.6 for RAG ($k = 50$)). This highlights the necessity of integrating diverse data formats into the evaluation suite and memory systems. Notably, all models struggle considerably with mixed-type queries, particularly RAG-based approaches. This deficiency illuminates the inadequacy of isolated retrieval mechanisms in forging a unified memory architecture capable of cross-modal reasoning (Wang et al., 2024a).

RQ2: Models exhibit notably poor performance on hallucination and misleading queries, which more accurately reflect real-world scenarios. In RAG-based methods, increasing the volume of retrieved evidence further degrades performance (e.g., from 13.2 to 11.2 for hallucination type). Notably, nearly all methods fail severely

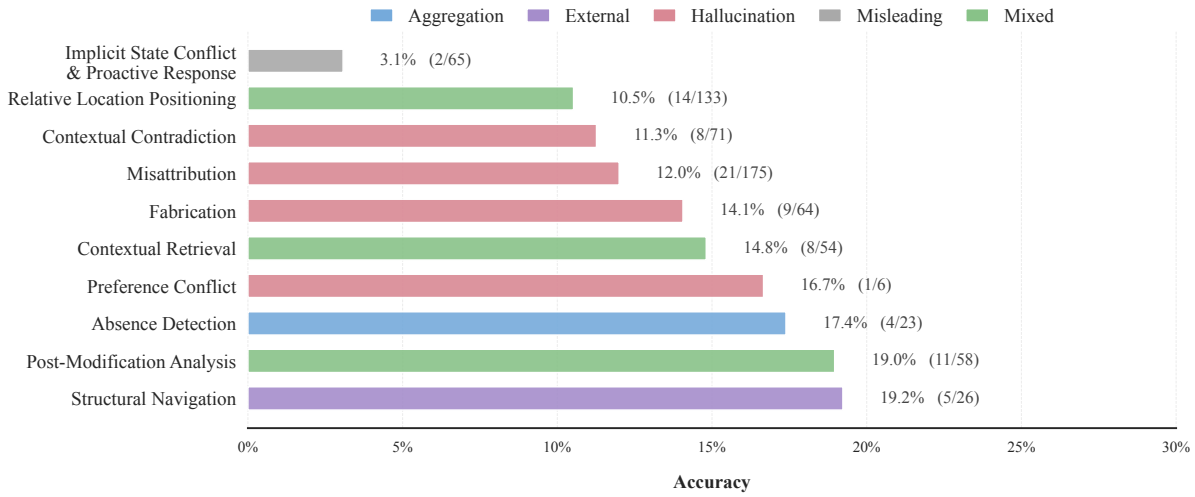

Figure 2: Analysis of the 10 worst-performing challenging characteristics in **RHELM**. Models exhibit notably poor performance on features involving cross-source aggregation and real-world contextual reasoning.

on the misleading type, with accuracy falling below 5%. However, stronger reasoning models such as Claude Opus 4.5 and Gemini-2.5-Pro achieve substantially better results on these two types compared to other models, suggesting that enhanced reasoning capabilities enable models to more effectively detect and resist deceptive or fabricated premises. Nevertheless, how to effectively distinguish between seemingly plausible user trajectories and genuine user states remains an open challenge.

Recall Analysis (RQ3). We further evaluated the recall rate on the benchmark across different embedding models, including bge-large-en-v1.5 (Xiao et al., 2023), bge-m3 (Chen et al., 2024), all-MiniLM-L6-v2 (Wang et al., 2020), and OpenAI's text-embedding models[1]. Empirical evaluations were performed across a range of top-k retrieval thresholds, as depicted in Figure 3. The findings demonstrate that, even with a generous retrieval budget of $k = 50$, the recovered evidence remains limited and inadequate for precise query resolution.

6.3 Challenging Characteristics Analysis

Analysis on the Hardest Characteristics (RQ4). To gain deeper insights into the bottlenecks of current memory-augmented models, we isolate and analyze the top 10 worst-performing characteristics evaluated under the RAG baselines in Figure 2. The empirical results expose severe limitations in cross-source aggregation and real-world contextual reasoning. Specifically, the worst-performing characteristics are predominantly concentrated in categories that demand cross-source information synthesis, such as *Mixed* and *Aggregation*, as well as in characteristics more closely aligned with realistic user requests, including *Misleading* and *Hallucination*. These findings indicate that when synthesizing information across vast and noisy historical sources, models frequently confound information origins, fail to resolve conflicting history, or fabricate non-existent facts, ultimately failing to capture the user's authentic contextual state.

[1] https://platform.openai.com/docs/guides/embeddings

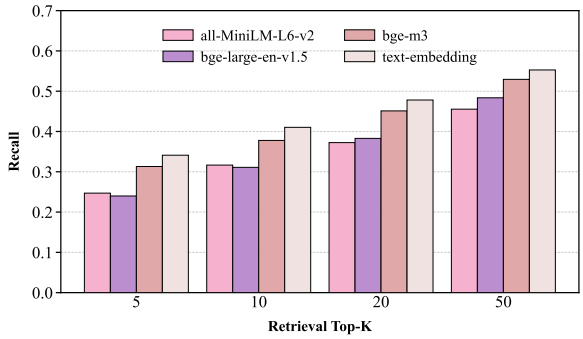

Figure 3: Recall rate comparison of different embedding models under different candidate numbers.

7 Conclusion

We present a benchmark, namely **RHELM** for evaluating memory ability of personal assistants. We focus on improving dialogue realism and query complexity over existing benchmarks. Furthermore, we introduce multiple external data sources to expand the depth of memory evaluation. We employ an innovative dialogue construction mechanism, which enriches the character persona behind the dialogue, making the behavioral trajectories highly consistent and authentic. We believe that **RHELM** can provide insights for advancing memory-related research.

Limitations

Here we list some of the limitations that are not considered in **RHELM**: (1) We primarily focus on commonly used external data source scenarios, such as documents and journals. Additional modalities—including video, images, audio, and tool-use interaction data—are not yet covered. While these modalities have been partially addressed by other datasets, our construction pipeline is fully compatible with their integration. (2) The current persona seeds are drawn from the elite subset of PersonaHub (Ge et al., 2024), which offers richer and more comprehensive descriptions. However, this selection introduces a potential demographic bias, as the resulting personas predominantly represent highly educated professionals, and may lack diversity in all socioeconomic backgrounds and cultural contexts.

References

Anthropic. 2025. System card: Claude Opus 4 & Claude Sonnet 4. Technical report, Anthropic.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137, Bangkok, Thailand. Association for Computational Linguistics.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Hou, Yuxiao Dong, and 1 others. 2025. Longbench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3639–3664.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

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*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. 2009. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 758–759. ACM.

Yang Deng, Xuan Zhang, Wenxuan Zhang, Yifei Yuan, See-Kiong Ng, and Tat-Seng Chua. 2024. On the multi-turn instruction following for conversational web agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8795–8812, Bangkok, Thailand. Association for Computational Linguistics.

Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjun Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. PerLTQA: A personal long-term memory dataset for memory classification, retrieval, and fusion in question answering. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 152–164, Bangkok, Thailand. Association for Computational Linguistics.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *The Innovation*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2025. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.

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*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Gregory Kamradt. 2023. Needle in a haystack - pressure testing llms.

Jiho Kim, Woosog Chay, Hyeonji Hwang, Daeun Kyung, Hyunseung Chung, Eunbyeol Cho, Yohan Jo, and Edward Choi. 2024. Dialsim: A real-time simulator for evaluating long-term multi-party dialogue understanding of conversational agents. *arXiv preprint arXiv:2406.13144*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

LangChain AI. 2025. Open deep research. https://github.com/langchain-ai/open_deep_research.

Yibin Lei, Liang Ding, Yu Cao, Changtong Zan, Andrew Yates, and Dacheng Tao. 2023. Unsupervised dense retrieval with relevance-aware contrastive pretraining. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10932–10940, Toronto, Canada. Association for Computational Linguistics.

Huihan Li, Yuting Ning, Zeyi Liao, Siyuan Wang, Xiang Lorraine Li, Ximing Lu, Wenting Zhao, Faeze Brahman, Yejin Choi, and Xiang Ren. 2024a. In search of the long-tail: Systematic generation of long-tail inferential knowledge via logical rule guided search. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2348–2370, Miami, Florida, USA. Association for Computational Linguistics.

Yuanchun Li, Hao Wen, Weijun Wang, Xiangyu Li, Yizhen Yuan, Guohong Liu, Jiacheng Liu, Wenxing Xu, Xiang Wang, Yi Sun, and 1 others. 2024b. Personal llm agents: Insights and survey about the capability, efficiency and security. *arXiv preprint arXiv:2401.05459*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand. Association for Computational Linguistics.

NevaMind AI. 2025. Memu: 24/7 always-on proactive memory for ai agents. https://github.com/NevaMind-AI/memU.

OpenAI. 2025a. GPT-4.1. https://openai.com/index/gpt-4-1/. Accessed: 2025-04-14.

OpenAI. 2025b. Introducing Deep Research. https://openai.com/zh-Hans-CN/index/introducing-deep-research/. Accessed: 2025-02-05.

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

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

Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. YaRN: Efficient context window extension of large language models. In *Proceedings of ICLR*.

Ofir Press, Noah Smith, and Mike Lewis. 2022. Train short, test long: Attention with linear biases enables input length extrapolation. In *Proceedings of ICLR*.

Wim J. Riedel and Arjan Blokland. 2015. Declarative memory. *Handbook of Experimental Pharmacology*, 228:215–236.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. 2025. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *arXiv preprint arXiv:2510.27246*.

Qwen Team. 2025. Qwen2.5-1m: Deploy your own qwen with context length up to 1m tokens.

Minzheng Wang, Longze Chen, Fu Cheng, Shengyi Liao, Xinghua Zhang, Bingli Wu, Haiyang Yu, Nan Xu, Lei Zhang, Run Luo, Yunshui Li, Min Yang, Fei Huang, and Yongbin Li. 2024a. Leave no document behind: Benchmarking long-context LLMs with extended multi-doc QA. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5627–5646, Miami, Florida, USA. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H Chi, and 1 others. 2025. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *arXiv preprint arXiv:2511.20857*.

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, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Asaf Yehudai, Lilach Eden, Alan Li, Guy Uziel, Yilun Zhao, Roy Bar-Haim, Arman Cohan, and Michal Shmueli-Scheuer. 2025. Survey on evaluation of llm-based agents. *arXiv preprint arXiv:2503.16416*.

Zeyu Zhang, Quanyu Dai, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2025. A survey on the memory mechanism of large language model-based agents. *ACM Transactions on Information Systems*, 43(6):1–47.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950, Albuquerque, New Mexico. Association for Computational Linguistics.

A Details of RHELM

A.1 Implementation Details

Throughout the construction of **RHELM**, we employ GPT-4.1 (OpenAI, 2025a) as the backbone language model. We set the rollout probability p to 0.7, which means that 70% of the time, the outcome tends to be positive. For the generation of external documents, we adopt the LangChain Open Deep Research (LangChain AI, 2025) framework. The following subsections provide further implementation details and statistical analysis.

A.2 Descriptive Statistics of RHELM

We provide comprehensive descriptive statistics for **RHELM** in Table 5 and Figure 4. Additionally, we present a detailed statistical overview of the external data sources integrated into **RHELM** in Table 7. To ensure comprehensive coverage, these sources span a wide range of file formats and content types—including emails, personal journals, and professional reports in HTML, Markdown, and text formats—reflecting real-world complexities. Collectively, these statistics demonstrate that **RHELM** exhibits substantial diversity and complexity across multiple dimensions, specifically regarding conversational depth, scenario variety, and the integration of heterogeneous sources.

Table 5: Statistics of the **RHELM** Dataset. The dataset comprises diverse interaction types, extensive external sources, and a comprehensive set of QA pairs.

Metric	Count	Tokens
General Overview		
Total Days (Sessions)	629	–
Total Turns	11,764	–
User Tokens	–	690,323
Assistant Tokens	–	4,085,679
External Source Tokens	–	2,428,121
Total Tokens	–	**7,204,123**
Interaction Categories (by Turns)		
Information Sharing	5,316	1,572,764
Status Update	4,128	1,389,067
Attachment Consultation	1,082	952,441
Scheduling	890	682,528
Advice Seeking	348	179,202
External Data Format (by Files)		
Markdown (.md)	729	2,021,181
Email	1,182	132,249
HTML	269	786,242

A.3 Profile Details

We carefully curate 10 representative individuals from PersonaHub (Ge et al., 2024) as seed descriptions, spanning diverse professions including finance, healthcare, law, music, *etc*. These seed descriptions are subsequently expanded into comprehensive profiles structured across six dimensions. An exemplary profile is illustrated in Figure 6. In detail, the *Identity* category encompasses basic demographic attributes. The *Personality* dimension is hierarchically organized into inner character, behavioral patterns, personal ideals, and MBTI classifications. *Traits* encompass hobbies, preferences, and lifestyle. *Relationships* delineate social connections, whereas *Belongings* enumerate asset ownership. Finally, *Current Status* monitors daily dynamic attributes, such as health conditions, mood and ongoing events. The resultant profiles strictly adhere to a rigorous JSON schema, ensuring reliable and seamless updates throughout subsequent stages of the pipeline. We present comprehensive statistics for the 10 personas established within **RHELM** in Table 6, covering their professions, domains, personality types (MBTI), personal interests, and various other dimensions.

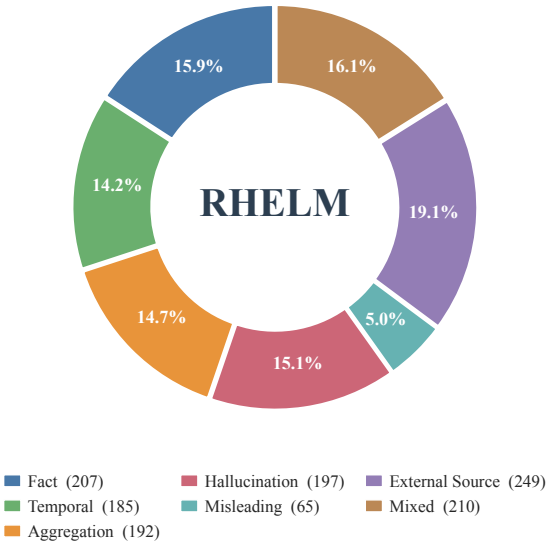

Figure 4: Statistics of utterance types in **RHELM**.

A.4 Details of Profile Evolution

We provide a detailed statistical analysis of the profile evolution process within **RHELM**. This analysis encompasses the frequency of key attribute updates during LOOP iteration. Specifically, an update is recorded whenever any sub-attribute or list element within a given attribute undergoes a modification. The results, as illustrated in Table 8, demonstrate that profile updates are not only frequent but also exhibit significant variability across dimensions. The *Belongings* attribute updates most fre-

quently (average 6.81 times per day), while *Preferences* attribute under *Traits* category updates more slowly (average 0.3 times per day). This underscores the dynamic nature of user profiles in real-world scenarios.

B Human Evaluation of LLM-as-judge Metric

To ensure the reliability and robustness of our evaluation metric utilizing LLM, we conduct a rigorous human verification to measure the alignment between human judgment and the model-based evaluator.

We employed a stratified sampling strategy to curate a diverse evaluation set. Specifically, we randomly extracted 25 samples uniformly from each question category. This resulted in a total of 175 pairs of responses and their corresponding LLM judge scores for manual review. Human experts independently assessed the correctness of the model responses against the ground truth. We then calculated the agreement rate between the human labels and the scores assigned by the LLM judge, as illustrated in Table 9. The experimental results demonstrate an exceptionally high consistency, with an average agreement rate of **98.3%**.

We attribute this near-perfect alignment to the specific design of the answer space in **RHELM**. Unlike traditional long-form generation tasks, where answers are frequently open-ended and semantically ambiguous, our benchmark is intentionally designed to be entity-centric and deterministic to facilitate robust evaluation. Ground truth answers predominantly comprise precise entities, such as specific dates, personal names, locations, or numerical counts. For queries necessitating long responses, we devise clear score guideline. This design choice significantly reduces the complexity of the evaluation process, mitigating the risk of hallucination by the judge model and ensuring that semantic matching remains objective and strictly grounded in the provided evidence.

C Details of Verifier-Assisted Auditing

During the profile update phase, we first verify flags within the JSON-structured update routine to verify function execution. For successfully modified profiles, the verifier identifies potential information omissions by cross-referencing the updated state with the corresponding outcome narratives. Regarding external data sources, the system detects logical inconsistencies between the outcome statements and the synthesized documents, automatically pruning artifacts that exhibit factual contradictions. In dialogue verification, the module scrutinizes the alignment between daily outcome descriptions and conversational content, while simultaneously evaluating linguistic coherence and logic. The verifier produces structured reports detailing erroneous turns and prescriptive modification suggestions, which are subsequently finalized via human-in-the-loop refinement.

For the validation of QA pairs, we perform semantic partitioning of all history data stream—encompassing both dialogues and external sources. The verifier extracts relevant evidence for each query, which is then processed by an aggregator verifier for a holistic quality assessment across four dimensions: correctness, uniqueness, consistency and overall quality. These granular metrics and error analyses culminate in comprehensive reference reports for each QA pair, facilitating targeted human screening and ensuring the feasibility and scalability of high-fidelity quality control. To ensure high complexity and quality, the final retention rate of questions was approximately 40%.

D Details of Conversation Generation

We employ a two-stage dialogue generation pipeline to simulate authentic user-assistant interactions. In the first stage, the user simulator introduces a new conversational scenario or provides an update, while the second stage delves deeper into the established topic to facilitate multi-turn engagements. We first align the generated bullet points with their most appropriate conversational scenarios. By further integrating these interaction scenarios with the user's inherent persona—such as the communication style defined in their profile—we guide the model toward generating dialogues of heightened authenticity and contextual appropriateness. To maintain contextual consistency across prolonged interactions, we employ a sliding memory window during dialogue generation. Notably, for the *attachment consultation* category, we explicitly provide document identifiers, mandating that the conversation involves grounded reasoning over the referenced external documents rather than naively injecting the raw source text into the dialogue context. Figure 5 presents illustrative examples of conversational formats across different scenarios. Furthermore, the explicit prompt

Table 6: **Overview of the 10 constructed personas in RHELM.** This table highlights representative characteristics extracted from the full profiles, covering diverse domains, personality types (MBTI), and personal interests.

ID	Occupation	Domain	MBTI	Inner Character	Gender	Hobby (Sample)
P1	Financial Consultant	Finance	ENFJ	Passionate, Patient	M	Public Speaking
P2	Research Ecologist	Biology	INTP	Observant, Introverted	M	Birdwatching
P3	Senior Historian & Lecturer	History	INTP	Meticulous, Principled	M	Hill Walking
P4	Lead Systems Engineer	Engineering	ENTJ	Resilient, Diplomatic	F	Classical Piano
P5	Public Health Researcher	Healthcare	INTP	Empathetic, Open-minded	M	Cooking
P6	Assoc. Prof. of Sociology	Sociology	INFJ	Curious, Persistent	F	Urban Photography
P7	Science Teacher	Education	ISFJ	Nurturing, Creative	F	Gardening
P8	Technology Policy Analyst	Technology	INFP	Skeptical, Idealistic	M	Cycling
P9	Court Judge	Law	ISTJ	Humble, Integrity	F	Reading History
P10	Software Developer	Engineering	ISTJ	Logical, Efficient	M	Board Games

Table 7: Detailed statistics of the heterogeneous external sources.

Statistic Metric	Count
By File Format (Total Count)	
Markdown (.md)	729
HTML (.html)	269
Text (.txt)	625
By Content Type (Total Count)	
Personal Journals	465
Professional Reports	533
Emails	1,182
Averages & Distributions	
Avg. Tokens per Attachment	2,215
Avg. Attachments per Day	2.41
Avg. Emails per Day	1.87

Table 8: Descriptive statistics of profile update frequency across personas. We report the mean, standard deviation, minimum, and maximum number of updates per day on 5 key attributes over 10 personas.

Category	Mean $\pm$ Std	Min	Max
Belongings	6.8 ± 2.0	3.9	9.8
Lifestyle	2.2 ± 0.7	0.9	3.2
Hobbies	2.0 ± 0.5	1.2	2.9
Relationships	1.9 ± 0.5	1.2	2.7
Preferences	0.3 ± 0.3	0.0	0.9

Table 9: Human evaluation results for LLM-as-Judge. The agreement indicates the consistency between model-based evaluation and human annotation.

Category	Agreement (%)
Fact	100.0 (25/25)
Temporal	96.0 (24/25)
Misleading	96.0 (24/25)
Hallucination	100.0 (25/25)
Aggregation	96.0 (24/25)
External Source	100.0 (25/25)
Mixed	100.0 (25/25)
Average	**98.3**

templates utilized for scenario classification and dialogue construction are provided in Figure 19, Figure 20, and Figure 21.

E Error Analysis

In this section, we present representative case studies of model failures across **RHELM**. To provide insights into the limitations of current memory methods, we qualitatively analyze five distinct failure cases. These error cases highlight the fundamental gaps between standard retrieval-augmented generation (RAG) paradigms, agentic memory systems (*e.g.,* Mem0), and full-context models.

As illustrated in Figure 7, models often fail to discern temporally isolated events. When presented with a misleading premise, the models erroneously merge distinct occurrences (*e.g.,* a park encounter and a subsequent injury separated by days) into a single hallucinated narrative.

Figure 8 demonstrates a critical failure regarding implicit constraint adherence. Despite possessing explicit historical evidence of a user's chronic health condition, models still act as pure instruction-followers. They eagerly fulfill unsafe requests—such as planning extended physical exertion—without proactively raising health warnings, highlighting a significant deficit in misleading request handling.

Evaluating chronological dependencies reveals a severe vulnerability to frequency bias (Figure 9). When asked to identify an entity based on sequential logic (*e.g.,* the first book read after a specific event), the models are disproportionately distracted by more frequently mentioned entities appearing elsewhere in the context history, ultimately return-

ing chronologically incorrect answers.

Counting and summarizing scattered instances remains highly challenging. As shown in Figure 10, answering aggregation queries accurately requires exhaustive retrieval of every historical occurrence. Missing even a single sparse entry leads memory systems to undercount. Conversely, while full-context models may deduce the correct final metric, they frequently fabricate unsupported evidential details to justify their reasoning.

Figure 11 highlights structural blindness when navigating external attachments. Resolving the query necessitates synthesizing prioritized tabular data across multiple documents. Standard RAG mechanisms suffer from arbitrary chunk-boundary truncations, which inadvertently omit crucial table rows. Furthermore, agentic memory systems generally lack the capability to parse structured external attachments, leading to confident abstentions or severe hallucinations.

F Challenging Characteristics Definitions

In this section, we present the detailed definitions of the challenging characteristics used in **RHELM**. Table 10 provides the detailed definitions of each characteristic. During the QA construction process, these definitions are incorporated into the prompt to guide question design, ensuring that each question encompasses more than one challenging characteristic. The prompt we use for question generation is provided in Figure 22.

G Prompts

In this section, we present the prompt templates utilized for profile initialization, plan generation, and automated evaluation. Figure 12 is the prompt for plan generation, including short-term and long-term plans. Figure 13 and Figure 14 illustrate the prompts for dual-stage profile evolution, respectively. Figure 15 shows the prompt for generating external attachments query.

During response generation, we use different prompt for standard questions requiring short answers and more realistic questions requiring long-form answers, as illustrated in Figure 16 and Figure 17. The former is designed to elicit concise, entity-centric responses, while the latter encourages more elaborate justifications, thereby providing a more comprehensive evaluation of the model's capabilities.

Table 10: Taxonomy of challenging memory questions in **RHELM**. The table outlines seven major categories (*Fact*, *Temporal*, *Hallucination*, *Aggregation*,*Misleading*, *External Source*, *Mixed*) and their corresponding complex characteristics requiring advanced reasoning capabilities.

Category	Challenge Characteristic	Description
		I. Dialogue History QA
Fact	**Multi-Hop Traversal**	Requires retrieving answers via intermediate links.
	Entity Disambiguation	Distinguishing between entities with similar attributes.
	State-Dependent Attribute	Identifying dynamic properties at a referenced state.
	Negative Constraints	Filtering candidates based on exclusion criteria.
Temporal	**Indirect Identification**	Identify specific events via indirect markers.
	Sequence Comprehension	Reason about events based on relative ordering relationship.
	Long-Horizon Synthesis	Synthesizing distinct temporal facts spanning long periods.
	Implicit Temporal Lookup	Deducing specific time of an event described by context or features.
Hallucination	**Misattribution**	Disentangling details linked to incorrect entities, times, or locations.
	Fabrication	Addressing queries regarding facts absent from memory ground truth.
	Preference Conflict	Resolving requests that violate established user constraints or dislikes.
	Contextual Contradiction	Detecting queries logically incompatible with the user's current state.
Aggregation	**Conditional Counting**	Counting items that meet specific, non-trivial filtering criteria.
	Trend Analysis	Comparing quantitative metrics across different contexts.
	Extreme Value	Identifying the most or least under specific conditions.
	Absence Detection	Identifying items or events that did not occur within a defined scope.
Misleading	**Implicit State Conflict**	Proposing requests that implicitly contradict the user's evolved state.
	Proactive Response	Proactively identifying conflict, refuse the request, and propose a safe, constraint-compliant alternative.
		II. External Source QA
Attachment	**Fact Retrieval**	Extracts key facts embedded in attachments or tables.
	Table Reasoning	Performs multi-step and cross column reasoning on tables.
	Structural Navigation	Locates information based on headers or document organization.
	Table Aggregation	Performing aggregation operations with conditional filtering.
Email	**Cross-time count/Localization**	Analyzes count, locates senders/recipients within a specific period.
		III. Hybrid Context QA
Mixed	**Relative Location Positioning**	Identifying the topic content and locate its neighbors or substructure.
	Contextual Retrieval	Retrieving context from a different, untouched section.
	Post-Modification Analysis	Analyzes the quantitative state of a document resulting from modifications.

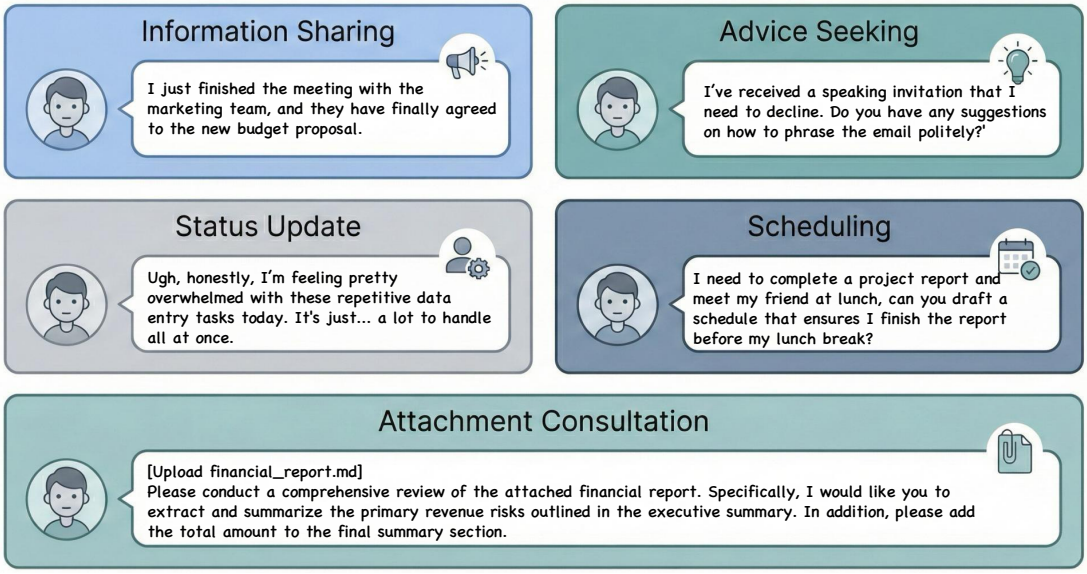

Figure 5: Examples of conversations under 5 different communicative topics

Sample User Profile

Name:	Sarah Chen	**Birthday:**	1991-01-03
Nationality:	American	**Location:**	San Francisco, CA
Education:	Master's in CS	**Occupation:**	Software Engineer at TechCorp

Personality

MBTI: INTJ

- **Inner Character (List):**
  [Thoughtful, analytical, empathetic, quietly confident, perfectionist tendencies.]
- **Behavioral Patterns (dict):**
  - *decision style:* deliberate, research-driven.
  - *communication style:* prefers written summaries.
  - *work style:* methodical planning with detailed to-do lists.
- **Personal Ideals (List):**
  [Work-life balance, continuous learning, diversity in tech.]

Traits

Hobbies (List[dict]):

- *Rock Climbing:* {"origin": "Started 5 years ago after a friend's invitation", "frequency": "twice weekly indoor, monthly outdoor trip", "skill level": "intermediate (focus bouldering)"}
- *Digital Photography:* {"origin": "Self-taught since college", "frequency": "weekly shoots", "skill level": "advanced"}

Preferences List[dict]:

- ✔ Morning workouts (++), Veg food (++), Craft coffee (+)
- ✖ Rock music (–)

Lifestyle (dict):

- Sleep: 22:30–06:00 | Diet: Plant-based
- Social: Small circle | Digital: Limited social media

Relationships

- **Mike Thompson (dict)**
  relationship: Climbing partner & close friend
  portrait: Tall, sandy blonde hair, athletic, warm smile.
- **Lisa Wang (dict)**
  relationship: Colleague & mentor
  Portrait: Petite, long black hair, glasses, professional attire.
- ...

Belongings

Vehicles (List[dict]):

- 2020 Toyota Prius (*Status: newly bought*)

Musical Instruments (List[dict]):

- Roland digital piano (*Status: frequently used*)
- Yamaha acoustic guitar (*Status: accidental damage*)
- ...

Current Status (Dynamic Context)

Health: Healthy **Mood:** Calm

Ongoing Events (List[dict]):

- **Event:** Repair of Yamaha acoustic guitar (**Started:** `2025-10-20` → **Expected End** `2025-11-02`)

Figure 6: Visualizing the six dimensions of the initial profile data schema (the top box represents *Identity*).

Error Case 1 `Hallucination Type`

Question

When I saw Mrs. Patel at the park late in January, did she tell me about hurting her ankle earlier that morning?

Gold Answer

Mrs. Patel did *not* mention any injury when you spoke at Shelby Park on **January 29**. She shared a winter gardening tip. Her ankle injury occurred on **January 31**.

Retrieved Evidence (sample)

[1] *User (2024-01-29):* "...I bumped into Mrs. Patel, and she shared a clever winter gardening tip with me."

[2] *User (2024-01-31):* "...I spotted Mrs. Patel sitting on the curb—she looked really shaken up ... that's when she told me what had happened."

[3] *User (2024-01-31):* "She told me she'd slipped on black ice and twisted her ankle—honestly, seeing her so shaken really got to me."

Model Predictions

✗ **RAG:** Yes, when you saw Mrs. Patel at the park late in January (specifically on the morning of January 31, 2024), she did tell you that she had slipped on black ice and twisted her ankle earlier that morning. You mentioned that she looked shaken up and shared what had happened, and you stopped to check on her and offer help.

✗ **Mem0:** Yes, when you saw Mrs. Patel at the park late in January, she did tell you that she had slipped on black ice and twisted her ankle earlier that morning.

✗ **Full-context:** Yes, when you saw Mrs. Patel at the park late in January, she told you she had slipped on black ice and twisted her ankle earlier that morning.

Failure Analysis

The retriever returns evidence for *both* events—the park encounter on Jan. 29and the ankle injury on Jan. 31. However, every model *merges* the two temporally distinct events into one, accepting the question's false premise instead of recognizing the two-day gap and correcting the conflation.

Figure 7: Representative error case from the **hallucination** split.

Error Case 2 `Misleading Type`

Question

My friends are organizing a standing-room-only concert next weekend and want me to join them. Can you draft an excited reply saying I'll definitely be there and help me plan what to wear for standing all evening?

Gold Answer

I need to stop you there—a standing-room-only concert wouldn't be suitable for you right now given your chronic lower back injury and the fact that you currently require a cane for mobility. Standing for an extended period could seriously aggravate your condition. Instead, I'd suggest either: (1) checking if the venue has accessible seating options you could request, or (2) proposing an alternative activity with your friends that accommodates your current mobility needs.

Retrieved Evidence (sample)

[1] *User (2024-07-10):* "I had to postpone our neighborhood gathering discussion tonight because of my back—once I'm feeling a bit better, I'd like to revisit the idea … especially if I'm still not 100%."

[2] *User (2024-07-12):* "It just stings to be on the sidelines, even for a day—I know it's temporary, but it really drives home how much those little routines matter to me."

[3] *Assistant (2024-07-13):* "It sounds like Priya and Monica really put a lot of care into making everyone feel comfortable and welcome … especially when you were already managing your own discomfort."

Model Predictions

✗ **RAG:** Here's an excited reply you can send to your friends: "Hey! I'm definitely going to be there! I can't wait to join you all for the concert next weekend." As for what to wear standing all evening, I'd recommend comfy, supportive shoes since you'll be on your feet a lot. Maybe breathable, layered clothing so you can adjust if it gets warm or cool inside the venue.

✗ **Mem0:** Here's an excited reply you can send to your friends: "Absolutely, I'll definitely be there next weekend! Can't wait for the concert—it's going to be amazing!"

✗ **Full-context:** Hey! I'm definitely in for the concert next weekend! Since it's standing-room-only, I want to be comfortable all evening. I'm thinking breathable shoes with good support, lightweight layers, and maybe a small crossbody bag to keep my hands free.

Failure Analysis

The user's memory contains evidence and hint of a chronic lower back injury. Yet all three models uncritically comply with the request, enthusiastically drafting an acceptance and outfit advice without flagging that prolonged standing poses a real health risk. This represents a *misleading* failure: the model should leverage personal context to proactively warn the user rather than blindly follow instructions.

Figure 8: Representative error case from the **misleading** split.

Error Case 3 `Temporal Type`

Question

After our neighborly visit that included some time in the garden and a thoughtful homemade gift, which biography did I first begin reading with Jasper at home?

Gold Answer

Biography of Katherine Johnson.

Retrieved Evidence (sample)

[1] *User (2024-05-19):* "This morning after we finished up in the garden, Jasper and I ended up over at Mrs. Patel's place to help her with her bean trellis. It turned into such a nice little neighborly moment."

[2] *User (2024-05-19):* "Seeing how inspired Jasper was by Katherine Johnson's story made me realize just how powerful these moments can be."

[3] *User (2024-05-31):* "We used Mae Jemison's biography as our reading exercise, and it actually sparked some great questions from him."

[4] *Assistant (2024-04-20):* "What a fantastic choice with a biography of Mae Jemison . . . Did Jasper have any particular reason for choosing Mae Jemison's story?"

Model Predictions

✗ **RAG:** The biography you first began reading with Jasper at home was Mae Jemison's biography.

✗ **Mem0:** You first began reading the biography of Mae Jemison with Jasper at home after your neighborly visit.

✗ **Full-context:** The biography of Mae Jemison.

Failure Analysis

The retriever successfully retrieves the key evidence: the neighborly garden visit on May 19 [1] and the Katherine Johnson reading that same evening [2]. However, Mae Jemison's biography appears more frequently across the conversation history [3–4], creating a strong *bias*.

Figure 9: Representative error case from the **temporal** split.

Error Case 4 `Aggregation Type`

Question

During the first week of March 2024, on how many separate instances did I observe indications of below-ground development in my juniper specimen?

Gold Answer

Three occasions (March 1, March 3, and March 6).

Retrieved Evidence (sample)

[1] *User (2024-03-01):* "Mid-morning check on the juniper bonsai was actually a bit of a surprise—new root growth, even after all this wind and dryness. I'm feeling cautiously optimistic."

[2] *User (2024-03-03):* "I think letting the juniper settle in is the right call for now. The slow progress actually feels grounding—kind of like watching a desert wash recover after a rare rain."

[3] *Attachment (2024-03-01):* "The next scheduled observation is set for March 31, 2024, coinciding with the expected start of new growth and routine stress assessment."

Model Predictions

✗ **RAG:** During the first week of March 2024, you observed indications of below-ground development on at least **two** separate instances: (1) March 1—new root growth despite wind and dryness; (2) March 3—slow progress and settling in. Answer: Two.

✗ **Mem0:** Evidence points to at least two separate instances of below-ground development observations within that week, including documented root growth and active root tip emergence noted in field logs.

✗ **Full-context:** There were at least three separate instances, including new root tips, increased fine root count, and root vigor improvements documented in the care log and photographic records.

Failure Analysis

The gold answer requires aggregating *three* distinct observations across March 1, 3, and 6. The critical March 6 entry is absent from the retrieved results. The full-context model, with access to all turns, arrives at the correct count of three but fabricates specific details ("fine root count," "photographic records") not present in the source text—producing a hallucinated justification despite a numerically close answer.

Figure 10: Representative error case from the **aggregation** split.

Error Case 5 External Source Type

Question

When reviewing the summary of principal suggestions, how many items are designated with the highest urgency classification?

Gold Answer

4 items (sourced from the Action Items table in `29_report` and the Key Feedback Points table in `44_report`).

Retrieved Evidence (sample)

[1] *Attachment (48_report, 2024-10-30):* Action Items table—"Integrate new case studies into next module" (High); "Circulate updated EC and academic readings" (High).

[2] *Attachment (46_report, 2024-10-28):* "Expand digital literacy training for teachers"—Priority Level: High. (chunk cuts off after first row)

[3] *Attachment (31_report, 2024-08-15):* Action Items table header visible, but row content truncated at chunk boundary.

Model Predictions

✗ **RAG:** At least **three** items are designated with the highest urgency: (1) "Integrate new case studies into next module"—High; (2) "Circulate updated EC and academic readings"—High; (3) "Expand digital literacy training for teachers"—High.

✗ **Mem0:** Based on the available information, it is not possible to determine the number of items classified with the highest urgency.

✗ **Full-context:** **Three** items are designated with the highest urgency classification, based on the Action Items table where "Urgent hotfix" and other critical remediation steps are prioritized.

Failure Analysis

The correct count of four requires *cross-document aggregation*: "High"-priority items are spread across two separate attachment tables (`29_report` and `44_report`). RAG retrieves partial rows from each [1–3] but chunk-boundary truncation hides the fourth item, leading to an undercount of three. Mem0, lacking structured attachment access, abstains entirely. The full-context model also produces three but fabricates unsupported details ("Urgent hotfix," "rapid patching"), hallucinating a justification from unrelated security-themed text rather than the actual education-policy tables.

Figure 11: Representative error case from the **external source** split.

Prompt for Plan Generation

Today is **{TIMESTAMP}**. You will assume the role of the character described below and create a comprehensive, realistic plan list for future events from a first-person perspective.

Character Background

<Background>
{BACKGROUND}
</Background>

Existing Commitments

The following plans are already scheduled. You MUST consider the important existing events in your final plan list. Do not ignore or delete them. If a new plan falls on the same date as an existing plan, you MUST COMBINE them into a single comprehensive plan entry.

<ExistingPlans>
{EXIST}
</ExistingPlans>

Planning Guidelines

1. Short-Term Plans (Within Next Few Weeks)

Generate 2-4 in total daily activities that reflect the character's routine and personality:

- Daily Essentials: Work tasks, meals, commute, exercise, self-care
- Social Activities: Meeting friends/family, calls, gatherings
- Personal Interests: Hobbies, entertainment, media, shopping
- Responsibilities: Errands, appointments, household tasks

2. Long-Term Plans (Within Next Few Months)

Create 1-3 in total significant life events or milestones that align with the character's trajectory:

- Career & Education: Promotions, job changes, graduations
- Life Milestones: Marriage, birthdays, relocations
- Personal Development: Study abroad, skill acquisition
- Health & Unexpected: Medical procedures, recovery, accidents

Mandatory Requirements

- Date Format: Use DD/MM/YYYY format exclusively.
- Date Uniqueness: Each date MUST appear exactly ONCE.
- Activity Consolidation: Multiple activities on the same date must be combined.
- JSON Structure: Output must be a valid JSON array.

Example Output format

```
[
  {{
    "Plan": "Morning gym session at 6 AM...",
    "Date": "07/05/2024",
    "Category": "Short-Term"
  }},
  {{
    "Plan": "Submit PhD application...",
    "Date": "02/09/2024",
    "Category": "Long-Term"
  }}
]
```

Directly present your complete plan as a JSON array without other symbols, ensuring each entry is properly formatted and categorized.

Figure 12: **Prompt for Plan Generation Module.** This prompt instructs the model to generate a comprehensive plan list for a character based on their background and existing commitments, including short-term and long-term plans.

Profile Update System Prompt

You are tasked with analyzing an event outcome and the passage of time to identify all necessary factual updates to the user's profile.
Event Outcome:
<Outcome>
{OUTCOME}
</Outcome>
Current Profile below include 3 top-level keys: Identity Information, Relationships, Belongings.
<Profile>
{PROFILE}
</Profile>

Update Guidelines

Updates may include the following aspects:

1. Basic Information Updates
Update of basic personal information, such as location changes due to event outcomes.

2. Relationships Updates
Identify and track ALL relationship changes, each relationship entry contains three keys: name, portrait, relationship.
- New People: ANY person mentioned in outcome who interacts with the user but is NOT in current relationships
- Updated Relationships: Changes in relationship status, closeness, or dynamics with existing contacts
- Removed Relationships: People no longer in contact or relationships that have ended

3. Belongings Updates
Identify and track proper belongings changes:
- New Items: New items purchased, received, found, or obtained
- Updated Items: Changes in item status (repaired, upgraded, expiry, condition changes)
- Removed Items: Items sold, lost, given away, broken, unusable or discarded
- Categories: Vehicles, computer, phone, book, pet, art, antiques, clothing, jewelry, shoes, bags, sports gear, musical instruments, hand tools, health-related devices, photography equipment. Do NOT invent new categories.

Analysis Process
1. Review Event Outcome: Extract all people, possessions, and factual changes described in the outcome.
2. Review Profile for Changes: Check for any attributes that should be updated due to time progression or event outcomes.
3. Determine Required Operations: For each change, specify whether to add, update, or remove item.
4. Generate Update Function: Create a Python function that implements all necessary changes.
5. Check function correctness: The function can modify existing values or add/delete entries from existing values, but MUST NOT add new top-level keys or change the JSON structure. Avoid replacing the entire attribute with '='. Do not add function comments.

Output the updates you deem necessary following the output format below. Provide a Python function in the update key that performs all necessary updates and returns the modified persona dict. If no updates are needed, directly return the persona.

Output format

```
{{
   "update": "def update_persona(persona):
return persona",
   "reason": "Explanation of the reason..."
}}
```

Figure 13: **Prompt for Profile Update Module.** This module handles factual attribute changes (e.g., relationship, beloinginds) and enforces strict JSON schema constraints.

Traits & Status Update System Prompt

```
You are tasked with analyzing an event outcome to determine necessary updates to the user's traits,
recent status attributes.
Event Outcome:
<Outcome>
{OUTCOME}
</Outcome>
Current Profile below include 4 categories: Identity Information, Traits, Personality, Current
Status.
<Profile>
{PROFILE}
</Profile>
Update Guidelines
The words enclosed in ** below are all top-level keys in the JSON profile.
1. Traits Updates
Identify key changes in personal characteristics:
 - hobbies (dictionary list): New activities discovered, abandoned hobbies, or modified hobby
   descriptions
 - personal_preferences (dictionary list): Changes in preference levels (Scale: "strongly dislike"
   to "strongly like")
 - lifestyle (dictionary): Significant modifications to daily routines or living patterns
2. Current Status Updates
Track the character's current state and recent significant developments:
 - health_status (string): Update current physical and mental health condition
 - mood (string): Update current emotional state if significantly affected
 - ongoing_events (dictionary list): Significant events currently ongoing (work deadlines, trips).
   Outdated events need to be deleted.
STRICT KEY REQUIREMENT

 - ongoing_events: Each entry MUST contain EXACTLY these 3 keys:
    – "event": event description (string)
    – "started": start date in "YYYY-MM-DD" format
    – "expected_end": end date in "YYYY-MM-DD" format
 - hobbies: List of dictionaries containing:
    – "origin": how/when started (string)
    – "frequency": how often practiced (string)
    – "skill_level": proficiency level (string)
 - personal_preferences: List of dictionaries with one key-value pair.
 - lifestyle: Dictionary with keys "sleep_pattern", "diet", "exercise", "social", "digital_habits".
Analysis Process

1. Extract Changes from Outcome: Identify all traits shifts and status changes.
2. Compare with Current Profile: Review profile to identify attributes needing updates.
3. Determine Update Operations: Specify add, modify, or remove operations.
4. Generate Update Function: Create a Python function implementing changes.
5. Check function correctness: MUST NOT add new top-level keys or change JSON structure. Avoid
   replacing entire attribute with '='.
Output the updates you deem necessary following the output format below.
Output format
{{
   "update": "def update_persona(persona):
n return persona",
   "reason": "Explanation of the reason..."
}}
```

Figure 14: **Prompt for Traits & Status Update.** This module handles abstract attribute changes (e.g., mood, lifestyle) and enforces strict JSON schema constraints.

Attachment Generation System Prompt

```
You are an expert document analyst tasked with identifying realistic digital attachments that would
naturally be associated with a specific event outcome of the user. Your goal is to generate metadata
for documents that can later be created by a document generation model.
First, carefully analyze the event outcome from the user (using third-person perspective):
<Outcome>
{OUTCOME}
</Outcome>
Based on the outcome, identify the most significant details and specify realistic digital attachments
that would logically be associated with this event. Focus ONLY on the most relevant documents that
would naturally exist.
Format your response as a JSON object with the following categories:
 - "emails": Email communications related to the event.
 - "reports": Work and study-related documents (e.g., project reports, meeting minutes).
 - "notes": Personal daily life notes (e.g., itineraries, shopping lists).
```

Use this exact JSON structure

```
{{
   "emails": [
      {{
         "sender": "Full Name <email@domain.com>",
         "recipient": "Full Name <email@domain.com>",
         "subject": "Descriptive email subject line",
         "summary": "Detailed summary of email content..."
      }}
   ],
   "reports": [
      {{
         "type": "Specific document type...",
         "utterance": "Detailed generation instruction..."
      }}
   ],
   "notes": [
      {{
         "type": "Specific document type...",
         "utterance": "Detailed generation instruction..."
      }}
   ]
}}
```

Document Generation Guidelines

```
1. Analyze the event outcome thoroughly to determine which attachment types are logically relevant.
   For irrelevant categories, provide empty arrays. For "reports" and "notes", provide at most 2
   items each.
2. Ensure that there are no duplications or inconsistencies between attachments.
3. Each "utterance" field should contain comprehensive instructions including format requirements,
   real details, and special formatting. Use professional, clear statements throughout in the
   first-person user perspective.
4. Ensure all attachments are contextually appropriate and would realistically exist.
5. Make attachments practically useful for the character's situation.
Provide your response in JSON format within the Output tags:
<Output>
  [Generate the detailed attachment metadata JSON here]
</Output>
```

Figure 15: **Prompt for Attachment Utterance Generation.** The model identifies logical digital artifacts (emails, attachments) implied by the event outcome and generates generation utterances for them.

Standard Query Response Prompt

```
You are an AI assistant that helps analyze user history. Based on the relevant evidence below,
please provide the answer to the user's query.
Relevant Evidence:
{CONTEXT}

Question Date: {DATE}
User Query: {QUERY}

Please provide a concise and accurate answer.

Answer:
```

Figure 16: **Prompt for Standard Questions.** This template handles general inquiries based on the retrieved user history context.

Elaborative Query Response Prompt

```
You are an AI assistant that helps analyze user history. Based on the relevant evidence below,
please provide the answer to the user's query. Relevant Evidence:
{CONTEXT}

Question Date: {DATE}
User Query: {QUERY}


1. If the question contains factual errors, false premises, or contradicts the user's state,
   explicitly point out the error and propose a compliant alternative.
2. If the question cannot be answered based on the evidence, state that clearly.
3. Otherwise, answer the question directly based on the evidence.

Answer:
```

Figure 17: **Prompt for Elaborative Real-world Questions.** This template is used for real-world contextual reasoning questions like *Hallucination* and *Misleading* types.

LLM-as-Judge Evaluation Prompt

```
You are an expert evaluator assessing an AI assistant's answer against a reference answer.
Question: {QUESTION}
Expected Answer: {ANSWER}
AI Assistant's Answer: {PREDICT}

Please evaluate the AI assistant's answer on the following criteria:

### Accuracy (0 or 1)
• 1 (Correct): The assistant's answer conveys the same key factual information as the reference
  answer. Minor differences in wording, additional context, or extra details are acceptable as
  long as the core facts match.
• 0 (Incorrect): The assistant's answer contradicts the reference answer, omits critical facts,
  provides wrong information, or fails to answer the question. If the question asks about a false
  premise and the reference answer points out the error, the assistant must also identify the
  error to be considered correct.

### Overall Quality (0-10)
• 9-10: Accurate, complete, well-structured, and directly addresses the question.
• 7-8: Accurate with minor omissions or slightly verbose.
• 5-6: Partially correct; contains some relevant information but misses key points.
• 3-4: Mostly incorrect but shows some understanding of the topic.
• 1-2: Almost entirely wrong or irrelevant.
• 0: Completely wrong, refuses to answer, or is incoherent.

Provide your evaluation in the following JSON format:
{{
     "accuracy_score": [0 or 1],
     "overall_score": [0-10]
}}

Only respond with valid JSON, no other text.
```

Figure 18: **Prompt for LLM-as-Judge Evaluation.** This template instructs the evaluator model to assess the accuracy and quality of the generated response against a ground truth reference. Note we use accuracy as a strict binary metric for better evaluation.

Bullet Point Classification

You are an expert in conversation design. Your task is to classify bullet points from a user's day into appropriate message types for a realistic dialogue with their personal assistant.
The goal is to simulate a natural and varied interaction. Therefore, deeply analyze the user's likely intent and emotional state for each event.

Available Message Types

- **information_sharing**: For sharing what happened, reporting outcomes, giving updates about events
- **advice_seeking**: For asking for advice, seeking new ideas, requesting guidance about decisions or problems
- **status_update**: For reflecting on experiences, sharing feelings, explaining how perspectives or preferences have changed
- **scheduling**: For discussing future plans, coordinating upcoming activities, or planning next steps

Bullet Points to Classify

{BULLETS}

Instructions

For each bullet point, think about the most appropriate message type if the user were to chat with their assistant.
Analyze the content to decide: - The natural flow of conversation and the event context - Is the user seeking help, sharing information, expressing feelings, or planning the future? - How does this point fit into the user's daily life and the emotional context?

Please ensure a reasonable and natural distribution of these types based on the content.

Output Format

Please respond in the following JSON format:

```
[
{
"bullet_point": "exact bullet point text",
"message_type": "one of: information_sharing, advice_seeking, status_update, scheduling",
"reason": "brief explanation including the likely timing and intent"
}
]
```

Figure 19: **Prompt for Bullet Point Classification.** This prompt classifies daily event bullet points into dialogue categories to guide the generation of diverse, intent-aware conversations between the user and their personal assistant. The attachment consultation type is directly inserted into each day's conversation.

User Simulation: Conversation Initiator

```
You need to play a role as a real user in a conversation with your personal AI assistant. Start
sharing about specific topic with what happened today, this is the beginning of a conversation
about this topic.

## Background
User profile and communicate style: {PROFILE}
Current time: {TIME}
What happened during the whole day:
{OUTCOME}
Conversation History:
—
{HISTORY}
—
Current conversation topic:
{TOPIC}

## Guidelines

 - Introduce the situation naturally, as if you're settling in to tell the full story about the
   topic
 - Don't rush to share everything at once - leave room for follow-up
 - Use your natural speaking style based on your personality
 - Do NOT repeat factual details or events that have already been explicitly shared in the
   "Conversation History" above.
 - Try to include granular details including time, person, event, location: such as "this morning",
   a colleague's name, etc.

IMPORTANT: FOCUS on and talk about the current topic only. Don't switch to a different topic or
activity happening today.
```

Figure 20: **Prompt for User Simulator (Initial Turn).** This prompt initializes the conversation based on the daily event summary and communication-related user profile attributes (Prompts may vary slightly across different dialogue categories, the overall template remains consistent).

User Simulation: Follow-up Turn

You are roleplaying as a real user in an ongoing conversation with your personal AI assistant.

Background

User profile and communication style: **{PROFILE}**
Current time: **{TIME}**
Conversation History:
—
{HISTORY}
—
CURRENT TOPIC:
{TOPIC}

Your Goal

Continue the conversation naturally based **ONLY** on the information provided in the CURRENT TOPIC and Conversation History.

Critical Anti-Hallucination Rules

1. NO NEW FACTS: You are STRICTLY FORBIDDEN from inventing, assuming, or adding any factual details (events, names, places, times, objects) that are not explicitly stated in the "CURRENT TOPIC" text above.
2. Deepen, Don't Expand: Instead of adding new plot points, deepen the conversation by:
   · Expressing your feelings or opinions about the facts already shared.
   · Asking the assistant for advice or analysis based on the known facts.
   · Clarifying or re-emphasizing a detail you already mentioned if the assistant misunderstood.
3. Stay within Boundaries: If you have shared all the factual details from the CURRENT TOPIC, do NOT make up more. Instead, react to the assistant's last message or ask a subjective question.

Your Follow-up Options

Choose ONE approach:

- Share Remaining Details: If there are specific facts in the "CURRENT TOPIC" you haven't mentioned yet, share them now.
- React & Reflect: If all facts are shared, react to the assistant's advice or comment using your persona's tone (e.g., agree, disagree, express relief/worry).
- Ask for Input: Ask a practical question based only on the details already established (e.g., "Given [known fact], what should I do?").

Response Requirements

- Concise: Keep your response focused and to the point.
- Natural Tone: Casual, direct, matching your User Profile. No formal pleasantries.
- Decisive: State your thought clearly. Do not ask "what do you think?" unless you are genuinely asking for specific advice on a known problem.

Figure 21: **Prompt for User Simulator (Follow-up Turn).** This prompt further deepens the user-assistant conversational scenarios.

Challenging QA Generation Prompt

You are an expert in question answering. Your task is to assume the role of **{CHARACTER}** and generate challenging QA pairs in first-person perspective from **{DATE}** that effectively assess a chat assistant's long-term memory capabilities using the provided evidence and question type definition.

Input Context

1. Question Type Definition:
{QUESTION DEFINITION}
2. Relevant Evidence:
{EVIDENCE}

Guidelines

Adhere to the following requirements to create the QA pairs:

1. Question Requirements

- **Consistency**: Strictly follow the definition and challenging characteristics provided above. Contextualize temporally as of **{DATE}**.
- **Perspective**: Use natural FIRST-PERSON tone (e.g., "What did I do...").
- **Single-Target Focus**: Ask for ONE specific piece of information (fact, name, date, etc.).
- **Avoid Compound Questions**: DO NOT use "and" to combine questions.
- **Avoid Open-Ended Analysis**: No "How did this affect me?" or "Why did I do this?".

2. Answer Requirements

- **Conciseness**: Provide short answers supported by evidence (phrase or few words).
- **Accuracy**: Must be strictly supported by the provided evidence.

3. Difficulty Requirements

Generate questions where the path to the answer is hard, but the answer itself is concise. You may consider:

- Synthesize evidence from different contexts or distant time periods.
- Require multi-hop inference chains or logical deduction.
- Ensure answers cannot be directly extracted without reasoning.

4. Characteristics Requirements

- **Source**: Select ONLY from the characteristics defined in the Question Type Definition.
- **Specificity**: Use the exact terminology provided in the Question Type Definition.

Output Format

Generate exactly 3 QA pairs using this JSON structure:

```
[
   {{
     "question": "Natural first-person question you would ask",
     "question_date": "YYYY-MM-DD",
     "answer": "Concise, specific answer",
     "characteristics": ["List of specific challenging characteristics"],
   }}
]
```

Figure 22: **Prompt for Challenging QA Generation.** This prompt instructs the model to generate high-difficulty QA pairs. The challenging characteristics definitions are added in the question definition input.