In Prospect and Retrospect: Reflective Memory Management for Long-term Personalized Dialogue Agents

Zhen Tan[1] [*], Jun Yan[2], I-Hung Hsu[2], Rujun Han[2], Zifeng Wang[2], Long T. Le[2], Yiwen Song[2], Yanfei Chen[2], Hamid Palangi[2], George Lee[2], Anand Iyer[3], Tianlong Chen[4], Huan Liu[1], Chen-Yu Lee[2] and Tomas Pfister[2]

[1]Arizona State University, [2]Google Cloud AI Research, [3]Google Cloud AI, [4]UNC Chapel Hill

Large Language Models (LLMs) have made significant progress in open-ended dialogue, yet their inability to retain and retrieve relevant information from long-term interactions limits their effectiveness in applications requiring sustained personalization. External memory mechanisms have been proposed to address this limitation, enabling LLMs to maintain conversational continuity. However, existing approaches struggle with two key challenges. First, rigid memory granularity fails to capture the natural semantic structure of conversations, leading to fragmented and incomplete representations. Second, fixed retrieval mechanisms cannot adapt to diverse dialogue contexts and user interaction patterns. In this work, we propose Reflective Memory Management (RMM), a novel mechanism for long-term dialogue agents, integrating forward- and backward-looking reflections: (1) Prospective Reflection, which dynamically summarizes interactions across granularities—utterances, turns, and sessions—into a personalized memory bank for effective future retrieval, and (2) Retrospective Reflection, which iteratively refines the retrieval in an online reinforcement learning (RL) manner based on LLMs' cited evidence. Experiments show that RMM demonstrates consistent improvement across various metrics and benchmarks. For example, RMM shows more than 10% accuracy improvement over the baseline without memory management on the `LongMemEval` dataset.

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in engaging in open-ended dialogue (Lee et al., 2023; Mendonça et al., 2024), yet their inherent statelessness poses a significant challenge for maintaining coherent, personalized conversations over time (Chen et al., 2024; Li et al., 2024b; Tseng et al., 2024), which are crucial across various real-world applications (*e.g.*, customer service (Kolasani, 2023), virtual assistants (Guan et al., 2024), and education platforms (Wen et al., 2024; Zhang et al., 2024d)). As illustrated in Figure 1, effective personalization requires not only understanding the immediate context but also recalling relevant information from the user's previous interactions (Dong et al., 2024; Whittaker et al., 2002; Williams and Hollan, 1981). The limitations with current LLMs to naturally retain and recall information from past
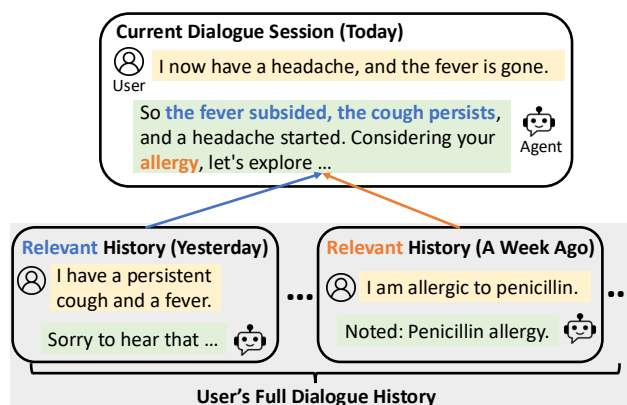

Figure 1 | An illustration of a personalized healthcare agent. Key information about a user's allergy and previous symptoms mentioned in the past sessions is needed to provide a more informed response in the current session.

interactions beyond their context windows sparked the development of external memory mechanisms for LLMs (Li et al., 2025; Ong et al., 2025; Zhang et al., 2024c). These memory systems serve as crucial components in personalized dialogue agents, enabling them to maintain consistent personality traits, remember user preferences, and build upon previous interactions.

While external memory mechanisms represent a significant step towards enabling persistent dialogue, current approaches suffer from two critical limitations. Firstly, existing systems digest information at a pre-defined granularity, such as turn, session, or time interval boundaries, which may not align with the inherent semantic structure of the conversation (e.g., topic shifts). This rigid approach can lead to fragmented or incomplete memory representations, hindering the LLM's ability to retrieve, utilize, and update relevant information effectively (Pan et al., 2025; Wu et al., 2025). Secondly, these systems rely on fixed retrievers (Li et al., 2025; Zhong et al., 2024), which struggle to adapt to the diverse retrieval demands of varying dialogue domains and individual user interaction patterns. Moreover, the expense associated with collecting labeled data for training personalized retrievers presents a substantial barrier to widespread adoption and scalability.

To address these limitations, we propose a novel **Reflective Memory Management** (**RMM**) mechanism to provide a more adaptable and granular approach to long-term dialogue memory. Our framework incorporates two key innovations. **Prospective Reflection** tackles the issue of fixed granularity by summarizing dialogue histories into decomposed topics, effectively integrating fragmented conversational segments into cohesive memory structures. This approach optimizes memory organization for future retrieval, allowing the LLM to access relevant information more effectively regardless of the original turn or session boundaries. Complementing this, **Retrospective Reflection** addresses the challenge of fixed retrievers by leveraging unsupervised attribution signals generated during the LLM's response generation to reflect on past retrieval. This allows for online refinement of the retriever as the conversation progresses, enabling the system to adapt to diverse dialogue domains and individual user interaction patterns without the need for costly labeled data.

By integrating these two reflective mechanisms, our approach enables LLMs to maintain a more nuanced and adaptable memory, leading to more coherent, personalized, and engaging dialogues. Experiments on `MSC` and `LongMemEval` benchmarks show that RMM achieves more than 5% improvement over the strongest baseline across memory retrieval and response generation metrics.

Our contributions are as follows: (1) We propose RMM as a novel memory management mechanism that employs topic-based memory management optimized for future retrieval and leverages attribution signal to reflect on past retrieval for unsupervised online retrieval refinement. (2) We conduct extensive experiments on two long-term personalized dialogue benchmarks to demonstrate the effectiveness of RMM over strong baselines. (3) We perform detailed analysis on the impact of various design choices to pinpoint the limitations of existing memory management mechanisms with fixed granularity and retrievers, shedding light on the room for future improvement.

2. Related Work

Long-term Conversations for LLMs. LLMs have demonstrated the ability to engage in extended, coherent dialogues, yet maintaining context and consistency over long-term interactions remains a challenge. Maharana et al. (2024) introduced the LoCoMo dataset to assess LLMs' performance in sustained dialogues, showing their struggles with long-range temporal and causal understanding. Existing solutions can be broadly categorized into two approaches: (1) Architectural modifications, such as enhancing attention mechanisms (Liu et al., 2024a; Zhang et al., 2024a), optimizing KV caches (LI et al., 2025; Liu et al., 2025), and refining position embeddings (Zhao et al., 2024; Zheng et al., 2024). These methods require white-box access to model internals, making them infeasible for

proprietary or API-based LLMs. (2) Summarization-based methods, which condense long contexts into structured events or topics for direct conditioning or retrieval (Jiang et al., 2025; Li et al., 2024a; Lu et al., 2023; Wang et al., 2023). RMM falls into this category but explicitly addresses the issue of fragmented topics arising from fixed granularity and incorporates retrospective reflection to refine the retrieval process, encouraging more coherent and contextual responses.

Memory-based Personalized Dialogue Agents. The development of memory-based personalized dialogue agents has further enhanced long-term interactions by enabling systems to retain and utilize information from past conversations (Bae et al., 2022). Traditional methods Walker et al. (1997); Weizenbaum (1966) laid the groundwork for understanding how systems can model user preferences, intentions, and behaviors across sessions, often using handcrafted rules, heuristics, or symbolic representations. Early approaches, such as CoMemNN (Pei et al., 2021), introduce mechanisms to incrementally enrich user profiles during dialogues. However, collecting substantial annotations for training a personalized system for long-term use is hard (Tseng et al., 2024). Recent advancements focus on integrating LLMs with memory modules (Chhikara et al., 2025; Packer et al., 2024; Rasmussen et al., 2025; Wang et al., 2025; Xu et al., 2025). For instance, the LD-Agent framework (Li et al., 2025) employs long-, short-term memory banks to manage conversational history for retrieval. MemoryBank (Zhong et al., 2024) incorporates a memory updating mechanism inspired by the Ebbinghaus Forgetting Curve, enabling models to retrieve relevant memories considering recency. Theanine (Ong et al., 2025) introduces timeline-based retrieval and utilizes an additional LLM for refinement. These methods typically deploy fixed retrievers with a pre-defined granularity. In contrast, the proposed RMM approach facilitates adaptive retrieval with a revised retrieval granularity.

3. Problem Formulation

We consider the task of building a personalized dialogue agent in a **multi-session** conversational setting. In this setting, an agent interacts with a user across multiple distinct sessions. A *session* represents a distinct interaction period, often delimited by user inactivity, explicit user confirmation of conversation completion, or the initiation of a new dialogue thread. Within each session, the conversation unfolds as a sequence of turns, where a *turn* consists of a user query and the agent's corresponding response. The agent is equipped with an external memory, serving as the sole repository for information gathered from previous sessions. The agent's objective is to generate contextually relevant and personalized responses to user queries, leveraging both the immediate conversational context within the current session and the relevant information retrieved from the memory.

This task presents two key challenges: first, the agent must proactively identify and store salient information from each session, anticipating *future* retrieval needs. Second, the agent must accurately retrieve relevant *past* information from the memory, as incorporating irrelevant context can distract the LLM and degrade response quality (Liu et al., 2024b; Shi et al., 2023). Effectively managing this balance between comprehensive storage and precise retrieval is critical for achieving personalized and coherent multi-session dialogues.

4. Framework Overview

To tackle the challenges, we introduce Reflective Memory Management (RMM), a novel framework that integrates two mechanisms. Prospective Reflection proactively decomposes dialogue history into topic-based memory representations, optimizing for future retrieval, while Retrospective Reflection dynamically refines the retrieval mechanism through online feedback signals generated during response generation. They together improve the quality of the retrieved memories, contributing to effective personalization.

Our framework comprises four key components. The **memory bank** stores dialogue history as a collection of memory entries, each represented as a pair (topic summary, raw dialogue), where the "topic summary" serves as the search key for retrieving the conversational segment. The **retriever** identifies relevant memories based on the current user query. To enable lightweight adaptation of the retrieval process, we incorporate a **reranker**, which refines the retriever's initial output by prioritizing the most pertinent memories. Finally, an **LLM** synthesizes the relevant memories with the current context to produce a personalized response. Crucially, the LLM also provides feedback signals based on its utilization of retrieved memories, which are used to refine the reranker through Retrospective Reflection. Our complete workflow is detailed in Algorithm 1.

Algorithm 1 Reflective Memory Management (RMM) for Dialogue Agents

---

Input: query q, past messages in current session S, memory bank B, retriever f_θ, reranker g_ϕ, LLM
Output: response a, updated S, g_ϕ, B
1: **Retrieve:** $\mathcal{M}_K \leftarrow f_\theta(q, B)$
2: **Rerank:** $\mathcal{M}_M \leftarrow g_\phi(q, \mathcal{M}_K)$, where $\mathcal{M}_M = \{m_i\}_{i=1}^M$
3: // Retrospective Reflection
4: **Generate:** $a, R_M \leftarrow \text{LLM}(q, S, \mathcal{M}_M)$ where $R_M = \{r_i\}_{i=1}^M$
5: $g_\phi \leftarrow \textbf{RL\_Update}(g_\phi, R_M)$
6: $S.append((q, a))$
7: // Prospective Reflection
8: **if** session S ends **then**
9:     $\mathcal{M} \leftarrow \textbf{ExtractMemory}(S)$
10:     **for** $m \in \mathcal{M}$ **do**
11:         $B \leftarrow \textbf{UpdateMemory}(B, m)$
12:     **end for**
13:     $S \leftarrow []$
14: **end if**

---

5. Prospective Reflection: Topic-Based Memory Organization

Traditional memory management systems often rely on fixed boundaries, such as session or turn delimiters, to structure dialogue history. However, these pre-defined boundaries may not align with the underlying semantic units of conversation. As a result, critical information may be fragmented across multiple memory entries, hindering effective retrieval. To address this, we introduce Prospective Reflection, a mechanism for organizing memory based on coherent topics, enabling more granular and semantically relevant future retrieval. Here "topic" refers to a semantically coherent unit of discussion that may span across one or multiple turns in a session. Each topic is associated with the raw dialogue segment(s) in which it was discussed. These topics can range from fine-grained user intents (e.g., asking about vegan recipes) to broader themes (e.g., travel planning). As illustrated in Figure 2, this process occurs at the conclusion of each session and consists of two key steps: memory extraction and memory update.
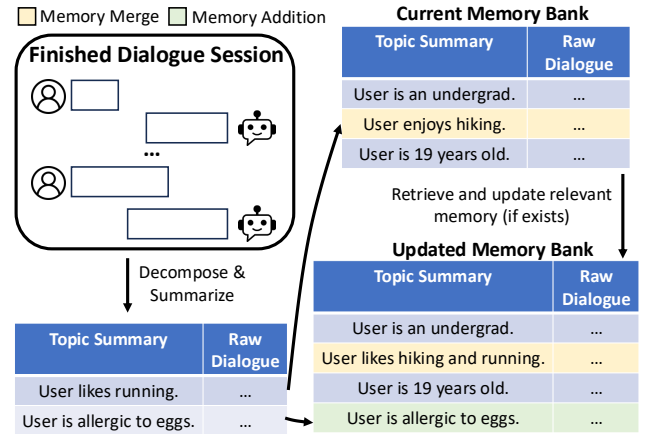

Figure 2 | Illustration of *Prospective Reflection*. After each session, the agent decomposes and summarizes the session into specific topics. These newly generated memories are compared with existing memories in the memory bank. Relevant memories are merged , while others are directly added . Prospective reflection ensures efficient organization of personal knowledge for future retrieval.

First, **memory extraction** is achieved by using an LLM (prompt in Appendix D.1.1) to extract dialogue snippets from the session with their corresponding summaries based on the distinct mentioned topics. Second, **memory update** involves integrating the extracted topic-based memories into the memory bank. Specifically, for each extracted memory, we retrieve the Top-K most semantically similar

memories already present in the memory bank. Subsequently, an LLM (prompt in Appendix D.1.2) determines whether the extracted memory should be directly **added** into the memory bank (e.g., when the extracted memory discusses a new topic) or **merged** with an existing memory into an updated one (e.g., when the extracted memory provides updated information to a previously discussed topic).

Through Prospective Reflection, the memory bank maintains a coherent and consolidated representation of the evolving dialogue history, organized around meaningful topic structures.

6. Retrospective Reflection: Retrieval Refinement via LLM Attribution

6.1. Reranker Design

While an off-the-shelf retriever can identify semantically-relevant memories, its performance can degrade across diverse dialogue domains and user interaction patterns. Instead of resorting to computationally expensive fine-tuning of the retriever, which requires extensive labeled data, we introduce a lightweight reranker to refine the retrieved memory list. This reranker allows for efficient adaptation to the nuances of specific dialogue domains and user preferences, enabling the system to dynamically adjust its retrieval strategy.

To be specific, the reranker processes the Top-K memory embeddings retrieved by the retriever, refining their relevance with respect to the user query and selecting the Top-M candidates. The whole process includes the following steps.

Embedding Adaptation. Let $\mathbf{q}$ represent the embedding of the query and $\mathbf{m}_i$ represent the embedding of the i-th memory entry retrieved by retriever. The embeddings are fed into the reranker to be refined via a linear layer with residual connections:

$$\mathbf{q}' = \mathbf{q} + \mathbf{W}_q\mathbf{q}, \quad \mathbf{m}'_i = \mathbf{m}_i + \mathbf{W}_m\mathbf{m}_i, \tag{1}$$

where $\mathbf{W}_q$ and $\mathbf{W}_m$ are linear transformation matrices for the query and memory, respectively.

Stochastic Sampling with Gumbel Trick. The adapted query embedding $\mathbf{q}'$ and memory embeddings $\mathbf{m}'_i$ are adopted to compute relevance scores via dot product: $s_i = \mathbf{q}'^{\top}\mathbf{m}'_i$. To select memory entries based on relevance scores, we employ the Gumbel Trick (Gumbel, 1954), which enables stochastic sampling from a discrete probability distribution while preserving gradients, making it particularly useful in reinforcement learning and differentiable ranking tasks (Jang et al., 2017). We add Gumbel noise g_i (Maddison et al., 2014) to the relevance scores s_i for each memory entry:

$$\tilde{s}_i = s_i + g_i, \quad g_i = -\log(-\log(u_i)), \tag{2}$$

where $u_i \sim \text{Uniform}(0, 1)$. The perturbed scores $\tilde{s}_i$ are then normalized using the softmax function to compute sampling probabilities: $p_i = \frac{\exp(\tilde{s}_i/\tau)}{\sum_{j=1}^{K} \exp(\tilde{s}_j/\tau)}$, where $\tau > 0$ is the temperature parameter controlling the sharpness of the distribution. Lower τ results in more deterministic sampling (approaching the maximum of s_i), while higher τ increases stochasticity, encouraging exploration.

By introducing a reranker, RMM ensures efficient retrieval refinement without modifying the retriever itself, making it adaptable to any pre-trained retrieval model while allowing task-specific optimizations through Reinforcement Learning (RL).

6.2. LLM Attribution as Rewards

Obtaining high-quality user-specific labeled data for refining the retrieval process is prohibitively expensive. To overcome this challenge, we propose leveraging the inherent capabilities of the LLM

generator itself to provide automated feedback on the quality of retrieved memories. Given the user query with context in the current session, and the retrieved memories, we prompt the LLM (prompt in Appendix D.2) to generate both the response and the associated citations to each individual memory in the context (Kenthapadi et al., 2024). This design uses a single LLM call for generating response and LLM attribution, reducing computational overhead. Moreover, the citations are generated conditioned on the response, which has been shown to be more effective compared to prior or post-hoc citations (Buchmann et al., 2024).

Rewards. As shown in Figure 3, each retrieved memory entry receives either a positive or negative reward based on its citation in the generated response. Specifically, we assign a reward of **+1 (Useful)** if the generator cites the memory in the final response, and **−1 (Not Useful)** otherwise. This reward assignment reflects the utility of each memory entry and allows the reranker to learn better retrieval strategies over time, aligning future selections with the generator's actual usage of retrieved evidence. We validate its effectiveness in Section 8.3.

6.3. Reranker Update

The reranker is fine-tuned using the REIN-FORCE algorithm (Williams, 1992) to optimize its relevance predictions based on these binary rewards with the following formulation:

$$\Delta\phi = \eta \cdot (R - b) \cdot \nabla_\phi \log P(\mathcal{M}_M | q, \mathcal{M}_K; \phi), \quad (3)$$

where R is the reward (+1 or −1), b is a baseline value set as a hyperparameter, and ϕ denotes the weights of the reranker.
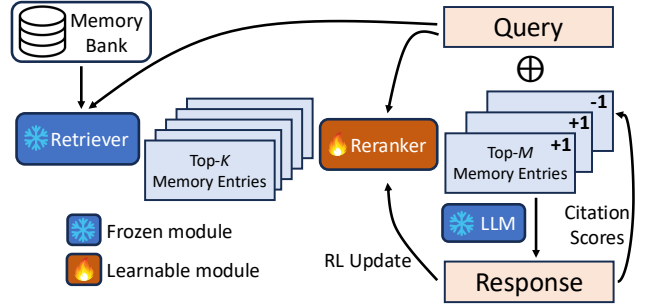

Figure 3 | Illustration of *Retrospective Reflection*. The Retriever fetches Top-K memory entries from the memory bank, which are refined by the learnable *Reranker* to select the Top-M most relevant entries. These entries are passed to the LLM along with the query to generate the final response. The LLM assigns binary *citation scores* (+1 for useful and −1 for not useful) to the retrieved memory entries based on their utility in the response. These scores are used as reward signals to update the reranker via an *RL update*, adapting the selection of relevant memory over time.

7. Experimental Setup

7.1. Implementation Details

In our experiments, we use Gemini-1.5-Flash as the generator and evaluate Gemini-1.5-Pro in Section 8.4. We equip RMM with the following dense retrievers with strong semantic representation capabilities and widespread adoption in personalized dialogue systems (Wu et al., 2025).

- Contriever (`facebook/contriever`) (Izacard et al., 2022): A dense retriever optimized for semantic search leveraging contrastive learning.
- Stella (`dunzhang/stella_en_1.5B_v5`) (Zhang et al., 2024b): A large embedding-based retriever, which is developed based on language models.
- GTE (`Alibaba-NLP/gte-Qwen2-7B-instruct`) (Li et al., 2023): A retriever designed for instruction-following queries, which is trained across a vast, multilingual text corpus spanning diverse domains.

Contriever is used as the default retriever. Following Wu et al. (2025), for experiments without a reranker, the Top-K is 5. Otherwise, the default Top-K is 20 and Top-M is 5. We explore the

impact of retrieval parameters in Appendix 8.7. For LongMemEval, we also consider the results of using an "Oracle" retriver which retrieves the ground-truth turns annotated in the dataset with the necessary personal knowledge to respond to a question. More implementation and training details are elaborated in Appendix A.

7.2. Datasets and Evaluation Metrics

We experiment on two publicly available benchmark datasets commonly used for personalized dialogue evaluation: MSC (Xu et al., 2022) and LongMemEval (Wu et al., 2025). Additional details about datasets can be found in Appendix B.

For MSC, the evaluation measures if the generated response matches the human-provided ground truth. We follow Li et al. (2025) to use **METEOR** (Banerjee and Lavie, 2005) for measuring lexical similarity and **BERTScore** (Zhang et al., 2020) for measuring semantic similarity. We also provide LLM judge results in Appendix 8.8.

For LongMemEval, we follow the original paper to use **Recall@K** to evaluate the model's ability to retrieve relevant information for the query from conversation histories and use an LLM judge to measure the **Accuracy** of the generated answer by comparing it to the human-provided ground truth using Gemini-1.5-Pro. The prompt is presented in Appendix D.3.

7.3. Compared Methods

To benchmark the performance of RMM, we compare it against the following baselines which represent different strategies for managing and retrieving long-term conversational memories, allowing for a comprehensive comparison with RMM.

- **No History**: No history session is used.
- **Long Context**: This method directly incorporate as much conversation history as possible into the context window. Older turns are truncated.
- **RAG**: These models retrieve relevant turns or sessions for a given user query, concatenate them with the query, and feed the resulting input to the LLM for response generation. We use turns as the default granularity for better performance.
- **Personalized Dialogue Agents**: We consider two agent systems: (1) MemoryBank (Zhong et al., 2024) treats conversation history as a fixed database and modulates retrieval using heuristics based on the forgetting curve. (2) LD-Agent (Li et al., 2025) employs fixed conversation databases with additional retrieval modulation using strategies such as keywords matching.

8. Experimental Results

8.1. Main Results

We present the main results shown in Table 1 and analyze each method's performance as follow.

History matters: Without any history, the LLM performs poorly, achieving a METEOR score of 5.2% on MSC and 0.0% accuracy on LongMemEval, showing the necessity of historical context.

Long context is not enough: Long-context models struggle due to fixed context windows and the inclusion of noisy context. On MSC, scores remain low (*e.g.,* METEOR below 20%, BERT score under 40%), and on LongMemEval, accuracy is lower than 58%. This limitation highlights their inability to retain and utilize long-term knowledge.

Method	Retriever	MSC		LongMemEval	
		METEOR (%) ↑	**BERT** (%) ↑	**Recall@5** (%) ↑	**Acc.** (%) ↑
No History	-	5.2	10.6	-	0.0
Long Context	-	14.8	31.9	-	57.4
	Contriever	24.8	50.8	54.3	58.8
RAG	Stella	26.2	51.6	59.2	61.4
	GTE	27.5	52.1	62.4	63.6
MemoryBank	Specific1	20.1	40.3	58.6	59.6
LD-Agent	Specific2	25.4	51.5	56.8	59.2
	Contriever	30.8	55.4	60.4	61.2
RMM (Ours)	Stella	31.9	56.3	65.9	64.8
	GTE	**33.4**	**57.1**	**69.8**	**70.4**
RAG	Oracle	-	-	100.0	90.2

Table 1 | Performance comparison of RMM with baseline methods on the MSC and LongMemEval datasets. Metrics include METEOR and BERT Scores for MSC, and Recall@5 and Accuracy (Acc.) scores for LongMemEval. RMM demonstrates superior performance across all metrics, highlighting its effectiveness in retrieval relevance and personalized response generation. No oracle retrieval is available for the MSC dataset. MemoryBank and LD-Agent utilize their specific methods for retrieval. Scores are averaged over 3 runs and are reported in percentage (%).

RAG Models: RAG models outperform Long-Context LLMs by only incorporating relevant histories. With strong retrievers like GTE, RAG achieves 27.5% METEOR and 52.1% BERT Scores on MSC and 62.4% recall and 63.6% accuracy on LongMemEval. We also observe that the performance is retriever-dependent, where stronger retrievers boost the performance.

Personalized Dialogue Agents: MemoryBank and LD-Agent show more moderate improvements over Long-Context LLMs. For instance, LD-Agent achieves 25.4% METEOR and 51.5% BERT score on MSC, but these models fall short of RAG and RMM. Their reliance on heuristic-based retrieval potentially limits adaptability to complex tasks.

Proposed RMM Framework: RMM consistently achieves the best results across datasets and metrics. With GTE, RMM achieves 33.4% METEOR and 57.1% BERT on MSC, and 69.8% recall and 70.4% accuracy on LongMemEval. Even with weaker retrievers like Contriever, RMM maintains competitive performance, demonstrating robustness. The improvements stem from RMM's ability to integrate dynamic memory management with adaptive retrieval optimization enables it to retrieve and utilize relevant knowledge effectively, outperforming all baselines.

To further assess the impact of memory integration, we calculate the proportion of test examples where memory improves response quality. On MSC, memory improves on 86% of responses, as the dataset frequently requires recalling prior discussion topics. On LongMemEval, where questions are deliberately designed to test historical recall, memory contributes to quality improvements in 100% of cases. These results show the necessity of memory mechanisms in maintaining long-term coherence. We provide case studies of the memory usage in Appendix C.

8.2. Ablation Study

We conduct ablation study to evaluate the contributions of key components in the RMM framework. we present the results in Table 2 and list our observations as below.

(***i***) Adding Prospective Reflection boosts performance by organizing the memory into structured topics, which reduces redundancy and improves relevance. (***ii***) Retrospective Reflection alone without a reranker misaligns retrieved content, leading to suboptimal results. Directly updating the retriever using RL rewards requires extensive amounts of training data for effective full fine-tuning, which is often difficult to obtain in real-world scenarios. Without sufficient data, it can lead to issues like catastrophic forgetting (McCloskey and Cohen, 1989). (***iii***) The addition of the reranker alongside RR significantly enhances alignment, achieving 27.5% METEOR and 58.8% Recall@5, demonstrating its effectiveness in refining retrieval quality. (***iv***) Finally, the complete RMM framework, which integrates Prospective Reflection, Retrospective Reflection, and the reranker, achieves the best results across all metrics, with a METEOR score of 30.8% on MSC and 60.4% Recall@5 on LongMemEval. This confirms that RMM enables more accurate and efficient future retrieval.

Variant	MSC		LongMemEval	
	METEOR	**BERT**	**Recall@5**	**Acc.**
RAG	24.8	50.8	54.3	58.8
+ PR	28.6	53.3	57.4	59.6
+ RR (W/O reranker)	20.3	31.8	34.2	31.0
+ RR	27.5	52.2	58.8	60.2
RMM	**30.8**	**55.4**	**60.4**	**61.2**

Table 2 | Ablation study on the datasets. Variants evaluate the impact of key components in RMM: Prospective Reflection (PR), Retrospective Reflection (RR), and the reranker. RR (W/O reranker) means the retriever is fine-tuned instead. Scores are obtained with Contriever and Gemini-1.5-Flash and in percentage (%).

8.3. Validation of Citation Scores

Our framework leverages LLM-generated citations to determine reward scores, guiding the retrieval refinement process. To assess the validity of the citation scores, we conduct evaluation on the LongMemEval dataset, using the Gemini-1.5-Pro model as the judge. The experiment tasks the LLM with determining whether cited memories were useful for response generation. The results, presented in Table 3, demonstrate high precision, recall, and F1, confirming the effectiveness of citation-based scoring in our framework.

8.4. Effect of Different LLMs

To examine the effect of different LLMs as generators, we evaluate both Gemini-1.5-Flash and Gemini-1.5-Pro in Long-Context LLMs and RMM. As shown in Table 4, for Long-Context models, Gemini-1.5-Pro achieves slightly better performance than Gemini-1.5-Flash across all metrics, suggesting that a stronger model improves response quality when relying solely on extended context windows. However, for RMM, Gemini-1.5-Flash outperforms Gemini-1.5-Pro, achieving higher METEOR and BERT scores on MSC and better accuracy on LongMemEval. Similar observations are reported by Wu

Metric	Precision	Recall	F1
Useful memory	89.4	91.1	90.2
Not useful memory	87.2	84.6	85.9
Overall	87.6	85.8	86.7

Table 3 | Evaluation of citation-based scoring in RR for useful memory identification on LongMemEval (results in %).

Method	LLM	MSC		LongMemEval
		METEOR	**BERT**	**Acc.**
Long Context	Gemini-1.5-Flash	14.8	31.9	**57.4**
	Gemini-1.5-Pro	**17.4**	**36.1**	56.6
RMM	Gemini-1.5-Flash	**30.8**	**55.4**	**61.2**
	Gemini-1.5-Pro	24.6	50.6	58.6

Table 4 | Effect of different LLMs on MSC and LongMemEval. Results (in %) compare Long-Context LLMs and RMM using the Contriever retriever with Gemini models as generators.
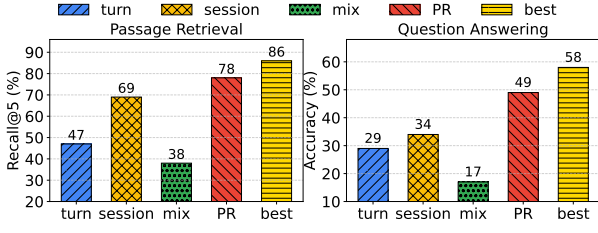
Figure 4 | Granularity analysis on randomly sampled 100 instances from `LongMemEval` with the GTE retriever and Gemini-1.5-Flash generator. "Turn" and "Session" indicate retrieval at a fixed granularity. "Mix" represents retrieving from a pool combining both turns and sessions. "PR" refers to the granularity resulting from the proposed Prospective Reflection, while "Best" corresponds to selecting the optimal granularity (either turn or session) for each instance.
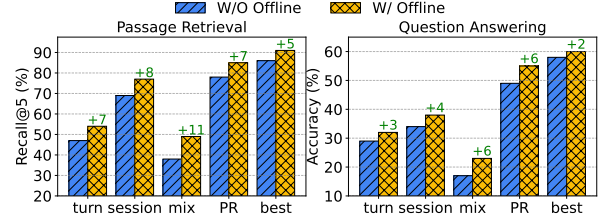

Figure 5 | Impact of offline pretraining on retriever performance for `LongMemEval` dataset with the same 100 random samples as Figure 4. Results without offline pretraining are shown in blue, while results with offline pretraining are shown in orange. Offline pretraining improves recall and accuracy across all settings.

et al. (2025), where GPT-4o-mini performs better than GPT-4o in personal knowledge QA. This trend can be attributed to stronger LLMs, such as Gemini-1.5-Pro, being more likely to abstain from answering queries involving personal information, possibly due to stronger alignment tuning aimed at enhancing privacy protection.

8.5. Effect of Different Granularities

We conduct experiments to show the advantage of the flexible granularity resulting from the proposed Prospective Reflection (PR) over pre-defined fixed granularities as baselines. Results in Figure 4 show that fixed granularities, such as "turn" and "session", achieve moderate performance, with session-level retrieval outperforming turn-level due to richer contexts. The "mixed" granularity underperforms, likely due to increased noise from a larger search space. The best configuration, which selects the optimal granularity per instance, achieves the highest scores, demonstrating the importance of adaptive memory organization. In contrast, PR improves performance by integrating fragmented conversational segments into cohesive memory structure, exhibiting an approaching performance with the best oracle granularity.

8.6. Offline Supervised Training

We further investigate the applicability of RMM in scenarios where a handful of labelled retrieval data is available, allowing for offline supervised pretraining (based on the off-the-shelf retriever) before online refinement. Figure 5 illustrates the impact of offline pretraining on retriever performance on the `LongMemEval` dataset. We randomly select 100 samples as test data with the rest as training and validation sets and apply vanilla supervised contrastive learning for the GTE retriever (Li et al., 2023). As the results show, across all settings, RMM consistently benefits from offline pretraining (orange bars) by outperforming retrievers without pretraining (blue bars). These results demonstrate that offline pretraining can enhance the retriever's ability to identify relevant information, providing a robust foundation for subsequent fine-tuning via RL.

8.7. The Impact of Top-*K* and Top-*M* for RMM

The results in Table 5 evaluate the impact of the number of retrieved memories (Top-K) and the number of reranked memories used for response generation (Top-M) in the RMM framework. Specifically, we analyze the performance on LongMemEval using Recall@5 (Top-K = 20, Top-M = 5), Recall@10 (Top-K = 50, Top-M = 10), and their corresponding QA accuracy scores.

The results demonstrate two key findings. First, increasing the number of memories (M) from 5 to 10 consistently improves both retrieval and accuracy metrics across all retrievers. For example, with the GTE retriever, Recall improves from 69.8% to 74.4%, and Accuracy increases from 70.4% to 73.8%. Second, the performance gain is most significant for stronger retrievers like GTE and Stella, highlighting the importance of retrieval quality. RMM with GTE achieves the best results of 70.4% Accuracy with Top-K = 20, Top-M = 5 and 73.8% Accuracy with Top-K = 50, Top-M = 10.

These observations emphasize that careful selection of Top-K and Top-M values can enhance both retrieval relevance and downstream QA performance. The combination of effective retrieval and reranking ensures that RMM efficiently leverages the most relevant information for long-term dialogue tasks.

8.8. Results for MSC with LLM-as-a-judge

Model	Retriever	LongMemEval			
		Recall@5	Acc.	Recall@10	Acc.
RMM	Contriever	60.4	61.2	67.2	66.8
	Stella	65.9	64.8	70.6	71.0
	GTE	**69.8**	**70.4**	**74.4**	**73.8**

Table 5 | Impact of Top-K (retrieved memories) and Top-M (reranked memories) on LongMemEval performance. Results include Recall@5 (Top-K = 20, Top-M = 5) and Recall@10 (Top-K = 50, Top-M = 10), and corresponding Accuracy across different retrievers. Results show that increasing the number of retrieved and reranked memories improves retrieval and QA performance on the LongMemEval dataset.

Method	LLM	METEOR	BERT	LLM-as-a-Judge (Yes%)
Long Context	Gemini-1.5-Flash	14.8	31.9	**25.4**
	Gemini-1.5-Pro	**17.4**	**36.1**	22.8
RMM	Gemini-1.5-Flash	**30.8**	**55.4**	**69.7**
	Gemini-1.5-Pro	24.6	50.6	65.4

Table 6 | Results for MSC with LLM-as-a-judge. RMM shows consistent advantages with more fine-grained evaluation.

Pair	Cohen's Kappa (κ)	Interpretation
Human A vs. Human B	0.82	Substantial agreement
LLM vs. Human A	0.71	Substantial agreement
LLM vs. Human B	0.69	Substantial agreement

Table 7 | Cohen's Kappa Agreement Between Annotators and LLM

For fair comparison, we follow prior work (Wu et al., 2025; Xu et al., 2022) and use METEOR and BERTScore. Here we include additional results using LLM-as-a-judge. Following Wu et al. (2025), we use Gemini-1.5-Pro to decide whether the generated answer matches the ground-truth as a binary annotation. The prompt we used in given in Appendix D.3. LLM-as-a-judge results also show the effectiveness of the proposed RMM.

Furthermore, we conducted a human evaluation on 100 randomly sampled instances from the MSC dataset. For each instance, annotators were shown the user query, the ground-truth response, and the model-generated response—mirroring the setup used for the LLM-based evaluations. We adopted the same instruction used in our LLM judge prompt, and asked human annotators to answer Yes/No to evaluate each instance. The annotation was independently performed by two NLP researchers. Both are PhD students working in NLP and not on the author list. We report inter-annotator agreement and agreement between human judgments and the LLM judge in Table 7.

9. Conclusion

We present RMM, a framework that integrates Prospective Reflection for structured, topic-based memory organization and Retrospective Reflection for dynamic memory reranking via reinforcement learning. Experimental results on benchmark datasets demonstrate that RMM outperforms state-of-the-art baselines in retrieval relevance and response quality for personalized dialogue tasks. By identifying limitations in existing memory management approaches—particularly those relying on fixed granularity and static retrievers, we highlight key challenges and avenues for future research in long-term dialogue memory modeling.

Limitations

While the proposed RMM framework demonstrates significant improvements in retrieval relevance and response quality, it is not without limitations. First, RMM relies on reinforcement learning for memory reranking, which can be computationally expensive, especially for large-scale datasets or real-time applications. Second, the current framework primarily focuses on textual data, limiting its applicability to multi-modal dialogue systems that incorporate images, audio, or video. Additionally, the memory updating mechanism may require further optimization to handle dynamically evolving long-term user interactions efficiently.

For future work, we plan to address these limitations by exploring more efficient reinforcement learning techniques and lightweight memory reranking strategies. We also aim to extend RMM to multi-modal dialogue systems to accommodate diverse user interactions. Furthermore, we will investigate privacy-preserving techniques to ensure safe deployment of RMM in real-world personalized dialogue applications where sensitive user data is involved.

Ethical Statement

This work focuses on developing a framework for long-term personalized dialogue systems to improve user experiences. However, we acknowledge the potential ethical implications of handling personal data in such systems. The RMM framework relies on historical conversations, which may contain sensitive or private information. To mitigate privacy risks, we recommend adopting robust encryption and privacy-preserving methods, such as differential privacy or federated learning, during data collection and model training.

Additionally, we emphasize the importance of transparent data usage policies and obtaining user consent when deploying personalized dialogue systems. Efforts should also be made to minimize biases in memory retrieval and response generation to ensure fairness and inclusivity across diverse user groups. Future work will continue to prioritize ethical considerations to promote the responsible development and deployment of personalized dialogue technologies.

References

S. Bae, D. Kwak, S. Kang, M. Y. Lee, S. Kim, Y. Jeong, H. Kim, S.-W. Lee, W. Park, and N. Sung. Keep me updated! memory management in long-term conversations. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3769–3787, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.276. URL https://aclanthology.org/2022.findings-emnlp.276.

S. Banerjee and A. Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In J. Goldstein, A. Lavie, C.-Y. Lin, and C. Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909.

J. Buchmann, X. Liu, and I. Gurevych. Attribute or abstain: Large language models as long document assistants. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8113–8140, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.463. URL https://aclanthology.org/2024.emnlp-main.463.

J. Chen, Z. Liu, X. Huang, C. Wu, Q. Liu, G. Jiang, Y. Pu, Y. Lei, X. Chen, X. Wang, K. Zheng, D. Lian, and E. Chen. When large language models meet personalization: perspectives of challenges and opportunities. *World Wide Web*, 27(4), June 2024. ISSN 1386-145X. doi: 10.1007/s11280-024-01276-1. URL https://doi.org/10.1007/s11280-024-01276-1.

P. Chhikara, D. Khant, S. Aryan, T. Singh, and D. Yadav. Mem0: Building production-ready ai agents with scalable long-term memory, 2025. URL https://arxiv.org/abs/2504.19413.

Y. R. Dong, T. Hu, and N. Collier. Can LLM be a personalized judge? In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.592. URL https://aclanthology.org/2024.findings-emnlp.592/.

Y. Guan, D. Wang, Z. Chu, S. Wang, F. Ni, R. Song, and C. Zhuang. Intelligent agents with llm-based process automation. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 5018–5027, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671646. URL https://doi.org/10.1145/3637528.3671646.

E. Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications: A Series of Lectures*. Applied mathematics series. U.S. Government Printing Office, 1954. URL https://books.google.com/books?id=SNpJAAAAMAAJ.

G. Izacard, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, and E. Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=jKN1pXi7b0.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=rkE3y85ee.

Z. Jiang, M. Sun, L. Liang, and Z. Zhang. Retrieve, summarize, plan: Advancing multi-hop question answering with an iterative approach. In *Companion Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1677–1686, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400713316. doi: 10.1145/3701716.3716889. URL https://doi.org/10.1145/3701716.3716889.

K. Kenthapadi, M. Sameki, and A. Taly. Grounding and evaluation for large language models: Practical challenges and lessons learned (survey). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6523–6533, New York, NY, USA, 2024.

Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671467. URL https://doi.org/10.1145/3637528.3671467.

S. Kolasani. Optimizing natural language processing, large language models (llms) for efficient customer service, and hyper-personalization to enable sustainable growth and revenue. *Transactions on Latest Trends in Artificial Intelligence*, 4(4), 2023. ISSN 3246-548X. URL https://ijsdcs.com/index.php/TLAI/article/view/476.

G. Lee, V. Hartmann, J. Park, D. Papailiopoulos, and K. Lee. Prompted LLMs as chatbot modules for long open-domain conversation. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4536–4554, Toronto, Canada, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.277. URL https://aclanthology.org/2023.findings-acl.277.

H. Li, P. Verga, P. Sen, B. Yang, V. Viswanathan, P. Lewis, T. Watanabe, and Y. Su. Alr^2: A retrieve-then-reason framework for long-context question answering. *ArXiv preprint*, abs/2410.03227, 2024a. URL https://arxiv.org/abs/2410.03227.

H. Li, C. Yang, A. Zhang, Y. Deng, X. Wang, and T.-S. Chua. Hello again! LLM-powered personalized agent for long-term dialogue. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *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 5259–5276, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.272. URL https://aclanthology.org/2025.naacl-long.272/.

Y. Li, H. Wen, W. Wang, X. Li, Y. Yuan, G. Liu, J. Liu, W. Xu, X. Wang, Y. Sun, et al. Personal llm agents: Insights and survey about the capability, efficiency and security. *ArXiv preprint*, abs/2401.05459, 2024b. URL https://arxiv.org/abs/2401.05459.

Y. LI, H. Jiang, Q. Wu, X. Luo, S. Ahn, C. Zhang, A. H. Abdi, D. Li, J. Gao, Y. Yang, and L. Qiu. SCBench: A KV cache-centric analysis of long-context methods. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=gkUyYcY1W9.

Z. Li, X. Zhang, Y. Zhang, D. Long, P. Xie, and M. Zhang. Towards general text embeddings with multi-stage contrastive learning. *ArXiv preprint*, abs/2308.03281, 2023. URL https://arxiv.org/abs/2308.03281.

H. Liu, M. Zaharia, and P. Abbeel. Ringattention with blockwise transformers for near-infinite context. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=WsRHpHH4s0.

N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, and P. Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12: 157–173, 2024b. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9.

X. Liu, Z. Tang, P. Dong, Z. Li, B. Li, X. Hu, and X. Chu. Chunkkv: Semantic-preserving kv cache compression for efficient long-context llm inference. *ArXiv preprint*, abs/2502.00299, 2025. URL https://arxiv.org/abs/2502.00299.

J. Lu, S. An, M. Lin, G. Pergola, Y. He, D. Yin, X. Sun, and Y. Wu. Memochat: Tuning llms to use memos for consistent long-range open-domain conversation. *ArXiv preprint*, abs/2308.08239, 2023. URL https://arxiv.org/abs/2308.08239.

C. J. Maddison, D. Tarlow, and T. Minka. A* sampling. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/309fee4e541e51de2e41f21bebb342aa-Paper.pdf.

A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. Evaluating very long-term conversational memory of LLM agents. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13851–13870, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.747. URL https://aclanthology.org/2024.acl-long.747/.

M. McCloskey and N. J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

J. Mendonça, A. Lavie, and I. Trancoso. On the benchmarking of LLMs for open-domain dialogue evaluation. In E. Nouri, A. Rastogi, G. Spithourakis, B. Liu, Y.-N. Chen, Y. Li, A. Albalak, H. Wakaki, and A. Papangelis, editors, *Proceedings of the 6th Workshop on NLP for Conversational AI (NLP4ConvAI 2024)*, pages 1–12, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.nlp4convai-1.1/.

K. T.-i. Ong, N. Kim, M. Gwak, H. Chae, T. Kwon, Y. Jo, S.-w. Hwang, D. Lee, and J. Yeo. Towards lifelong dialogue agents via timeline-based memory management. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *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 8631–8661, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.435. URL https://aclanthology.org/2025.naacl-long.435/.

C. Packer, S. Wooders, K. Lin, V. Fang, S. G. Patil, I. Stoica, and J. E. Gonzalez. Memgpt: Towards llms as operating systems, 2024. URL https://arxiv.org/abs/2310.08560.

Z. Pan, Q. Wu, H. Jiang, X. Luo, H. Cheng, D. Li, Y. Yang, C.-Y. Lin, H. V. Zhao, L. Qiu, and J. Gao. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xKDZAW0He3.

J. Pei, P. Ren, and M. de Rijke. A cooperative memory network for personalized task-oriented dialogue systems with incomplete user profiles. In J. Leskovec, M. Grobelnik, M. Najork, J. Tang, and L. Zia, editors, *WWW '21: The Web Conference 2021, Virtual Event / Ljubljana, Slovenia, April 19-23, 2021*, pages 1552–1561. ACM / IW3C2, 2021. doi: 10.1145/3442381.3449843. URL https://doi.org/10.1145/3442381.3449843.

P. Rasmussen, P. Paliychuk, T. Beauvais, J. Ryan, and D. Chalef. Zep: A temporal knowledge graph architecture for agent memory, 2025. URL https://arxiv.org/abs/2501.13956.

F. Shi, X. Chen, K. Misra, N. Scales, D. Dohan, E. H. Chi, N. Schärli, and D. Zhou. Large language models can be easily distracted by irrelevant context. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR, 2023. URL https://proceedings.mlr.press/v202/shi23a.html.

Y.-M. Tseng, Y.-C. Huang, T.-Y. Hsiao, W.-L. Chen, C.-W. Huang, Y. Meng, and Y.-N. Chen. Two tales of persona in LLMs: A survey of role-playing and personalization. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024,* pages 16612–16631, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.969. URL https://aclanthology.org/2024.findings-emnlp.969/.

M. A. Walker, D. J. Litman, C. A. Kamm, and A. Abella. PARADISE: A framework for evaluating spoken dialogue agents. In *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*, pages 271–280, Madrid, Spain, July 1997. Association for Computational Linguistics. doi: 10.3115/976909.979652. URL https://aclanthology.org/P97-1035/.

Q. Wang, L. Ding, Y. Cao, Z. Tian, S. Wang, D. Tao, and L. Guo. Recursively summarizing enables long-term dialogue memory in large language models. *ArXiv preprint*, abs/2308.15022, 2023. URL https://arxiv.org/abs/2308.15022.

Z. Z. Wang, J. Mao, D. Fried, and G. Neubig. Agent workflow memory. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=NTAhi2JEEE.

J. Weizenbaum. Eliza—a computer program for the study of natural language communication between man and machine. *Commun. ACM*, 9(1):36–45, Jan. 1966. ISSN 0001-0782. doi: 10.1145/365153.365168. URL https://doi.org/10.1145/365153.365168.

Q. Wen, J. Liang, C. Sierra, R. Luckin, R. Tong, Z. Liu, P. Cui, and J. Tang. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '24, page 6743–6744, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704901. doi: 10.1145/3637528.3671498. URL https://doi.org/10.1145/3637528.3671498.

S. Whittaker, Q. Jones, and L. Terveen. Managing long term communications: conversation and contact management. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences*, pages 1070–1079, 2002. doi: 10.1109/HICSS.2002.994063. URL https://doi.org/10.1109/HICSS.2002.994063.

M. D. Williams and J. D. Hollan. The process of retrieval from very long-term memory. *Cognitive Science*, 5(2):87–119, 1981. ISSN 0364-0213. URL https://www.sciencedirect.com/science/article/pii/S0364021381800286.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.*, 8(3–4):229–256, May 1992. ISSN 0885-6125. doi: 10.1007/BF00992696. URL https://doi.org/10.1007/BF00992696.

D. Wu, H. Wang, W. Yu, Y. Zhang, K.-W. Chang, and D. Yu. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=pZiyCaVuti.

J. Xu, A. Szlam, and J. Weston. Beyond goldfish memory: Long-term open-domain conversation. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.356. URL https://aclanthology.org/2022.acl-long.356.

W. Xu, K. Mei, H. Gao, J. Tan, Z. Liang, and Y. Zhang. A-mem: Agentic memory for llm agents, 2025. URL https://arxiv.org/abs/2502.12110.

C. Zhang, Y. Sun, J. Chen, J. Lei, M. Abdul-Mageed, S. Wang, R. Jin, S. Park, N. Yao, and B. Long. Spar: Personalized content-based recommendation via long engagement attention. *ArXiv preprint*, abs/2402.10555, 2024a. URL https://arxiv.org/abs/2402.10555.

D. Zhang, J. Li, Z. Zeng, and F. Wang. Jasper and stella: distillation of sota embedding models. *ArXiv preprint*, abs/2412.19048, 2024b. URL https://arxiv.org/abs/2412.19048.

T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

Z. Zhang, R. A. Rossi, B. Kveton, Y. Shao, D. Yang, H. Zamani, F. Dernoncourt, J. Barrow, T. Yu, S. Kim, et al. Personalization of large language models: A survey. *ArXiv preprint*, abs/2411.00027, 2024c. URL https://arxiv.org/abs/2411.00027.

Z. Zhang, D. Zhang-Li, J. Yu, L. Gong, J. Zhou, Z. Liu, L. Hou, and J. Li. Simulating classroom education with llm-empowered agents. *ArXiv preprint*, abs/2406.19226, 2024d. URL https://arxiv.org/abs/2406.19226.

L. Zhao, X. Feng, X. Feng, W. Zhong, D. Xu, Q. Yang, H. Liu, B. Qin, and T. Liu. Length extrapolation of transformers: A survey from the perspective of positional encoding. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9959–9977, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.582. URL https://aclanthology.org/2024.findings-emnlp.582/.

C. Zheng, Y. Gao, H. Shi, M. Huang, J. Li, J. Xiong, X. Ren, M. Ng, X. Jiang, Z. Li, and Y. Li. Dape: Data-adaptive positional encoding for length extrapolation. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 26659–26700. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/2f050fa9f0d898e3f265d515f50ae8f9-Paper-Conference.pdf.

W. Zhong, L. Guo, Q. Gao, H. Ye, and Y. Wang. Memorybank: Enhancing large language models with long-term memory. In M. J. Wooldridge, J. G. Dy, and S. Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 19724–19731. AAAI Press, 2024. doi: 10.1609/AAAI.V38I17.29946. URL https://doi.org/10.1609/aaai.v38i17.29946.

A. Implementation and Training Details

A.1. Parameter Setup

We use the following hyper-parameters for all experiments:

- **Reranker:** The reranker is an MLP with a residual connection. The training setup is:
  - Batch size: 4
  - Top-M: 5
  - Top-K: 20
- **Reinforcement Learning:** Retrospective Reflection uses REINFORCE with:
  - Batch size: 4
  - Gumbel temperature (τ): 0.5
  - Reward (R): +1 for cited entries, −1 for non-cited entries
  - Baseline value (b): 0.5
  - Learning rate for policy gradient updates (η): 1×10^{-3}
- **LLM:** Gemini-1.5-Flash/-Pro is used for response generation with:
  - Context window size: 128k tokens
  - Temperature: 0.0
- **Retriever:** GTE for experiments in Section 8.6 is pretrained with supervised contrastive learning using the following configuration:
  - Learning rate: 1×10^{-4}
  - Training epochs: 10
  - Batch size: 32
  - Top-K: 5

A.2. Dependencies

Our implementation relies on the following tools and libraries:

- **Programming Language**: Python 3.10.13
- **Core Libraries**: PyTorch 2.4.1+cu121, Hugging Face Transformers 4.44.2
- **Utilities**: NumPy, Pandas, Sklearn and Matplotlib for data processing and visualization

A.3. Hardware and Reproducibility

All experiments are conducted on a server with the following hardware configuration:

- **GPUs**: 16 NVIDIA A100 GPUs
- **RAM**: 40 GB
- **CUDA Version**: 12.2

A.4. Details for MemoryBank and LD-Agent Baselines

We integrate MemoryBank and LD-Agent as baselines, with key features implemented using the LongMemEval codebase[1]. We use Contriever as the default retriever. Particularly, they differ in the way for structuring and accessing stored information.

MemoryBank (Zhong et al., 2024) retrieves historical context by maintaining a structured memory where both conversational summaries and round-level utterances are stored as key-value pairs. The retrieval process involves directly matching user queries to the most relevant stored information, ensuring efficient context retrieval for response generation.

LD-Agent (Li et al., 2025), on the other hand, enhances retrieval by incorporating keyphrase-based queries. In addition to storing factual and summarized information, its retrieval is based on queries with key phrases

Figure 6 | Convergence of usefulness scores (ratio of useful memories cited) over RL training steps. The score improves as the reranker is updated based on Retrospective Reflection, indicating enhanced alignment between retrieved memory and generated responses.

extracted from past interactions. This enables the model to adapt more effectively to diverse query formulations, retrieving context that aligns with the underlying semantic meaning of the user input.

For both methods, retrieval operates in a non-hierarchical manner, meaning that all stored data is accessed through a uniform search mechanism without additional interaction-based refinement. The retrieved content is then used to provide historical grounding for response generation.

A.5. The Convergence of Citation Scores in RL

Figure 6 illustrates the convergence of citation scores (usefulness scores) during reinforcement learning. The x-axis represents the RL training steps, while the y-axis measures the ratio of useful memories cited by the LLM generator. Initially, the usefulness score starts at a low value around 0.2, reflecting the misalignment between retrieved memories and response generation. As training progresses, the score steadily increases, converging to approximately 0.4 by step 1000. This trend highlights the effectiveness of Retrospective Reflection in updating the reranker, allowing the retrieval process to better align with the generator's citation behavior. The gradual convergence indicates stable learning and suggests that RL fine-tuning improves retrieval quality without overfitting.

B. Dataset Description

We conduct experiments on two publicly available datasets: MSC (Xu et al., 2022) and LongMemEval (Wu et al., 2025). MSC is a benchmark dataset for multi-session conversations, providing turn-level and session-level conversational data with annotations for relevance and response quality. On this dataset, following Li et al. (2025), we evaluate the ability of an LLM agent to produce human-like personalized responses. Each response can be grounded in historical context across multiple previous sessions. The focus is on accurately generating personalized responses by leveraging relevant user preferences and conversation patterns. We followed the methodology outlined by Li et al. (2025) to construct the data for our experiments. Specifically, we use the first 1000 sessions as chat history and the rest for
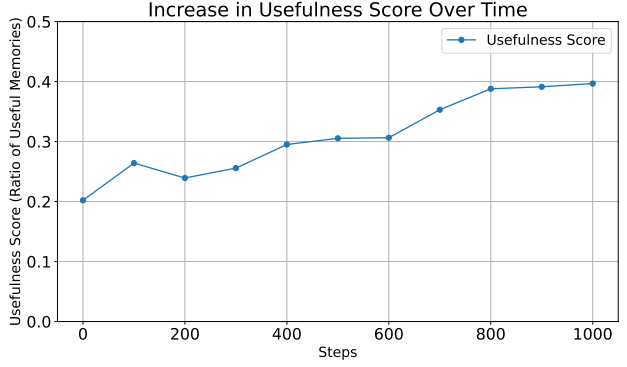
---

[1] https://github.com/xiaowu0162/LongMemEval

evaluation.

`LongMemEval` is designed for long-term conversational evaluation. It includes extended histories across turn, session, and mixed granularities. For experiments in Section 8.6, we randomly sample 100 test instances and use the remaining data for training and validation. On this dataset, following Li et al. (2025), we evaluate the system's ability to answer human-designed questions about specific personal knowledge described in the historical sessions. For example, given a query like, "What car did Mary buy last summer?", the system must retrieve and synthesize information scattered across multiple sessions. The task emphasizes accurately identifying and leveraging relevant details from long-term memory.

C. Case Studies

We present case studies to illustrate how RMM effectively integrates relevant memory fragments to enhance response quality. The following examples highlight scenarios where historical context is essential for maintaining coherence and accuracy in long-term dialogue.

C.1. Case 1: Revisiting Fitness Choices (MSC)

Tracking personal preferences and habits across multiple conversations is essential for maintaining coherent and personalized dialogue. In this case, the user initially considers purchasing a treadmill (Session A), later expresses a preference for using the gym treadmill due to weather constraints (Session B), and finally confirms their gym-going routine (Session C). An effective memory mechanism should correctly track this evolving decision and retrieve the most up-to-date preference.

Case 1: Revisiting Fitness Choices (MSC)

Session A, Turn 3:

- **Speaker_1:** Ah, got it. Well, maybe one of the older gyms will work out better for you — or I guess you could get that **treadmill** you were talking about before.
- **Speaker_2:** I'm leaning towards the **treadmill**. I think it will work better for my lifestyle.

Session B, Turn 2:

- **Speaker_1:** I go to the gym at least five times a week, and I lift weights at least three of those days. When I need to give my arms a break, I work on my leg muscles. I run around the track or just ride the stationary exercise bicycle.
- **Speaker_2:** That sounds like a good plan. I definitely need to add some weights to my routine. I will be on the **treadmills** a lot, especially since it is hard for me to run outdoors daily due to the weather.

Session C, Turn 2:

- **[Question] Speaker_1:** They are great also, thanks for asking. Are you still going to the gym?
- **[Answer] Speaker_2:**
  - **Ground-truth:** Yes, every night. I run on a **treadmill**.
  - **Output (RMM):** Yes, I go to the gym and run on the **treadmill**. It has become a key part of my routine.
  - **Output (Long Context):** I have been considering getting a **treadmill** for home, but I am still unsure. I haven't decided yet.

Analysis: The user's decision about treadmill usage shifts across sessions. Initially, in Session A, they express interest in buying a treadmill. By Session B, they reconsider and decide that using the gym treadmill would be sufficient and confirm that they run on the treadmill at the gym. Without memory management, the model generates an outdated response, assuming the user is still undecided about purchasing a treadmill.

C.2. Case 2: Tracking Chronological Order of Events (LongMemEval)

In long-term interactions, correctly recalling the sequence of past events is essential for maintaining factual consistency. This case examines whether the model can track the order in which the user attended two different events.

Case 2: Tracking Chronological Order of Events (`LongMemEval`)

Session A, Turn 1:

- **Speaker_1:** I recently attended a workshop on **Effective Time Management** at the local community center. It was incredibly insightful and gave me some strategies for managing tasks efficiently. What are your thoughts on time management workshops?
- **Speaker_2:** They're quite helpful for understanding basic techniques, but it also depends on how you apply the strategies in daily life.

Session B, Turn 3:

- **Speaker_1:** I also attended a webinar on **Data Analysis using Python** two months ago. The instructor shared some incredible resources for data visualization. Have you explored advanced visualization techniques?
- **Speaker_2:** Yes, they can greatly enhance how you present your data. Libraries like **Matplotlib** and **Seaborn** are good starting points for creating professional visuals.

[Question] Which event did I attend first, the "Effective Time Management" workshop or the "Data Analysis using Python" webinar?

[Answer]

- **Ground-truth:** "Data Analysis using Python" webinar.
- **Output (RMM):** You attended the **Data Analysis using Python** webinar two months ago. The **Effective Time Management** workshop happened later at the local community center.
- **Output (Long Context):** I'm not sure, but you mentioned both events in previous conversations.

Analysis: The correct response requires linking the time reference ("two months ago") with the corresponding event. Without RMM, the model fails to retrieve this detail, resulting in an uncertain and incomplete answer. With RMM, the model correctly recalls the chronological order, demonstrating the advantage of structured memory retrieval in tracking event sequences.

D. Prompts

D.1. Prospective Reflection

D.1.1. Memory Extraction

Function: Memory extraction for `SPEAKER_1`

Task Description: Given a session of dialogue between `SPEAKER_1` and `SPEAKER_2`, extract the personal summaries of `SPEAKER_1`, with references to the corresponding turn IDs. Ensure the output adheres to the following rules:

- Output results in **JSON format**. The top-level key is "extracted_memories". The value should be a list of dictionaries, where each dictionary has the keys "summary" and "reference":
  - **summary**: A concise personal summary, which captures relevant information about `SPEAKER_1`'s experiences, preferences, and background, across multiple turns.
  - **reference**: A list of references, each in the format of [turn_id] indicating where the information appears.
- If no personal summary can be extracted, return NO_TRAIT.

Example:
INPUT:

- **Turn 0:**
  - **SPEAKER_1:** Did you check out that new gym in town?
  - **SPEAKER_2:** Yeah, I did. I'm not sure I like the vibe there, though.
- **Turn 1:**
  - **SPEAKER_1:** What was wrong with it?
  - **SPEAKER_2:** The folks there seemed to care more about how they looked than working out. It was a little too trendy for me. I'm pretty plain.
- **Turn 2:**
  - **SPEAKER_1:** Ah, got it. Well, maybe one of the older gyms will work out better for you—or I guess you could get that treadmill you were talking about before. Are you leaning one way or the other yet?
  - **SPEAKER_2:** I'm leaning towards the treadmill. I think it will work better for my lifestyle. I just don't know which type to get. There are so many choices out there. Do you use a treadmill at your gym? Do you have a suggestion for a home one?
- **Turn 3:**
  - **SPEAKER_1:** I usually just lift weights there, to be honest. But I think I've heard good things about the NordicTrack?
  - **SPEAKER_2:** Yeah, I've heard good things about that, too. I like the idea of a multi-exercise piece of equipment. As long as the weather isn't too bad, then I prefer to go for a run. But since it rains quite a bit here, I like the idea of an inside option. How is the weather in New England?
- **Turn 4:**
  - **SPEAKER_1:** Oh, it can get pretty foggy and rainy here too, I'm afraid. But as I'm sure you've heard, it's really beautiful in the fall! Are there four distinct seasons where you are, too?
  - **SPEAKER_2:** Yes, I've heard about the fall colors. I may get there one day. Yes, we have seasons—rain, lighter rain, summer, and more rain! Ha!
- **Turn 5:**
  - **SPEAKER_1:** Haha! I lived overseas in the tropics once. Sounds just like it!
  - **SPEAKER_2:** The tropics sound great. It's not as warm as the tropics, but I like it. I'm from Alaska, so I'm pretty weather-tough.

OUTPUT:

```
{
    "extracted_memories": [
        {
            "summary": "SPEAKER_1 asked about a new gym in town and suggested older gyms or
    a treadmill as alternatives.",
            "reference": [0, 2]
        },
        {
            "summary": "SPEAKER_1 usually lifts weights at the gym rather than using a
    treadmill.",
            "reference": [3]
        },
        {
            "summary": "SPEAKER_1 has heard good things about the NordicTrack treadmill.",
            "reference": [3]
        },
        {
            "summary": "SPEAKER_1 lives in New England and experiences foggy and rainy
    weather but enjoys the fall season.",
            "reference": [4]
        },
        {
```

```
            "summary": "SPEAKER_1 has lived overseas in the tropics before.",
            "reference": [5]
        }
    ]
}
```
Task: Follow the JSON format demonstrated in the example above and extract the personal summaries for SPEAKER_1 from the following dialogue session.
Input: {}
Output:

Function: Memory extraction for SPEAKER_2

Task Description: Given a session of dialogue between SPEAKER_1 and SPEAKER_2, extract the personal summaries of SPEAKER_2, with references to the corresponding turn IDs. Ensure the output adheres to the following rules:

- Output results in **JSON format**. The top-level key is "extracted_memories". The value should be a list of dictionaries, where each dictionary has the keys "summary" and "reference":
    - **summary**: A concise personal summary, which captures relevant information about SPEAKER_2's experiences, preferences, and background, across multiple turns.
    - **reference**: A list of references, each in the format of [turn_id] indicating where the information appears.
- If no personal summary can be extracted, return NO_TRAIT.

Example:
INPUT:

- **Turn 0:**
    - **SPEAKER_1:** Did you manage to go out on a run today?
    - **SPEAKER_2:** Yes, I actually was able to. I am considering joining the local gym. Do you prefer going to the gym?

- **Turn 1:**
    - **SPEAKER_1:** I do actually. I like the controlled environment. I don't want to have to depend on the weather considering where I live.
    - **SPEAKER_2:** That's why I am thinking about it. I hate to have to run when it's raining, and I feel like it rains here all the time.

- **Turn 2:**
    - **SPEAKER_1:** A lot of gyms have tracks so that you can run indoors. Hey, have you thought about maybe buying a treadmill and using that at home?
    - **SPEAKER_2:** I am definitely considering getting one. I'm just trying to figure out what I would do more—go to the gym and actually do more than just running, or stick to what I know and get a treadmill.

- **Turn 3:**
    - **SPEAKER_1:** Oh, that's true. I hadn't thought about all of that. You're right. With a gym, there are a whole lot of options for what you can do. Do you have some good gyms near you?
    - **SPEAKER_2:** They just built one in the small town really close to me, and it looks pretty decent. Before that, it was like an hour drive.

- **Turn 4:**
    - **SPEAKER_1:** With you not owning a car, going to any others would probably be difficult. Well, do you have any good parks and running trails nearby?

 – **SPEAKER_2:** Yeah, exactly. There is a super nice little running trail that is pretty decent.

· **Turn 5:**

 – **SPEAKER_1:** Hey, do you run with anyone? I mean, have you joined a club, or will you if you haven't?
 – **SPEAKER_2:** There isn't any around here; maybe I could start one. Thank you for that idea.

OUTPUT:
```json
{
    "extracted_memories": [
        {
            "summary": "SPEAKER_2 is considering joining a local gym due to frequent rain
    affecting outdoor runs.",
            "reference": [0, 1]
        },
        {
            "summary": "SPEAKER_2 is debating between buying a treadmill for home use or
    going to the gym for more workout variety.",
            "reference": [2]
        },
        {
            "summary": "A new gym was recently built nearby SPEAKER_2, replacing a previous
     one that was an hour away.",
            "reference": [3]
        },
        {
            "summary": "SPEAKER_2 has access to a nice local running trail.",
            "reference": [4]
        },
        {
            "summary": "SPEAKER_2 notices there is no local running club but is considering
    starting one.",
            "reference": [5]
        }
    ]
}
```
Task: Follow the JSON format demonstrated in the example above and extract the personal summaries for SPEAKER_2 from the following dialogue session.
Input: {}
Output:

D.1.2. Memory Update

Task Description: Given a list of history personal summaries for a specific user and a new and similar personal summary from the same user, update the personal history summaries following the instructions below:

- **Input format:** Both the history personal summaries and the new personal summary are provided in JSON format, with the top-level keys of "history_summaries" and "new_summary".
- **Possible update actions:**
  - **Add:** If the new personal summary is not relevant to any history personal summary, add it.
    Format: Add()
  - **Merge:** If the new personal summary is relevant to a history personal summary, merge them as an updated summary.
    Format: Merge(index, merged_summary)
    Note: index is the position of the relevant history summary in the list. merged_summary is the merged summary of the new summary and the relevant history summary. Two summaries are considered relevant if they discuss the same aspect of the user's personal information or experiences.
- If multiple actions need to be executed, output each action in a single line, and separate them with a newline character ("\n").
- Do not include additional explanations or examples in the output—only return the required action functions.

Example:
INPUT:

- **History Personal Summaries:**
  - {"history_summaries": ["SPEAKER_1 works out although he doesn't particularly enjoy it."]}
- **New Personal Summary:**
  - {"new_summary": "SPEAKER_1 exercises every Monday and Thursday."}

OUTPUT ACTION:
Merge(0, SPEAKER_1 exercises every Monday and Thursday, although he doesn't particularly enjoy it.)

Task: Follow the example format above to update the personal history for the given case.
INPUT:

- **History Personal Summaries:** {}
- **New Personal Summary:** {}

OUTPUT ACTION:

D.2. Retrospective Reflection

Task Description: Given a user query and a list of memories consisting of personal summaries with their corresponding original turns, generate a natural and fluent response while adhering to the following guidelines:

- Cite **useful** memories using [i], where i corresponds to the index of the cited memory.
- Do **not cite memories that are not useful**. If no useful memory exist, output [NO_CITE].
- Each memory is independent and may repeat or contradict others. The response must be directly supported by cited memories.
- If the response relies on multiple memories, list all corresponding indices, e.g., [i, j, k].
- The citation is evaluated based on whether the response references the original turns, **not the summaries**.

Examples:
Case 1: Useful Memories Found
INPUT:

- **User Query:** SPEAKER_1: What hobbies do I enjoy?
- **Memories:**
    - **Memory [0]:** SPEAKER_1 enjoys hiking and often goes on weekend trips.
        * Speaker 1: I love spending my weekends hiking in the mountains.
          Speaker 2: That sounds amazing! Do you go alone or with friends?
        * Speaker 1: Last month, I hiked a new trail and it was amazing.
          Speaker 2: Nice! Which trail was it?
    - **Memory [1]:** SPEAKER_1 plays the guitar and occasionally performs at open mics.
        * Speaker 1: I've been practicing guitar for years and love playing at open mics.
          Speaker 2: That's awesome! What songs do you usually play?
        * Speaker 1: I performed at a local cafe last week and had a great time.
          Speaker 2: That must have been fun! Were there a lot of people?
    - **Memory [2]:** SPEAKER_1 is interested in astronomy and enjoys stargazing.
        * Speaker 1: I recently bought a telescope to get a closer look at planets.
          Speaker 2: That's so cool! What have you seen so far?
        * Speaker 1: I love stargazing, especially when there's a meteor shower.
          Speaker 2: I'd love to do that sometime. When's the next one?

Output: You enjoy hiking, playing the guitar, and stargazing. [0, 1, 2]

Case 2: No Useful Memories
INPUT:

- **User Query:** SPEAKER_1: What countries did I go to last summer?
- **Memories:**
    - **Memory [0]:** SPEAKER_1 enjoys hiking and often goes on weekend trips.
        * Speaker 1: I love spending my weekends hiking in the mountains.
          Speaker 2: That sounds amazing! Do you go alone or with friends?
        * Speaker 1: Last month, I hiked a new trail and it was amazing.
          Speaker 2: Nice! Which trail was it?
    - **Memory [1]:** SPEAKER_1 plays the guitar and occasionally performs at open mics.
        * Speaker 1: I've been practicing guitar for years and love playing at open mics.
          Speaker 2: That's awesome! What songs do you usually play?
        * Speaker 1: I performed at a local cafe last week and had a great time.
          Speaker 2: That must have been fun! Were there a lot of people?

```
        – Memory [2]: SPEAKER_1 is interested in astronomy and enjoys stargazing.
            * Speaker 1: I recently bought a telescope to get a closer look at planets.
              Speaker 2: That's so cool! What have you seen so far?
            * Speaker 1: I love stargazing, especially when there's a meteor shower.
              Speaker 2: I'd love to do that sometime. When's the next one?

Output: I don't have enough information to answer that. [NO_CITE]

Additional Instructions:

    · Ensure the response is fluent and directly answers the user's query.
    · Always cite the useful memory indices explicitly.
    · The citation is evaluated based on whether the response references the original turns,
      not the summaries.
    · Follow the format of the examples provided above.


Input:

    · User Query: {}
    · Memories: {}

Output:
```

D.3. LLM-as-a-Judge

```
You are an expert language model evaluator. I will provide you with a question, a
ground-truth answer, and a model-generated response. Your task is to determine whether
the response correctly answers the question by following these evaluation rules:

    · Answer Yes if the response contains or directly matches the correct answer.
    · Answer Yes if the response includes all necessary intermediate steps leading to the
      correct answer.
    · Answer No if the response provides only a partial answer or omits essential
      information.
    · Answer No if the response does not sufficiently address the question.

Examples:
Example 1: Correct Response

    · Question: What is the capital of France?
    · Ground-truth Answer: Paris
    · Response: The capital of France is Paris.

Evaluation:

    · Output: Yes


Example 2: Incorrect Response

    · Question: What is the capital of France?
    · Ground-truth Answer: Paris
    · Response: France is a country in Europe.

Evaluation:

    · Output: No


Additional Instructions:

    · Apply the evaluation criteria consistently.
```

- Base your decision strictly on the information in the response.
- Avoid subjective interpretations and adhere to the provided examples.

Input:

- **Question:** {}
- **Ground-truth Answer:** {}
- **Response:** {}

Output: