MemBench: Towards More Comprehensive Evaluation on the Memory of LLM-based Agents

Haoran Tan[1*‡§], **Zeyu Zhang**[1*‡§], **Chen Ma**[1‡§], **Xu Chen**[1†‡§], **Quanyu Dai**[2†], **Zhenhua Dong**[2]

[1]Gaoling School of Artificial, Renmin University of China, Beijing, China,
[2]Huawei Noah's Ark Lab
{tanhaoran1321,zeyuzhang,xu.chen}@ruc.edu.cn, daiquanyu@huawei.com

Abstract

Recent works have highlighted the significance of memory mechanisms in LLM-based agents, which enable them to store observed information and adapt to dynamic environments. However, evaluating their memory capabilities still remains challenges. Previous evaluations are commonly limited by the diversity of memory levels and interactive scenarios. They also lack comprehensive metrics to reflect the memory capabilities from multiple aspects. To address these problems, in this paper, we construct a more comprehensive dataset and benchmark to evaluate the memory capability of LLM-based agents. Our dataset incorporates factual memory and reflective memory as different levels, and proposes participation and observation as various interactive scenarios. Based on our dataset, we present a benchmark, named MemBench, to evaluate the memory capability of LLM-based agents from multiple aspects, including their effectiveness, efficiency, and capacity. To benefit the research community, we release our dataset and project at https://github.com/import-myself/Membench.

1 Introduction

In recent years, large language models (LLMs) have demonstrated remarkable capabilities in processing natural languages and performing complex tasks across various domains (Zhao et al., 2023; Wu et al., 2024b). However, vanilla LLMs typically operate in static scenarios, without interacting with external environments, thereby limiting their potential advancement toward artificial general intelligence (AGI). To address this limitation, many recent works propose LLM-based agents with extra modules besides the foundation models, enabling them to interact with environments with autonomous learning and dynamic adaptation (Wang et al., 2024a; Xi et al., 2025). Among them, the memory module serves as an essential foundation for saving critical information and accumulating experiences. It empowers LLM-based agents to better meet the demands of dynamic tasks, as well as evolve within their environments continuously (Zhang et al., 2024a).

Some previous studies evaluate the memory capability of LLM-based agents in a subjective way, which adopts human evaluators or LLMs to score the memory process (Zhong et al., 2024). Other studies focus on the evaluation in an indirect way (Packer et al., 2023). They measure the task performances of agents conditional on different memory mechanisms, where a better memory mechanism generally leads to better performances. Recently, some studies introduce long-term dialogue datasets, which can be used to evaluate the long-term memory capabilities of LLM-based agents objectively (Wu et al., 2024a).

However, previous works have some limitations on evaluating the memory capability of LLM-based agents. First of all, most of them provide insufficient evaluation of the different levels of memory capabilities, which primarily focus on factual memory while neglecting reflective memory. Here, we define the factual memory as a low-level type of memory that involves information that is explicitly provided. In contrast, the reflective memory stands a higher level, which is not explicitly stated but can be implicitly reflected. For example, a user's taste preferences represent reflective memory, while their preference for specific dishes is factual memory. Second, most of them are limited to participation scenarios, where the agent interacts with the user from a first-person perspective. However, in the agent's daily usage, there are also observation scenarios, where the agent observes and

[*]Co-first authors.
[†]Corresponding authors.
[‡]Beijing Key Laboratory of Research on Large Models and Intelligent Governance
[§]Engineering Research Center of Next-Generation Intelligent Search and Recommendation, MOE

Findings of the Association for Computational Linguistics: ACL 2025, pages 19336–19352
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

records the user's messages from a third-person perspective. Moreover, most of them are just focusing on the effectiveness of memory mechanisms without considering their efficiency and capacity, which is also significant in real-world applications.

To address these limitations, we propose a more comprehensive dataset and benchmark to evaluate the memory capability of LLM-based agents. The major features of our dataset and benchmark are presented as follows:

Multi-scenario Dataset. To evaluate the agent's memory capabilities across different scenarios, our dataset includes data from two common usage scenarios. The first is the participation scenario, where the agent interacts with the user. The second is the observation scenario, where the agent is assumed as the role of an observer and required to record the information provided by the user.

Multi-level Memory Content. Our dataset focuses on both factual memory and reflective memory, enabling a comprehensive evaluation of the memory capability of LLM-based agents. It allows for the evaluation of memory capabilities in tasks including information extraction, cross-session reasoning, knowledge updating, temporal reasoning, as well as reflective summarization.

Multi-metric Evaluation. Based on our dataset, we introduce a multi-metric benchmark to evaluate the memory capabilities of LLM-based agents. To provide a comprehensive and obvious assessment of the various aspects of the agent's memory performance, we offer four evaluation metrics, including accuracy, recall, capacity, and temporal efficiency.

In summary, we introduce a dataset featuring multi-scenario and multi-level content, which is distinctively different from previous datasets. Additionally, we introduce a more comprehensive benchmark with mutli-metric evaluations. To benefit the research community, we have released our dataset and project at Github repository*. In the following parts, we provide the related works in Section 2. We illustrate the process of data construction in Section 3, and present the benchmark with analyses in Section 4. Finally, we draw conclusions of our paper in Section 5.

2 Related Works

In recent years, LLM-based agents have been widely applied in many fields, such as recommendation (Wu et al., 2024b), finance (Ding et al., 2024) and personal assistants (Li et al., 2024), because of their great capabilities in solving complex tasks and interactive scenarios (Wang et al., 2024a). Among the various abilities of agents to solve problems, memory is one of the most important, which is responsible to store observed information and recall relevant experiences, in order to support LLM inference (Zhang et al., 2024a). The evaluation on the memory capability of LLM-based agents is a critical problem for developing advanced memory.

Table 1: The comparision among different datasets. PS indicates Participation Scenario. OS indicates Observation Scenario. FM indicates Factual Memory. RM indicates Reflective Memory.

Datasets	Profiles	Scenarios	Levels
PerLTQA	✓	PS	FM
LoCoMo	×	PS	FM
LongMemEval	×	PS	FM
MemBench	✓	PS & OS	FM & RM

Previous datasets used for memory evaluation mainly come from dialogue datasets designed to evaluate chat assistants, focusing on assessing the factual memory capabilities of the assistant. LoCoMo (Maharana et al., 2024) constructs long-term conversations with LLM-expanded personalized descriptions and temporal event graphs, creating conversations for multiple evaluation tasks. LongMemEval (Wu et al., 2024a) builds a user-assistant interaction dataset with attribute ontology and timestamped history. Xu et al. (2022) adapts PersonaChat (Zhang et al., 2019) through translation and role-playing, with annotated personalization usage and partial information visibility. Bai et al. (2023) propose a bilingual benchmark in English and Chinese, covering comprehensive tasks like Q&A, summarization, and code completion. An et al. (2023) provide a dataset designed to evaluate long-context language models across diverse domains and input lengths. PerLTQA (Du et al., 2024) generates character profiles and events with ChatGPT and Wikipedia, creating profiles, events, and QA pairs after manual validation.

As we have shown in Table 1, most of these datasets lack diverse scenarios, focusing only on participation scenarios (PS) and overlooking the agent's observation scenarios (OS). Additionally, they only focus on factual memory (FM), neglecting reflective memory (RM). Some works do not include the profiles of users. Previous studies typically use these datasets in long-context evaluation methods that do not align with the agent's mem-

*https://github.com/import-myself/Membench ory process. Moreover, the evaluation metrics used with these datasets are not comprehensive.

Compared with previous works, our work is the first study that emphasizes reflective memory, puts forward observation scenarios, adopts evaluation methods that are better suited to the agent's memory process, with more comprehensive metrics.

3 Dataset Construction

3.1 Pipeline of Data Generation

Inspired by MemSim (Zhang et al., 2024b), we expand the dataset for memory evaluation based on this framework. Building upon the question types of factual memory included in Memsim, we extend the evaluation of the ability for memory knowledge updating and extracting information from the assistant's response in single or multi-session. In addition, we incorporate reflective memory generation methods and extend observation scenario to participation scenario. The dataset creation process is as follows and as shown in Figure 1.

User's Relation Graph Sampling. Following the approach of Memsim, we create a relation graph composed of user profiles and their related entities including individuals, events, places, and items. Based on Memsim's method for sampling attributes related to factual memory, we propose a method for sampling of high-level attributes related to reflective memory. To better fit the distribution of high-level attributes in the real world, we leverage user-item relationship pairs and relevant ratings from three recommendation datasets, including MovieLens (Harper and Konstan, 2015), Food (Majumder et al., 2019), and Goodreads (Wan and McAuley, 2018; Wan et al., 2019). We extract each user's high-level preferences in each recommendation dataset by identifying the most frequent category of items with which he or she likes or rates positively. If there is no category information, we utilize LLMs(GPT-4o-mini) to summarize the high-level preferences corresponding to these positive relation items. We assign high-level preference attributes using either random matching or matching based on identical attributes, and then we obtain the user's relation graph shown in Figure 2. At the same time, we construct three one-to-many mappings between high-level preferences and low-level factual attributes with LLMs or the item-category relationships from the recommendation datasets.

Memory Dataset Construction. Memsim provides a data creation process for the observation scenario. We expand it to the participation scenario using the self-dialogue method. When selecting high-level preference attributes, we should choose multiple low-level preference attributes from the mapping dictionary and use these low-level preference attributes to generate the corresponding evidence dialogues. For example, a user might say, "I like the movie *Star Wars*". To ensure fluency of the conversation, the specific discussions about the low-level preferences, such as the discussion about content of the movie, are inserted between the key dialogues to form a complete conversation. Finally, the remaining relevant attributes are used to generate multi-turn dialogues. The evidence dialogues will be inserted into them to form a complete session. We introduce a time-based session division approach, where the timestamp within a session is assigned continuously to each turn dialogue, typically with short intervals, such as one minute. The timestamp across different sessions maintains a sequential order, but the time gap between two adjacent sessions is typically longer, such as one day.

3.2 Multi-scenario Memory

The interactive scenarios of the agent can be categorized into two types, including the participation scenario and the observation scenario. In the participation scenario, the agent interacts with the user, while in the observation scenario, the agent serves only as an observer, recording user-inputted messages. In the participation scenario, other modules of the agent, such as the reasoning module, will affect the memory module. However, in the observation scenario, the agent does not perform actions and thus does not influence memory. These two scenarios cannot be considered the same. Therefore, we provide the following two types of datasets:

Participation Memory Scenario. The participation memory scenario is represented by the dialogue between the user and the agent, which is the agent's typical usage scenario. To eliminate the influence of other modules of the agent, we predefine the agent's responses to the user's expressions. In the user-agent dialogue interaction, the agent's memory not only needs to remember the message expressed by the user but also needs to store the message of the agent's responses, such as the agent's reply when the user requests a recommendation. In our dataset, the data for participation scenario dataset consists of sessions composed of many turns in dialogues.

Observation Memory Scenario. The observa-

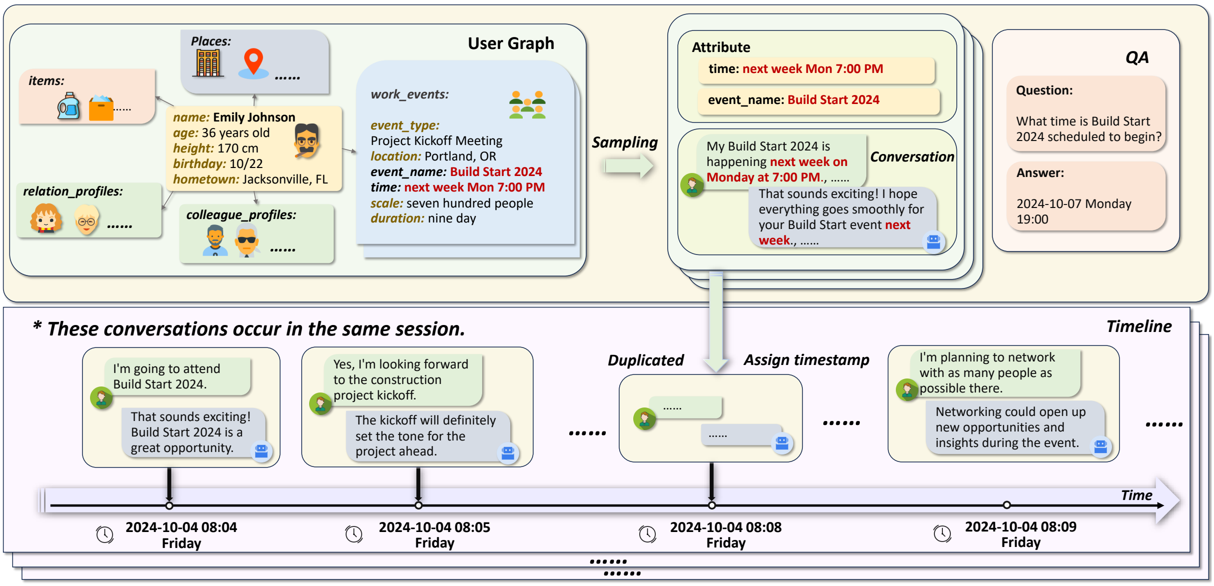


Figure 1: An example of generating dialogue data. First, the event "Build Start 2024" is extracted with the time "next week Mon 7:00 PM," which is then used to generate evidence dialogues and questions. It's merged with dialogues generated from other attributes to form a complete dialogue, and an answer is generated based on the provided time label "2024-10-07 Monday 19:00".

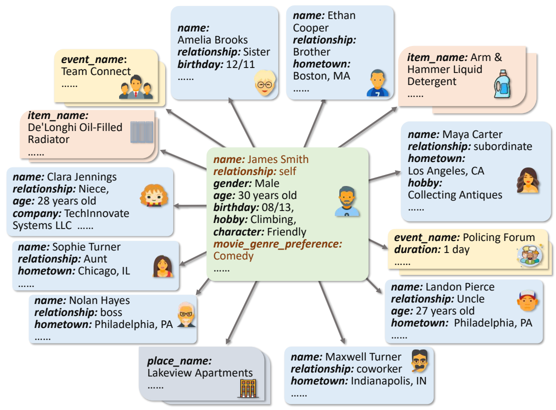


Figure 2: A user relation graph is composed of a user's profile and his or her associated entities including individuals, events, items and places.

tion memory scenario is represented by the flow of message input from the user to the agent. In this process, the agent passively receives the user's message flow over time and does not interact with the user. This scenario focuses on the agent's role as an observer, where the agent only needs to remember the message expressed by the user without taking any action. In our dataset, the data for observation scenario dataset consists of message lists composed of many messages.

3.3 Multi-level Memory

In the daily usage of LLM-based agents, we expect it to have factual memory capabilities, while also hoping that its memory mechanism can summarize and reason at higher levels to generate reflective memory. Reflective memory enables the agent to gain a more comprehensive understanding of the user, thereby improving the satisfaction of subsequent interactions. From this perspective, we divide the types of memory data and questions in our dataset into two categories:

Factual Memory. It refers to the specific factual attributes of the users or the entities associated with them, such as their relative's age or occupation, the time details of events and so on. This information will be expressed in daily dialogues between users and agents. Asking questions about these attributes can test various memory abilities of the agent. For example, in dialogues, the user may not directly express the time of an event but might use indirect references, such as "next Monday", we can evaluate the agent's ability to extract information and instantly convert time-related information by asking it the exact time of the day of the month of the event. In addition, we can also evaluate its ability to update knowledge based on different expressions of the same attribute over time. Furthermore, by designing questions that require the integration of multiple entities' attributes for answers, we evaluate the agent's memory capability in terms of its memory reasoning abilities in both single-session and multi-session contexts. These question examples are shown in Figure 3.

Reflective Memory. Reflective memory refers

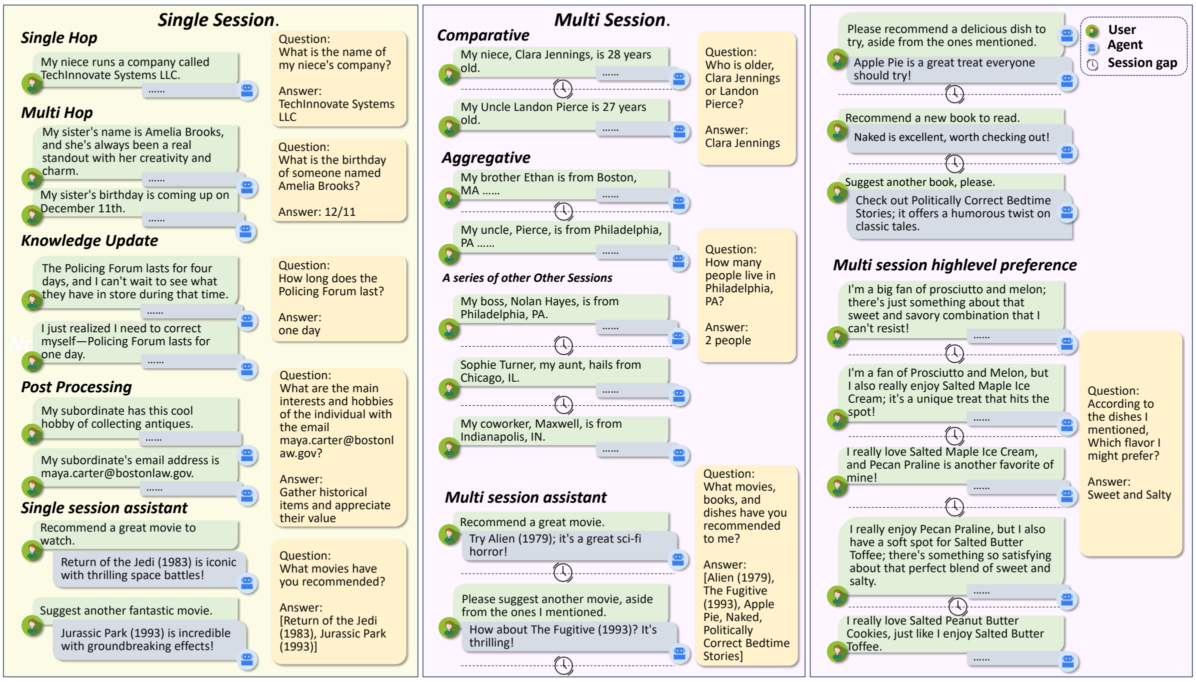
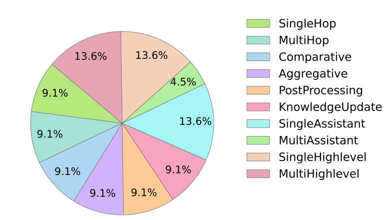
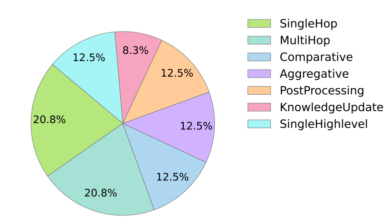


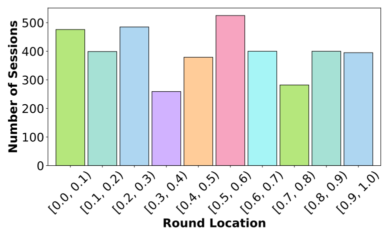
Figure 3: An overview of part categories of data used to test different abilities.



(a) The distribution of different types of questions in the participation dataset.



(b) The distribution of different types of questions in the participation dataset.



(c) The distribution of the number of different round locations of key evidence dialogue turn within a session.

Figure 4: The details of the category distribution and answer distribution in the dataset.

to the extraction and summarization of high-level preferences based on the user's expression of low-level preferences, including some factual attributes in the dialogue. For example, the user's taste preferences are inferred from his expressions of liking for different dishes. To enhance the credibility of the answers, our memory content is reinforced through multiple expressions of different factual preferences or attributes to strengthen the understanding of the agent. We can evaluate the agent's memory mechanism's ability to extract and summarize preferences at different levels.

3.4 Multi-metric Evaluation

In order to more comprehensively assess the memory mechanism of the agent, we employ a total of four evaluation metrics as follows.

Memory Accuracy. To avoid misjudgments caused by the agent's flexible expression of answers, in our evaluation dataset, all questions are set as multiple-choice questions. After the agent completes the memory process, both the questions and options will be submitted to the agent. The accuracy score of the memory is calculated by comparing the agent's choice with the true choice.

Memory Recall. For retrieval-based memory mechanisms, the accuracy of retrieval is also an important metric that needs to be measured. It not only reflects the agent's ability to effectively store and organize memory content, but also indicates the agent's efficient use of memory when answering questions. In the process of creating dialogues, we first generate key evidence dialogues for answering questions, which enables the measurement of retrieval accuracy.

Memory Capacity. We consider that the agent's memory mechanism might have a capacity limit, which is reflected in a sharp decline in accuracy when the amount of memory content reaches a certain point. This critical threshold represents the

Table 2: The statistics of our dataset. TPT indicates the average number of tokens per trajectory. PS indicates Participation Scenario. OS indicates Observation Scenario. RM indicates Reflective Memory. FM indicates Factual Memory

Data Type	# Session	# Question	# Trajectory	TPT
PS-RM	3.5k	3.5k	3.5k	2,195
PS-FM	51k	39k	8k	10,285
OS-RM	2k	2k	2k	745
OS-FM	8.5k	8.5k	8.5k	617

capacity of the memory. This phenomenon might not exist, because, for example, when evaluating the retrieval-based memory mechanisms, their accuracy depends on the effectiveness of retrieval.

Memory Efficiency. Regarding the design of the agent's memory mechanism, we need to focus not only on the accuracy and completeness of the memory but also on efficiency. Some memory mechanisms may result in excessively high processing time costs for the agent, which could be unacceptable in practical applications.

3.5 Dataset Statistics

The dataset consists of two parts: (1) 500 graphs composed of user profiles and profiles of entities associated with users, and (2) multiple dialogues between users and assistants, multiple users' messages, and corresponding questions. The quantity is shown in Table 2. In order to better simulate the distribution of the location of answer in real-world conversation, the key evidence rounds in a session are almost evenly distributed in each round in a session. As shown in Figure 4, we can see the quantity distribution of different categories and the quantity distribution of key evidence rounds in the session.

4 Benchmark

In this section, we create a benchmark based on our dataset to evaluate the memory capabilities of LLM-based personal agents. To better evaluate the upper bounds of the agent's memory mechanism capabilities, we also utilize the News dataset (DataGuy and Amoako, 2022) to generate a large amount of dialogues and messages serving as noise memory content that is irrelevant to the questions. We ensures that the content of noise data does not contain factual conflicts with memory messages, or dialogues in our evaluation dataset. It also allows us to control the difficulty of the evaluation by adjusting the proportion of noise data.

4.1 Experimental Settings

To better align with the memory process of agents in real-world scenarios, particularly the flow of time, we simulate the interaction process between the user and the agent to input the content that needs to be remembered. At time t, we input the user's statement from the t-th round, while content from the previous $t-1$ round and earlier can only be recalled through memory. In the participation memory scenario, at the t-th round, the agent not only needs to remember the user's messages but also needs to remember the response it has generated, which is predefined by us. Meanwhile, in the observation scenario, the agent only needs to remember the user's messages.

To set different levels of difficulty, we utilize the noise dataset to randomly insert some noise sessions into the adjacent sessions. By controlling the proportion, we create a dataset with an average length of over 100k tokens for each individual test. Due to the large number of datasets we have created and the complexity of the agent's memory mechanism design, we perform uniform sampling on each subset of the two different sized datasets as the final tests in this paper. In the dataset of ordinary size, we extract 120 reflective memory and 360 factual memory data in participation test data (each session has about 10K tokens), as well as 60 reflective memory data and 280 factual memory data in observation test data (each message list has about 1K tokens), formulating as Sub-dataset 1. For the 100k dataset, we extract 30 reflective memory data and 90 factual memory data in participation test data (each session has about 100K tokens), as well as 15 reflective and 84 factual memory data in observation test data (each message list has about 10K tokens), denoted as Sub-dataset 2.

To eliminate the influence of other memory mechanism designs on the agent in the evaluation results, we make no modifications to the agent's action modules or other components.Based on MemEngine (Zhang et al., 2025), we implement seven memory mechanisms, using Qwen2.5-7B as the base model for the agent applications on our benchmark, including FullMemory, RetrievalMemory, RecentMemory, GenerativeAgent (Park et al., 2023), MemoryBank (Zhong et al., 2024), MemGPT (Packer et al., 2023), and Self-Controlled Memory (SCMemory) (Wang et al., 2023). In our experiments, all methods that involve retrieval use the multilingual-e5-

Table 3: The results of different memory mechanisms on factual memory dataset. The read time (RT) and write time (WT) are presented in seconds per operation.

Method	Participation-Accuracy		Paticipation-Efficiency		Observation-Accuracy		Observation-Efficiency	
	10k	100k	RT	WT	1k	100k	RT	WT
FullMemory	0.647	0.489	0.001	<0.001	0.786	0.631	<0.001	<0.001
RecentMemory	0.639	0.422	0.001	<0.001	0.8	0.512	<0.001	<0.001
RetrievalMemmory	0.692	0.833	0.041	0.058	0.883	0.933	0.024	0.026
GenerativeAgent	0.478	0.455	0.045	6.116	0.779	0.476	0.031	6.239
MemoryBank	0.442	0.456	0.035	8.047	0.721	0.488	0.037	18.243
MemGPT	0.455	0.411	4.549	0.106	0.789	0.488	1.541	2.480
SCMemory	0.355	0.444	1.531	2.276	0.529	0.429	0.085	0.535

Method	Participation-Recall@10		Observation-Recall@10	
	10k	100k	10k	100k
RetrievalMemmory	0.776	0.749	0.847	0.769

Table 4: The results of different mechanisms on reflective memory dataset. The read time (RT) and write time (WT) are presented in seconds per operation.

Method	Participation-Accuracy		Paticipation-Efficiency		Observation-Accuracy		Observation-Efficiency	
	10k	100k	RT	WT	1k	100k	RT	WT
FullMemory	0.733	0.533	<0.001	<0.001	0.883	0.333	<0.001	<0.001
RecentMemory	0.700	0.333	<0.001	<0.001	0.867	0.400	<0.001	<0.001
RetrievalMemmory	0.692	0.833	0.036	0.057	0.883	0.933	0.026	0.028
GenerativeAgent	0.742	0.333	0.028	6.064	0.883	0.200	0.030	6.019
MemoryBank	0.692	0.400	0.033	15.705	0.900	0.333	0.032	12.827
MemGPT	0.733	0.367	1.042	<0.001	0.883	0.200	0.921	<0.001
SCMemory	0.542	0.267	0.036	0.057	0.783	0.333	0.025	0.028

small (Wang et al., 2024b) for retrieval.

4.2 Evaluations on Factual Memory

The test results for factual memory are shown in Table 3. FullMemory, RetrievalMemory, and RecentMemory perform better than other memory mechanisms on Sub-dataset 1. However, on Sub-dataset 2, FullMemory and RecentMemory exhibit a certain degree of decline, as the target message may fall outside the memory window. Due to the smaller window size of RecentMemory, the decline is more obvious. Other designed memory mechanisms did not show significantly superior performance in our evaluation, which might be due to flaws in these memory mechanisms. Additionally, it is important to note the time consumed by these memory mechanisms when reading and writing each round of message, especially MemGPT, which takes a longer time to read information, and MemoryBank, which takes longer to write information.

The previous evaluation works are not focusing on the design of agent memory mechanisms and solely provided factual memory datasets, so it could not adequately discuss the agent's ability to summarize reflective memory. In the following part, we evaluate on the reflective memory.

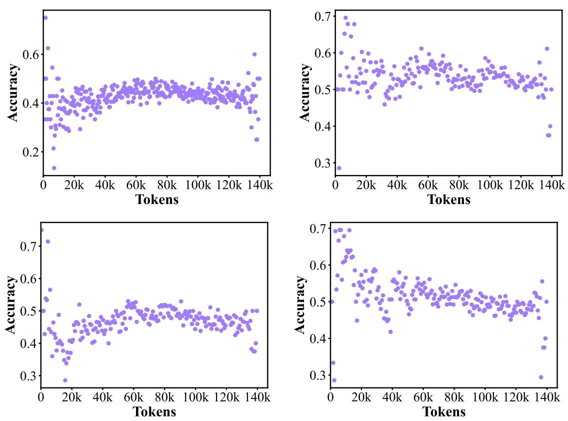


Figure 5: The accuracy of SCMemory(top-left), MemGPT(top-right), GenerativeAgent(bottom-left) and RecentMemory(bottom-right) as the memory token increases.

4.3 Evaluations on Reflective Memory

The test results for reflective memory are shown in Table 4. It can be observed that GenerativeAgent, MemGPT, and MemoryBank performed very well on Sub-dataset 1, but their performance significantly declines on Sub-dataset 2. Only the retrieval-based RetrievalMemory achieved the remaining good results. It is likely due to the limited con-

Table 5: The results of memory mechanisms with different LLMs on our sub-dataset 1. The read time (RT) and write time (WT) are presented in seconds per operation. For the reflective memory, P-Accuracy means the accuracy under the participant scenario, and O-Accuracy refers to the accuracy under the observation scenario.

Method	Factual-Participation			Factual-Observation			Reflective Memory	
	Accuracy	RT	WT	Accuracy	RT	WT	P-Accuracy	O-Accuracy
Qwen2.5-7B-Instruct								
FullMemory	0.647	0.001	<0.001	0.786	<0.001	<0.001	0.733	0.883
RecentMemory	0.639	0.001	<0.001	0.800	<0.001	<0.001	0.700	0.867
RetrievalMemmory	0.692	0.041	0.058	0.883	0.024	0.026	0.692	0.883
GenerativeAgent	0.478	0.045	6.116	0.779	0.031	6.239	0.742	0.883
GPT-4o-mini								
FullMemory	0.736	0.001	<0.001	0.864	<0.001	<0.001	0.783	0.883
RecentMemory	0.697	0.001	<0.001	0.864	<0.001	<0.001	0.758	0.900
RetrievalMemmory	0.633	0.003	0.031	0.857	0.023	0.023	0.767	0.900
GenerativeAgent	0.592	0.107	0.970	0.846	0.030	0.998	0.758	0.900
Meta-Llama-3.1-8B-Instruct								
FullMemory	0.519	0.001	<0.001	0.779	<0.001	<0.001	0.708	0.817
RecentMemory	0.461	0.001	<0.001	0.779	<0.001	<0.001	0.683	0.850
RetrievalMemmory	0.500	0.050	0.062	0.700	0.044	0.049	0.733	0.833
GenerativeAgent	0.430	0.036	6.551	0.725	0.065	12.322	0.725	0.850
glm-4-9b-chat								
FullMemory	0.475	0.001	<0.001	0.775	<0.001	<0.001	0.658	0.850
RecentMemory	0.539	0.001	<0.001	0.746	<0.001	<0.001	0.708	0.850
RetrievalMemmory	0.483	0.032	0.037	0.739	0.025	0.025	0.742	0.800
GenerativeAgent	0.439	0.050	0.165	0.718	0.030	0.111	0.675	0.900

text window of the models or the incorporation of forgetting mechanisms in these memory systems, which leads to the loss of important memories. However, these findings still suggest that well-designed memory mechanisms are capable of effectively capturing reflective memory. How to maintain this ability after prolonged interactions may pose a challenging research problem.

4.4 Evaluations on Memory Capacity

To explore the capacity of the agent's memory mechanism, we test the answering accuracy of each round after the key evidence turns on the observation scenario in Sub-dataset 2(100k). In order to observe the changes in the accuracy of MemGPT and Self-Controlled Memory with the number of tokens increases, we drew Figure 5. From the results, we can observe that both memory mechanisms exhibit a sharp decline, which may be due to the upper limit of memory performance retention capacity for these memory mechanisms in Qwen2.5-7B-Instruct (Yang et al., 2024; Team, 2024).

4.5 Comparison of Different Inference Models

In practical applications of agents, different models may be selected for different memory mechanisms. Therefore, we evaluate the performance of several common models across various memory mechanisms. Specifically, we selected Qwen2.5-7B-Instruct, gpt-4o-mini, Meta-Llama-3.1-8B-Instruct and glm-4b-chat (GLM et al., 2024) for evaluation. The results are shown in Table 5. Under the same context window length, the choice of base model significantly affects the agents' performence. In most cases, GPT-4o-mini performs as the best model compared to others. Although the factual memory capability of Meta-Llama-3.1-8B-Instruct is notably inferior to that of other models, its reflective memory ability is still relatively good. Interestingly, for GenerativeAgent, choosing GPT-4o-mini as the base model results in a significantly higher time consumption compared to other models in our experiments. However, in most cases, the time consumption differences between the three models are not substantial.

5 Conclusion

This paper provides a more comprehensive and scalable dataset for evaluating LLM-based agent's memory mechanisms. It includes a dataset with multi-scenarios (both participation and observation), and multi-level memory content include reflective memory and factual memory. Based on this dataset, we constructed a time-aware evaluation framework that simulates the daily interactions between users and agents with multi-metric include accuracy, recall, capacity and temporal efficiency. We evaluate the performance of seven common memory mechanisms in agents on our benchmark.

Limitations

The dataset proposed in this paper consists of a graph formed by the profiles of users and relevant entities, enabling further exploration of the agent's memory mechanism. Our evaluation method is limited by an assessment of memory for structured data. However, by comparing the construction of relevant entity profiles or the capture of specific attribute information in the agent's memory during user-agent interactions, we can investigate the agent's ability to structure memory. In addition, there are still many areas to explore in reflective memory, such as users' emotional memory.

Ethics Statement

The data used in this article to construct the dataset includes data from publicly available, authorized datasets. All publicly available data are used in accordance with their respective licenses for research purposes. The LLM-generated content may pose risks, including the potential for unintended biases or harmful output. Although we have taken steps to minimize these risks, we encourage users to apply the dataset responsibly to avoid ethical risks.

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (No. 62422215 and No. 62472427), Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, Public Computing Cloud, Renmin University of China, fund for building world-class universities (disciplines) of Renmin University of China. This work is also sponsored by Huawei Innovation Research Programs. We gratefully acknowledge the support from Mindspore[†], CANN(Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

[†] https://www.mindspore.cn

References

Chenxin An, Shansan Gong, Ming Zhong, Mukai Li, Jun Zhang, Lingpeng Kong, and Xipeng Qiu. 2023. L-eval: Instituting standardized evaluation for long context language models. *arXiv preprint arXiv:2307.11088*.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, et al. 2023. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*.

DataGuy and Gordon Amoako. 2022. twitter-news.

Han Ding, Yinheng Li, Junhao Wang, and Hang Chen. 2024. Large language model agent in financial trading: A survey. *arXiv preprint arXiv:2408.06361*.

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 synthesis in question answering. *arXiv preprint arXiv:2402.16288*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

F Maxwell Harper and Joseph A Konstan. 2015. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19.

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

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Bodhisattwa Prasad Majumder, Shuyang Li, Jianmo Ni, and Julian McAuley. 2019. Generating personalized recipes from historical user preferences. *arXiv preprint arXiv:1909.00105*.

Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

Bing Wang, Xinnian Liang, Jian Yang, Hui Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. Enhancing large language model with self-controlled memory framework. *arXiv preprint arXiv:2304.13343*.

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. 2024a. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

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

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, et al. 2024b. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences*, 68(2):121101.

Xinchao Xu, Zhibin Gou, Wenquan Wu, Zheng-Yu Niu, Hua Wu, Haifeng Wang, and Shihang Wang. 2022. Long time no see! open-domain conversation with long-term persona memory. *arXiv preprint arXiv:2203.05797*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2019. Neural personalized response generation as domain adaptation. *World Wide Web*, 22:1427–1446.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024a. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

Zeyu Zhang, Quanyu Dai, Luyu Chen, Zeren Jiang, Rui Li, Jieming Zhu, Xu Chen, Yi Xie, Zhenhua Dong, and Ji-Rong Wen. 2024b. Memsim: A bayesian simulator for evaluating memory of llm-based personal assistants. *arXiv preprint arXiv:2409.20163*.

Zeyu Zhang, Quanyu Dai, Xu Chen, Rui Li, Zhongyang Li, and Zhenhua Dong. 2025. Memengine: A unified and modular library for developing advanced memory of llm-based agents. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 821–824.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

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

A Case Studies

A.1 User Relation Graph Example

In this section, we present examples of the components of our dataset, including the user graph and test cases. For the user graph, we show profile examples of the user itself, related individuals, events, items, and places.

UserProfile: "gender": "Male", "relationship": "self", "name": "James Smith", "age": "30 years old", "height": "164 cm", "birthday": "08.13", "hometown": "San Francisco, CA", "work location": "Boston, MA", "education": "Associate Degree", "occupation": "Police Officer", "position": "Community Policing Officer", "company name": "Boston Law Enforcement Agency", "hobby": "Climbing", "character": "Friendly", "contact number": "4150430511", "email address": "james.smith@bostonlawenforcement.gov", "ssn": "914610199408130162", "passport number": "PUP4822676", "bank account": "6222022865544246", "driver license": "914EAPRDV5F", "highlevel preference": ["movie genre preference": ("Comedy", "Romance", "Action", "Drama", "Thriller"), "taste preference": "Umami and Sweet", "book preference": "Humor"].

RelativeRoleProfile."gender":"Male", "relationship":"Brother","name": "Ethan Cooper","age": "28 years old","height": "165cm","birthday": "01/20","hometown": "Boston, MA", "work location": "Los Angeles, CA", "education": "Associate Degree", "occupation": "Electrician", "position": "Electrical Maintenance Technician", "company name": "SparkLight Electric Services", "hobby": "Running", "character": "Thoughtful", "contact number": "20103787263","email address": "ethan.cooper@sparklightelectric.com"

ColleagueRoleProfile. "gender": "Male", "relationship": "boss", "name": "Nolan Hayes", "age": "39 years old","height": "170cm","birthday": "03/24","hometown": "Philadelphia, PA","work location": "Boston, MA","education": "Associate Degree","occupation": "Police Officer","position": "Police Sergeant","company name": "Boston Law Enforcement Agency", "hobby": "Attending Concerts","character": "Empathetic", "contact number": "30503926075","email address": "nolan.hayes@bostonlawagency.gov"

WorkEventProfile. "event type": "Company Team Building", "main content": "Community engagement workshop for team bonding.", "location": "Los Angeles, CA", "time": "the week after next Sat 9:00 AM", "event name": "Team Connect", "scale": "one hundred people", "duration": "eight day"

RestEventProfile."event type": "Community Fair", "main content": "Join us at the Community Fair for climbing challenges, equipment demos, and safety workshops! Engage with fellow enthusiasts, explore local climbing spots, and enjoy inspiring talks from seasoned climbers. Perfect for all skill levels and climbing enthusiasts!", "location": "Miami, FL", "time": "2024-10-12 19:00", "event name": "Climb Fest", "scale": "nine hundred people", duration": "three day"

ItemProfile: "relationship": "Own", "item type": "Laundry Detergent", "item name": "Arm & Hammer Liquid Detergent", "item review": "As a police officer, I'm always on the go, and I need products that can keep up with my busy lifestyle. I've been using Arm & Hammer Liquid Detergent for a while now, and I have to say, it's been a game changer for me. Not only does it tackle tough stains from my uniforms and gear with ease, but it also leaves my clothes smelling fresh and clean. The added baking soda really helps to neutralize odors, which is a must when you're working in various environments. Plus, I appreciate that it's available in eco-friendly options, making it easier to care for the planet while looking after my laundry. Definitely a solid choice for anyone looking for effective and reliable detergent!"

PlaceProfile. Place example: "relationship": "Visited","place type": "Mall", "place name": "The Grove", "place review": "I recently visited The Grove and I have to say, it was a really refreshing experience. The vibe there is incredibly friendly and welcoming, just like the community I strive to serve as a police officer. The shops and restaurants offer a great variety, and I especially enjoyed grabbing a bite at one of the local eateries. The layout is easy to navigate, making it a perfect spot to relax and enjoy some fresh air.As someone who loves climbing, I appreciated the green spaces where you can unwind and enjoy nature. It's a fantastic place to spend time with family or friends. The only downside I found was that it got a bit crowded during peak hours, but that's to be expected in such a popular location. Overall, I'd highly recommend The Grove to anyone looking for a fun and friendly outing!"

Table 6: Overview of Factual Memory questions.

Types	Descriptions
Single-hop	Rely on one message to answer the question directly.
Multi-hop	Require multiple messages to answer the question jointly.
Comparative	Compare two entities on a shared attribute with multiple messages.
Aggregative	Aggregate messages about more than two entities on a common attribute.
Post-processing	Involve extra reasoning steps to answer with multiple messages.
Knoewledge-updating	The basis for answering questions will be updated over time with different messages.
Single-session-assistant	Rely on a single message from the assistant to directly answer the question.
Multi-session-assistant	Require Multiple messages from the assistant to collectively answer the question.

Table 7: Overview of Reflective Memory questions.

Types	Descriptions
Preference	Rely on multiple messages to actively express the user's lower-level preferences.
Emotion	Rely on multiple consecutive messages within a specific time to express the user's emotional state.

A.2 FM-RM Directionary Example

When creating the correspondence between factual memory attribute and reflective memory attribute, we simultaneously created a dictionary mapping the two. Below, we provide examples from each category of reflective memory.

Movie. "Action": ["Star Wars (1977)","Godfather, The (1972)","Raiders of the Lost Ark (1981)","Titanic (1997)","Empire Strikes Back, The (1980)","Boot, Das (1981)","Godfather: Part II, The (1974)","African Queen, The (1951)","Princess Bride, The (1987)","Braveheart (1995)", "Glory (1989)", "Fugitive, The (1993)","Alien (1979)","Return of the Jedi (1983)","Terminator 2: Judgment Day (1991)", "Butch Cassidy and the Sundance Kid (1969)","Aliens (1986)","Magnificent Seven, The (1954)","Terminator, The (1984)","Apollo 13 (1995)","Indiana Jones and the Last Crusade (1989)","Die Hard (1988)","Hunt for Red October, The (1990)","Good, The Bad and The Ugly, The (1966)","Blues Brothers, The (1980)","Ben-Hur (1959)", "Cyrano de Bergerac (1990)", "Star Trek: The Wrath of Khan (1982)","In the Line of Fire (1993)", "Adventures of Robin Hood, The (1938)","Jaws (1975)","Face/Off (1997)","Men in Black (1997)","Diva (1981)","Jurassic Park (1993)","Rock, The (1996)","Full Metal Jacket (1987)", "Perfect World, A (1993)","Star Trek: First Contact (1996)","Speed (1994)","Air Force One (1997)", "Crying Game, The (1992)", "True Romance (1993)","Abyss, The (1989)","Clear and Present Danger (1994)","Heat (1995)","True Lies (1994)","Get Shorty (1995)","Last of the Mohicans, The (1992)","Supercop (1992)"]

Food. "Sweet": ["Candy", "Honey", "Fruit", "Maple Syrup Pancakes", "Baklava", "Chocolate Cake", "Custard", "Jelly", "Pecan Pie", "Apple Pie", "Brownies", "Banana Bread", "Donuts", "Rice Krispies"]

Book."Health & Fitness": ["What to Expect When You're Expecting (Revised Edition)", "Make the Connection: Ten Steps to a Better Body and a Better Life", "The South Beach Diet: The Delicious, Doctor-Designed, Foolproof Plan for Fast and Healthy Weight Loss", "Dr. Atkins' New Diet Revolution", "Dr. Atkins' New Diet Revolution", "Prescription for Nutritional Healing: A Practical A-Z Reference to Drug-Free Remedies Using Vitamins, Minerals, Herbs & Food Supplements", "8 Weeks to Optimum Health", "Body for Life: 12 Weeks to Mental and Physical Strength", "Your Pregnancy: Week by Week (Your Pregnancy Series)", "Fat Land: How Americans Became the Fattest People in the World"]

A.3 Question Description

In Tab 6, we provide an explanation for each type of Factual Memory questions. In Tab 7, we provide an explanation for each type of Reflective Memory questions.

A.4 Participation Example

In this section, we provide more detailed examples from the participation scenarios in our dataset. To make the presentation clearer, we have only listed the key evidence dialogue rounds necessary for answering the questions and omitted any unnecessary information.

The detailed examples are as follows:

Single Hop

User: My niece runs a company called TechInnovate Systems LLC.

Assistant:...

Question: What is the name of my niece's company?

Answer: TechInnovate Systems LLC

Multi Hop

User: My sister's name is Amelia Brooks, and she's always been a real standout with her creativity and charm.

Assistant:...

User: My sister's birthday is coming up on December 11th.

Assistant:...

Question: What is the birthday of someone named Amelia Brooks?

Answer: 12/11

Knowledge Updating

User:The Policing Forum lasts for four days, and I can't wait to see what they have in store during that time. Assistant:

User: I just realized I need to correct myself—Policing Forum only lasts for one day. Assistant:...

Question: How long does the Policing Forum last?

Answer: one day

Post Processing

User: My subordinate has this cool hobby of collecting antiques.

Assistant:...

User: My subordinate's email address is maya.carter@bostonlaw.gov.

Assistant:...

Question: What are the main interests and hobbies of the individual with the email address maya.carter@bostonlaw.gov?

Answer: Gather historical items and appreciate their value

Single session assistant

user: I'm looking for a great movie to watch, aside from the ones I've mentioned earlier.

assistant: I'm all about Return of the Jedi; it's such an iconic film with those thrilling space battles and unforgettable characters that make it a must-see!

user: I'm looking for a fantastic movie to watch, aside from the ones I've mentioned earlier.

assistant: I recommend the Jurassic Park from 1993; it's such an incredible film with its groundbreaking special effects and thrilling adventure that really grabs my attention!

Question: What movies have you recommended to me before?

Answer: ["Return of the Jedi (1983)","Jurassic Park (1993)"]

Comparative

User: My niece, Clara Jennings, is 28 years old.

Assistant: ...

User: My Uncle Landon Pierce is 27 years old.

Assistant: ...

Question: Who is older, Clara Jennings or Landon Pierce?

Answer: Clara Jennings

Aggregative

User: My brother Ethan Cooper hails from Boston, MA

Assistant:...

User: My sister, Amelia Brooks, is from Washington, DC.

Assistant:...

User: My niece, Clara Jennings, is from Jacksonville, FL.

Assistant:...

User: My uncle, Landon Pierce, hails from Philadelphia, PA.

Assistant:...

User: Sophie Turner, my aunt, is from Chicago, IL.

Assistant:...

User: My boss, Nolan Hayes, is from Philadelphia, PA.

Assistant:...

User: Maya Carter, who works under me, is from Los Angeles, CA.

Assistant:...

User:My coworker, Maxwell Turner, is from Indianapolis, IN.

Assistant:...

Question: How many people live in Philadelphia, PA?

Answer: 2 people

Multi session assistant

User: I'm looking for a great movie to watch, excluding the ones mentioned earlier.

Assistant: I definitely think Alien (1979) is a must-see; it's an intense blend of science fiction and horror that really keeps you on the edge of your seat!

User: I'm looking for a great movie to watch, aside from the ones I mentioned earlier.

Assistant: "I'm all about The Fugitive from 1993; it's a thrilling ride that really keeps you on the edge of your seat!

User: I'm looking for a delicious dish to try, aside from the ones mentioned earlier.

Assistant: I'm a big fan of Apple Pie; it's such a delightful treat that everyone should definitely give it a try!

User: I'm looking for a good book to read, excluding the ones mentioned earlier.

Assistant: I really think Naked is a fantastic read; it's definitely worth checking out!

User: I'm looking for a good book to read, aside from the ones mentioned earlier.

Assistant: I really think everyone should check out Politically Correct Bedtime Stories: Modern Tales for Our Life and Times; it offers a hilarious twist on classic tales that resonate with today's issues.

Question: What movies, books and dishes have you recommended to me?

Answer: ["Alien (1979)","Fugitive, The (1993)","Apple Pie","Naked","Politically Correct Bedtime Stories: Modern Tales for Our Life and Times"]

Multi session highlevel preference

User: I'm a big fan of prosciutto and melon; there's just something about that sweet and savory combination that I can't resist!

Assistant:...

User: I'm a fan of Prosciutto and Melon, but I also really enjoy Salted Maple Ice Cream; it's a unique treat that hits the spot!

Assistant: ...

User: I really love Salted Maple Ice Cream, and Pecan Praline is another favorite of mine!

Assistant: ...

User: I really enjoy Pecan Praline, but I also have a soft spot for Salted Butter Toffee; there's something so satisfying about that perfect blend of sweet and salty.

Assistant: ...

User: I really love Salted Peanut Butter Cookies, just like I enjoy Salted Butter Toffee.

Assistant:...

Question: According to the dishes I mentioned, Which flavor I might prefer?

Answer: Sweet and Salty

A.5 Observation Example

The only difference between the data for participation and the data here is the absence of responses from "assistant," so specific examples are not provided here.

Table 8: The detail statistics of our Participation dataset. RM indicates Reflective Memory. FM indicates Factual Memory. The types we have include ssh(sigle-hop), mh(multi-hop), comp(comparative), agg(aggregative), pp(post-processing), ku(knowledge-update), ssa(signle-session-assistant), msa(multi-session-assistant).

Data Type	# Session	# Question	# Trajectory
RM-Pr	3.0k	3.0k	3.0k
RM-Em	0.5k	0.5k	0.5k
FM-sh	8k	8k	1k
FM-mh	8k	8k	1k
FM-comp	8k	4k	1k
FM-agg	8k	1k	1k
FM-ku	8k	8k	1k
FM-pp	8k	8k	1k
FM-ssa	1.5k	1.5k	1.5k
FM-msa	1.5k	0.5k	0.5k

Table 9: The detail statistics of our Participation dataset. RM indicates Reflective Memory. FM indicates Factual Memory. The types we have include sh(sigle-hop), mh(multi-hop), comp(comparative), agg(aggregative), pp(post-processing), ku(knowledge-update).

Data Type	# Session	# Question	# Trajectory
RM-Pr	1.5k	1.5k	1.5k
RM-Em	0.5k	0.5k	0.5k
FM-sh	1.5k	1.5k	1.5k
FM-mh	1.5k	1.5k	1.5k
FM-comp	1.5k	1.5k	1.5k
FM-agg	1.5k	1.5k	1.5k
FM-pp	1.5k	1.5k	1.5k
FM-ku	1k	1k	1k

B Detail Data Statics

In Tab 8, we provide the detail statistics of Participation dataset. In Tab 9, we provide the detail statistics of Observation dataset.

C Data Creation Prompt

C.1 Profile Prompt

Flavour Reflective Memory Attribute Please choose user's taste from [Tastes] according to the dishes he likes below. [Dishes]:{Dishes} [Tastes]: ["Sweet", "Sour", "Spicy", "Salty", "Umami", "Bitter", "Sweet and Salty", "Sweet and Sour", "Salty and Umami", "Sour and Spicy", "Sweet, Salty, and Spicy", "Sour and Salty", "Sour, Sweet, and Salty", "Salty, Umami, and Spicy", "Numbing and Spicy", "Creamy and Sweet", "Umami and Sweet", "Bitter and Sweet", "Astringent", "Numbing", "Rich and Fatty", "Cool", "Warm and Spicy"] example output:{{'taste': 'sweet'}}

C.2 Self-dialogue Prompt

Role Dialogue Generation Prompt. Please generate a {round_length}-round interactive conversion between the user and assistant, with a total of {sentence_length} sentences. The dialogue's main content should based on the given information about the user's {entity}. Ensure that no information beyond what is provided is introduced in the dialogue. **Note that the user cannot ask the assistant for information because the assistant does not know the information.** Note that the assistant is the user's personal assistant, so it should only respond passively to the user's dialogue, but it can reply with varied content. Please return the conversation in a JSON list format as shown in the example, ensuring that the result can be directly parsed by json.loads. **Every json must includes both user and assistant with their words! Every json's format is {{"user": user's words, "assistant": assistant's words}}** #[Information]: {information}

#example: [{{"user": "I wanted to talk to you about my cousin, Ethan Parker. He's 39 years old.", "assistant": "Certainly! Ethan is 39. Is there something specific you'd like to discuss about him?"}}, {{"user": "Well, he's actually pretty short for a guy, only 162 cm. He always jokes about it, though.", "assistant": "It sounds like Ethan has a good sense of humor about his height. That's always a great quality!"}}, {{"user": "Yes, he really does. It's one of the things that makes him so fun to be around.", "assistant": "I can imagine! Having a cousin with a lighthearted attitude must make family gatherings enjoyable."}}, {{"user": "Definitely. And since we're so close in age, it feels like he's more of a friend than just family.", "assistant": "That sounds wonderful. It must be special to have that kind of bond with your cousin."}} {{"user": "It really is. We've shared a lot of memories growing up together.", "assistant": "Those shared memories must make your relationship even stronger. It sounds like Ethan has been a big part of your life."}}]

Event Dialogue Generation Prompt. Please generate a {round_length}-round interactive conversion between the user and assistant, with a total of {sentence_length} sentences. The dialogue's main content should based on the given information about the {event_name}. Ensure that no information beyond what is provided is introduced in the dialogue. Note that the assistant is the user's personal assistant, so it should only respond passively to the user's dialogue, but it can reply with varied content. Note that the user cannot ask the assistant for information because the assistant does not know the information. You can start with user saying I'm going to attend {event_name} Please return the conversation in a JSON list format as shown in the example, ensuring that the result can be directly parsed by json.loads. #[Information]: {information}

#example: [{{"user": "I wanted to talk to you about my cousin, Ethan Parker. He's 39 years old.", "assistant": "Certainly! Ethan is 39. Is there something specific you'd like to discuss about him?"}}, {{"user": "Well, he's actually pretty short for a guy, only 162 cm. He always jokes about it, though.", "assistant": "It sounds like Ethan has a good sense of humor about his height. That's always a great quality!"}}, {{"user": "Yes, he really does. It's one of the things that makes him so fun to be around.", "assistant": "I can imagine! Having a cousin with a lighthearted attitude must make family gatherings enjoyable."}}, {{"user": "Definitely. And since we're so close in age, it feels like he's more of a friend than just family.", "assistant": "That sounds wonderful. It must be special to have that kind of bond with your cousin."}} {{"user": "It really is. We've shared a lot of memories growing up together.", "assistant": "Those shared memories must make your relationship even stronger. It sounds like Ethan has been a big part of your life."}}]

C.3 Observation Prompt

Role Message Prompt. [User Message]: {message} Please rewrite the above user message into a colloquial declarative sentence. Ensure it is smooth and free of grammatical errors, without changing the original information. Only output the rewritten user message, without including the original message. Do not output any other description. Output example: Lucas Grant, who is my boss, has a Master's degree. """

Event Message Prompt. [User Message]: {message} Please rewrite the above user message into a colloquial declarative sentence. Ensure it is smooth and free of grammatical errors, without changing the original information, and avoid using 'you'. Don't forget use I , me or my Only output the rewritten user message, without including the original message. Do not output any other description. Output example: Climb Fest draws a crowd of around nine hundred people.

D Result Details

D.1 Reflective Result

In Tab 10, we show the detailed results of our 10k-Reflective memory dataset.

D.2 Facutal Result

In Tab 11, we show the detailed results of our 10k-Factual-Partipation memory dataset. In Tab 12, we show the detailed results of our 10k-Factual-Observation memory dataset.

Table 10: The results of different mechanisms on different types of our 10k-Reflective memory dataset.

Method	Participation-Accuracy		Observation-Accuracy	
	preference	emotion	preference	emotion
FullMemory	0.733	0.593	0.883	0.630
RecentMemory	0.700	0.481	0.867	0.556
RetrievalMemmory	0.692	0.556	0.883	0.593
GenerativeAgent	0.742	0.412	0.883	0.676
MemoryBank	0.692	0.296	0.900	0.481
MemGPT	0.733	0.471	0.883	0.556
SCMemory	0.542	0.294	0.783	0.333

Table 11: The results of different mechanisms on different types of our 10k-Factual-Participation dataset. Including sh(sigle-hop), mh(multi-hop), comp(comparative), agg(aggregative), pp(post-processing), ku(knowledge-update), ssa(signle-session-assistant), msa(multi-session-assistant).

Method	Participation-Accuracy							
	sh	mh	comp	agg	pp	ku	ssa	msa
FullMemory	0.825	0.8	0.55	0.275	0.625	0.75	0.7	0.55
RecentMemory	0.85	0.75	0.425	0.45	0.65	0.725	0.717	0.5
RetrievalMemmory	0.875	0.775	0.55	0.275	0.475	0.675	0.4	0.3
GenerativeAgent	0.75	0.675	0.3	0.35	0.525	0.525	0.267	0.55
MemoryBank	0.575	0.7	0.25	0.25	0.475	0.55	0.417	0.4
MemGPT	0.625	0.625	0.275	0.225	0.45	0.625	0.367	0.45
SCMemory	0.575	0.475	0.05	0.275	0.525	0.475	0.217	0.1

Table 12: The results of different mechanisms on different types of our 10k-Factual-Observation memory dataset. Including sh(sigle-hop), mh(multi-hop), comp(comparative), agg(aggregative), pp(post-processing), ku(knowledge-update).

Method	Observation-Accuracy					
	sh	mh	comp	agg	pp	ku
FullMemory	0.92	0.92	0.667	0.233	0.82	0.6
RecentMemory	0.92	0.92	0.667	0.367	0.82	0.65
RetrievalMemmory	0.92	0.92	0.633	0.367	0.78	0.45
GenerativeAgent	0.88	0.94	0.7	0.3	0.82	0.4
MemoryBank	0.8	0.78	0.633	0.233	0.800	0.6
MemGPT	0.94	0.92	0.667	0.233	0.82	0.600
SCMemory	0.46	0.68	0.133	0.133	0.78	0.65