CoMem: A Benchmark for Continual Memory and Dynamic Preference Evolution in Long-Context Agents

Anonymous ACL submission

Abstract

The transition of Large Language Models (LLMs) from stateless engines to lifelong agents requires robust capabilities to track dynamic user preferences amidst temporal noise. However, most existing benchmarks predominantly focus on static retrieval fidelity, neglecting the longitudinal evolution of user states. To bridge this gap, we introduce **CoMem**, a benchmark designed to evaluate continual memory and dynamic preference evolution across sequential dialogue checkpoints. CoMem incorporates two distinct scenarios: user-assistant dyadic interactions and multi-party dialogues, across three context lengths (up to 128k tokens). We introduce a rigorous evaluation protocol, adapting metrics from continual learning—specifically Forgetting Measure and Forward Transfer—applied to sequential dialogue checkpoints. Through rigorous evaluation of diverse LLMs and memory architectures, our experiments yield four critical insights: (1) **Native Context Dominance:** Native long-context attention mechanisms significantly outperform external retrieval systems, which tend to discard subtle evolutionary signals; (2) **The Forgetting Trap:** Most systems suffer from severe catastrophic forgetting and "Forward Transfer Saturation" as dialogue complexity increases, failing to update outdated beliefs; (3) **The Oscillation Phenomenon:** High average accuracy often masks underlying volatility, where agents inconsistently flip between correct and incorrect answers across checkpoints; and (4) **Reasoning Limits:** While reasoning-enhanced models act as denoising filters for noisy retrieval, they encounter cognitive load thresholds in ultra-long contexts. The CoMem data construction pipeline and evaluation toolkit provide essential insights for developing next-generation personalized agents.

1 Introduction

The transition of Large Language Models (LLMs) from stateless query-response engines to lifelong personalized agents represents a paradigm shift in artificial intelligence. Agents deployed in domains such as healthcare (Wang et al., 2025), financial services (Batra et al., 2025), and digital companionship (Gao et al., 2025) must maintain a coherent and evolving internal model of the user. Unlike traditional systems bounded by short session windows, lifelong agents navigate non-monotonic data streams where user preferences shift (e.g., a dietary restriction changing from "omnivore" to "vegetarian") and context becomes saturated with obsolete information.

Although long-context modeling and Retrieval-Augmented Generation (RAG) (Lewis et al., 2020) have expanded token capacity, capacity does not equate to effective personalization. The core challenge lies in *curation*: distinguishing valid, evolving signals from "temporal noise." Current memory systems, often relying on semantic similarity, struggle to resolve conflicts between outdated high-confidence retrieval (e.g., "I love steak") and recent updates. Without robust reasoning to resolve these temporal conflicts, agents risk constructing inconsistent personas, eroding user trust.

Despite the criticality of this issue, existing benchmarks fail to capture these dynamics. While benchmarks like LoCoMo (Maharana et al., 2024) and Personamem (Jiang et al., 2025) evaluate static trait retention, they do not systematically assess how an agent's understanding fluctuates as a dialogue evolves. RAG-oriented benchmarks (e.g., RGB (Chen et al., 2023)) focus on retrieval fidelity rather than the longitudinal modeling of user state. More detailed discussion of memory system benchmarks and dynamics of user preference is in Appendix A.2.

To address these limitations, we introduce **CoMem**, a comprehensive benchmark for evaluating dynamic preference evolution and noise resilience. Unlike prior works that use static reading comprehension, CoMem evaluates memory sys-

Table 1: Comparison of memory benchmarks across user preference modeling, interaction settings, and evaluation dimensions.

Benchmark	Involves User Pref.	Primarily Pref.-Focused	Pref. Following	User–Assistant Dialogue	Multi-User Dialogue	>3 Participants	Continuous Mem. Eval.	Checkpoint Mem. Eval.
Personamem	✓	✓	✓	✓	×	×	×	×
Dialsim	✓	×	×	×	✓	✓	×	×
PrefEval	✓	✓	✓	✓	×	×	×	×
Evo-Memory	×	×	×	×	×	×	✓	×
MemBench	✓	×	×	✓	×	×	×	×
REALTALK	✓	×	×	×	✓	×	×	×
CoMem (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

tems at multiple checkpoints within a single continuous dialogue. Our contributions are:

- **Synthesized Multi-Session Dataset:** derived from Personamem (Jiang et al., 2025) and Dialsim (Kim et al., 2024), specifically curated via a two-stage LLM-verification pipeline to isolate preference evolution and entity tracking in both dyadic and multi-party settings.
- **Continual Learning Metrics:** We adapt Forgetting Measure (Lopez-Paz and Ranzato, 2017) and Forward Transfer (Chaudhry et al., 2018) to the dialogue domain, providing nuanced quantification of memory robustness beyond simple accuracy.
- **Benchmarking & Analysis:** We evaluate four memory systems across four LLMs. Results indicate that while context window expansion helps, most memory architectures exhibit high forgetting rates after the first dialogue quarter, highlighting the need for better memory update mechanisms.

2 Related Works

2.1 RAG Systems & Benchmarks

RAG enhances LLMs by retrieving relevant context. RAG benchmarks typically evaluate the system's ability to locate and utilize static knowledge. Benchmarks such as RGB (Chen et al., 2023) and CRUD-RAG (Lyu et al., 2024) focus on fundamental operations like retrieval precision and answer faithfulness. Further advancements like RAGAS (Es et al., 2024) and ARES (Saad-Falcon et al., 2024) introduce automated metrics for these qualities. Others address robustness: RECALL (Liu et al., 2023) and MIRAGE (Park et al., 2025) test resilience against counterfactual or noisy contexts. Critically, these benchmarks operate on *static snapshots* where ground truth is fixed (e.g., Wikipedia facts). They do not evaluate the *non-monotonic* nature of personalized agents, where a piece of information (e.g., a user's favorite food) may become invalid "temporal noise" as the dialogue progresses—a core focus of our work.

2.2 Memory Systems & Benchmarks

Recent memory architectures, such as Mem0 (Chhikara et al., 2025) and Memp (Fang et al., 2025), aim to manage the full lifecycle of memory (extraction, storage, deletion). Corresponding benchmarks have emerged to test these capabilities. The LTM Benchmark (Castillo-Bolado et al., 2024) and LongMemEval (Wu et al., 2024) assess general information retention over long contexts. However, these often resemble "needle-in-a-haystack" tasks rather than coherent persona tracking. Evo-Memory (Wei et al., 2025) introduces a task-stream evaluation for continuous learning, but it focuses on task completion (e.g., solving math problems in sequence) rather than the nuances of user personalization or social preference evolution.

2.3 User Preference & Multi-Party Dynamics

The closest precursors to CoMem are benchmarks focusing on persona and social dynamics. Personamem (Jiang et al., 2025) and PrefEval (Zhao et al., 2025) explicitly evaluate whether agents can follow user preferences. While Personamem provides high-quality dyadic interaction data, it treats evaluation as a static reading comprehension task over a finished dialogue, failing to measure *when* or *how* a system captures (or forgets) a preference change during the conversation. To evaluate resilience against social noise, benchmarks like Dialsim (Kim et al., 2024) and LoCoMo (Maharana et al., 2024) simulate multi-user interactions. While Dialsim effectively tests entity tracking (who said what), it does not explicitly benchmark the longitudinal evolution of user models.

CoMem bridges these gaps by synthesizing the

preference-centric focus of Personamem and the multi-party complexity of Dialsim into a *continual evaluation framework*. Unlike prior works, we assess performance at multiple checkpoints, utilizing metrics like Forgetting and Forward Transfer to quantify the stability of the user model over time. Table 1 shows the comparison between CoMem and multiple currently popular benchmarks.

3 CoMem

To rigorously evaluate memory systems under dynamic preference evolution, we introduce CoMem. The benchmark is constructed through a multi-stage pipeline designed to ensure that ground-truth answers are valid at specific temporal checkpoints within a dialogue.

3.1 Task Formulation

We formalize the memory evaluation as a checkpoint-based question answering task. Let $\mathcal{D} = \{u_1, a_1, u_2, a_2, \ldots, u_T, a_T\}$ represent a dialogue history of T turns. We define a set of normalized temporal checkpoints $\tau \in \{0.25, 0.50, 0.75, 1.0\}$. For a specific checkpoint τ, the accessible context is truncated to $C_\tau = \mathcal{D}_{1:\lfloor \tau \cdot T \rfloor}$. Given a question q and a set of candidate options $\mathcal{O}$, the memory system must select the correct option $o^* \in \mathcal{O}$ based solely on C_τ. Unlike standard QA where the context is static, CoMem evaluates the system's ability to map a continuous function $f(C_\tau, q) \rightarrow o^*_\tau$, where the ground truth o^*_τ may shift as τ increases.

3.2 Data Construction Pipeline

Our dataset is synthesized from two primary sources: *Personamem* (Jiang et al., 2025) for dyadic user-assistant interactions, and *Dialsim* (Kim et al., 2024) for multi-party social dynamics. Figure 1 shows the general evaluation procedure across our datasets. We then apply a four-step curation pipeline (Figure 2) to adapt these datasets for continual evaluation.

3.2.1 Context Standardization

Personamem (Dyadic): We utilize the 32k and 128k splits. Since Personamem contains explicit 'persona' profiles, we interleave these sparsely into the dialogue history to simulate natural disclosure.
Dialsim (Multi-Party): Dialsim consists of TV show scripts (e.g., *Friends*, *The Office*, *The Big Bang Theory*). To create long-context continuity, we concatenate scripts from the same season chronologically. We retain speaker tags to challenge the memory system's entity tracking capabilities (e.g., distinguishing between Ross's preferences and Joey's).

3.2.2 Temporal Checkpoint & Annotation

A critical challenge is that an answer valid at the end of a dialogue may be unknown or different at an earlier stage. We employ a Model-in-the-Loop (HITL) verification strategy to establish ground truth at each checkpoint:

Truncation: For each dialogue, we generate four truncated contexts corresponding to checkpoints $\tau \in \{0.25, 0.5, 0.75, 1.0\}$.

Round 1: Initial Annotation: We feed the truncated context C_τ and the original question q to GPT-5. The model is instructed to identify the answer *based strictly on visible information* or declare "Unknown."

Round 2: Consistency Verification: To mitigate hallucinations, the answers from Round 1 are verified by a second model (Claude 4.5 Sonnet) using a Chain-of-Thought prompt.

Filtering: We retain only those samples where both models achieve exact agreement on the answer key across all valid checkpoints. Questions with unstable or ambiguous reasoning paths are discarded.

A more detailed description of our dataset construction process and some notices for our dataset construction is in Appendix A.3.

3.3 Evaluation Metrics

We move beyond simple accuracy by adapting metrics from Continual Learning (CL) to the dialogue domain. We define accuracy at checkpoint n as A_n.

Forgetting Measure (F_n) This metric quantifies the loss of previously acquired knowledge. It specifically penalizes cases where a system correctly answers a question at an earlier stage but fails as new (potentially noisy) context is added. We define the forgetting at checkpoint n (where $n > 1$) as:

$$F_n = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}\left(\exists i < n : z_{q,i} = 1 \wedge z_{q,n} = 0\right)$$

where $\mathbb{I}$ is the indicator function and $z_{q,\tau} \in \{0, 1\}$ denote the binary correctness of question q at

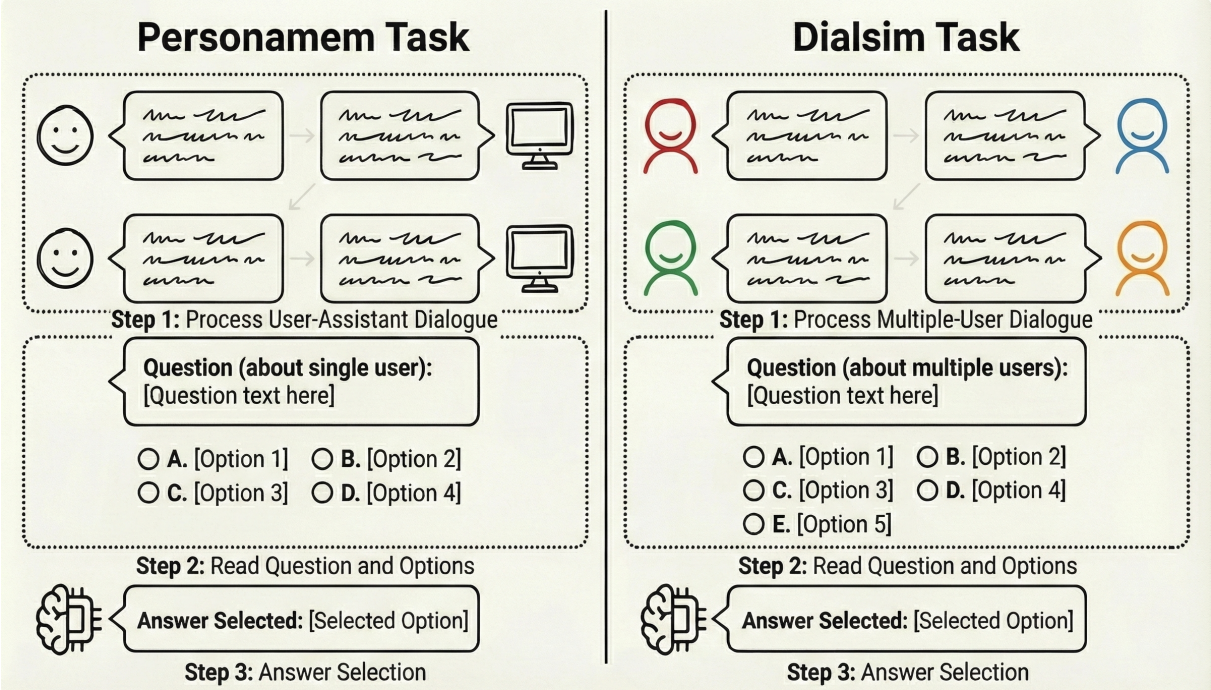

Figure 1: Task Procedure for Memory Systems on Personamem and Dialsim Datasets. Both task procedures include a 3 step frame from processing dialogue to generate answer. However, Personamem task procedure is under a user-assistant scenario, but Dialsim task procedure is under a multiple-user scenario with more than two users.

checkpoint τ (1 if correct, 0 otherwise). A high F_n indicates catastrophic forgetting, where new dialogue turns overwrite valid past memories.

Forward Transfer (FWT_n) This metric measures the system's ability to effectively utilize new information to resolve previously unanswerable or incorrectly answered questions.

$$FWT_n = \frac{1}{|Q|} \sum_{q \in Q} \mathbb{I}\left(\forall i < n : z_{q,i} = 0 \wedge z_{q,n} = 1\right)$$

A positive FWT score signifies effective learning: the system successfully extracted the necessary information from the dialogue segment between checkpoint $n-1$ and n.

Stability Ratio (SR) To capture more information, we also report the **All-Correct Ratio** (proportion of questions correctly answered at *all* checkpoints $0.25 \rightarrow 1.0$) and **All-Wrong Ratio**, highlighting the system's consistency.

3.4 Dataset Documentation and Statistics

We utilize data from Dialsim (Kim et al., 2024) and Personamem (Jiang et al., 2025) to construct our benchmark. Both are English datasets. Personamem contains LLM synthesized persona and dialogue information, while Dialsim contains a large amount of TV show scripts. After being processed and cleaned by our data processing pipeline, here the evaluation data includes:

Personamem 32k: 37 contexts, 200 questions, with an average context length of 26k tokens.

Personamem 128k: 110 contexts, 322 questions, with an average context length of 118.8k tokens.

Dialsim: 15 contexts, 723 questions, with an average context length of 120.5k tokens.

4 Experiments

4.1 Experimental Setup

Base Models: We evaluate four base models representing diverse capabilities and context windows: **GPT-4o-mini** (GPT, 2024), **Gemini-2.5-Flash** (Comanici et al., 2025), **DeepSeek-V3.2** (AI et al., 2025), and **DeepSeek-V3.2-think** (AI et al., 2025).

Memory Baselines: We compare four distinct memory architectures: **Full Memory** feeds the entire truncated dialogue history into the context window; **RAG Memory** (Tan et al., 2025) is a standard retrieval baseline using `text-embedding-3-large` model, which retrieves

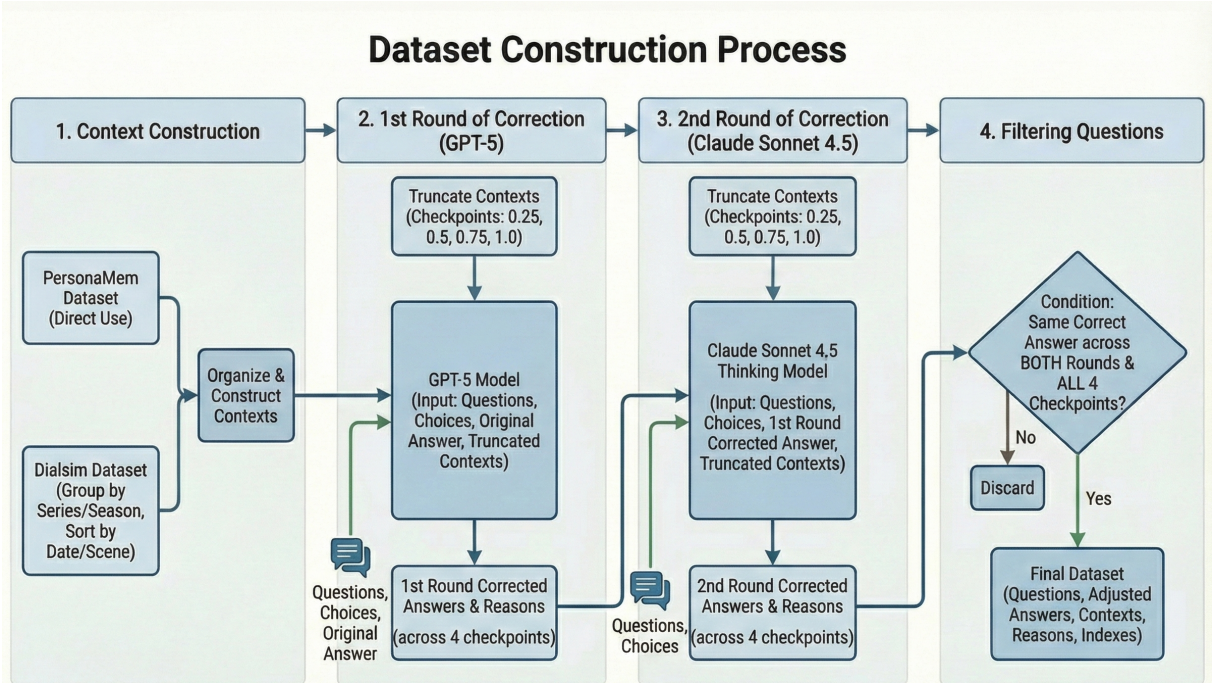

Figure 2: Dataset Construction Process. Here is a 4-step process from context construction to question filtering with 2 LLM based answer correction round in between. Notice that for Dialsim dataset we need to reconstruct scripts in it to form long dialogue contexts similar to those of Personamem dataset. Also, notice that we only keep questions with exactly the same answers in 2 rounds over ALL 4 checkpoints for our final benchmark dataset.

top-k chunks (chunk size=300 tokens) based on cosine similarity to the question; **Memp** (Fang et al., 2025) is an episodic memory system (Fang et al., 2025) that compresses history into concise summaries. **Mem0** (Chhikara et al., 2025) is a full-lifecycle memory system (Chhikara et al., 2025) that performs extraction, updates, and deletion. Notably, Mem0 results are reported for 32k and Dialsim due to its API rate limits on 128k. All generation is performed with temperature $T = 0$ to ensure reproducibility.

4.2 Main Results

We first analyze the aggregate accuracy across all checkpoints. Figure 3 summarizes the performance. Here we highlight the main insights, with more details in our Appendix A.4.

Native Long-Context Dominance: The combination of **Gemini-2.5-Flash + Full Memory** consistently achieves the highest performance across all datasets. This suggests that for tasks involving distributed preference cues, current external memory modules (RAG/Mem0) still lag behind the native attention mechanisms of state-of-the-art long-context models.

The "Context-Memory" Trade-off: For smaller models like **GPT-4o-mini**, external memory systems provide significant gains. As shown in the Personamem-128k task, RAG Memory outperforms Full Memory by an average of **4.5%** across checkpoints. However, for stronger models (DeepSeek-V3.2/Gemini-2.5-Flash), applying RAG or Memp often degrades performance. This indicates that aggressive retrieval pruning often discards subtle preference evolution signals that strong models could otherwise detect in full context.

4.3 Analysis I: Temporal Dynamics & Forgetting

A core contribution of CoMem is evaluating stability over time. We observe a concerning trend of **Catastrophic Forgetting** in memory systems.

Increasing Forgetting Rates: As shown in Table 3 and Table 2 (The same pattern holds for other tasks as well, see Appendix A.1 for full details), the Forgetting Measure (F_n) typically rises as the dialogue progresses. On Personamem-128k, the $F_{1.0}$ score for DeepSeek-V3.2 increases to **0.264**, meaning over 26% of questions correctly answered early in the conversation are answered incorrectly by the end. This confirms that current memory systems

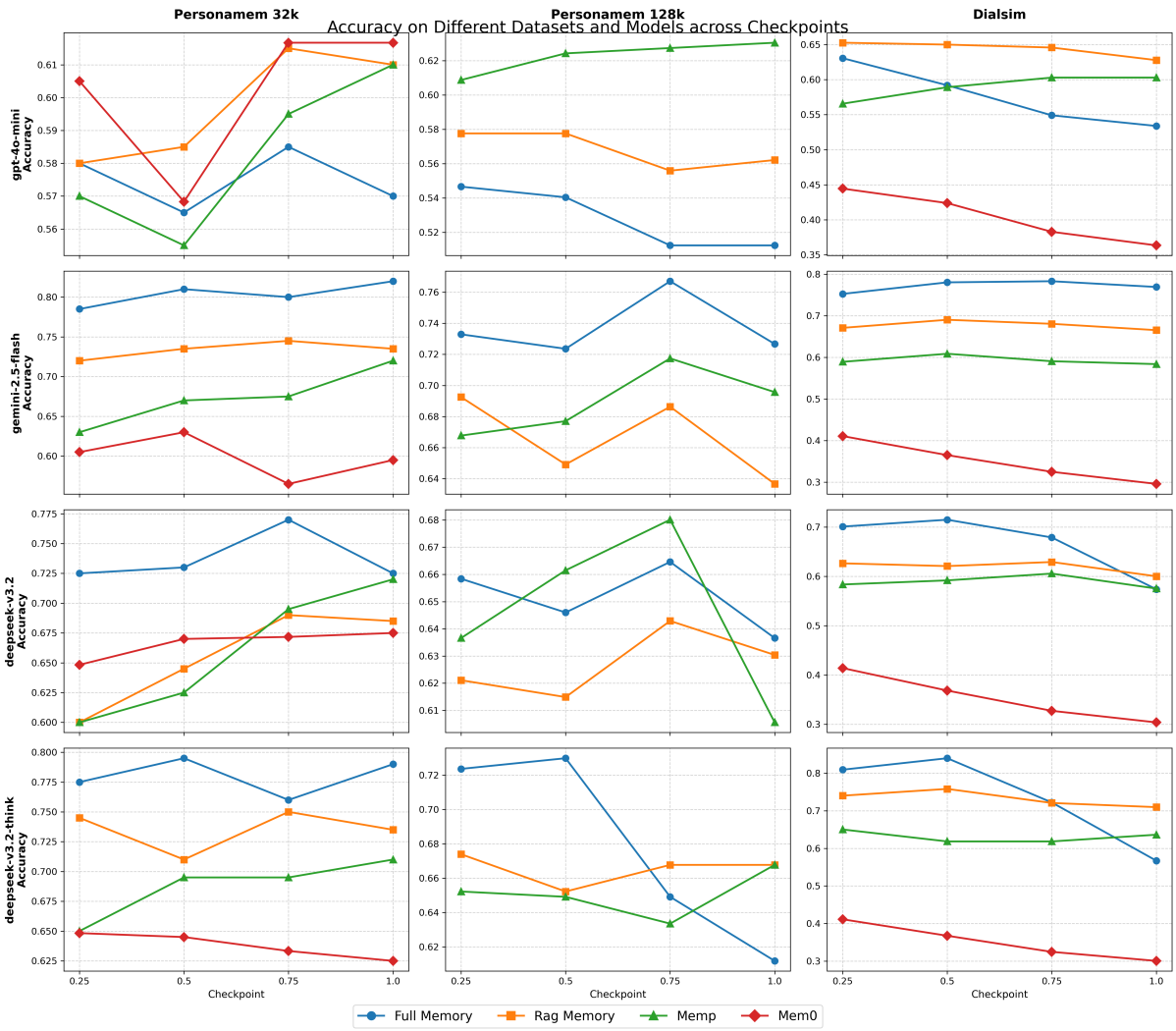

Figure 3: Accuracy Plots for all experiments. 12 figures here are accuracy results over all 4 LLMs GPT-4o-mini,Gemini-2.5-flash, Deepseek-v3.2 and Deepseek-v3.2-think AND on all 3 datasets Personamem 32k, Personamem 128k and Dialsim. Each figure corresponds to results with 4 memory systems: Full Memory, RAG Memory, Memp and Mem0 and over all 4 checkpoints.

struggle with **destructive interference**, where new noisy context overwrites correct prior beliefs.

Forward Transfer Saturation: Ideally, FWT_n should remain stable, indicating continuous learning. However, we observe a phenomenon of **"Context Saturation."** Forward Transfer is highest at Checkpoint 0.25 (> 0.15 on average) but drops sharply to < 0.05 after Checkpoint 0.5 across most systems. This implies that memory systems rapidly construct a coarse user profile in the first quarter of the dialogue but fail to effectively integrate finer-grained updates in later stages, treating them as noise rather than signal updates.

4.4 Analysis II: Scenario Resilience (Dialsim)

The Dialsim dataset introduces multi-party noise, revealing distinct failure modes.

Mem0's Attribution Failure: While Mem0 performs competitively on single-user tasks (i.e., the Personamem-32k), it suffers a catastrophic performance drop in Dialsim (see the first line in Figure 3)[1]. Error analysis reveals that Mem0's extraction module (powered by GPT-4o) frequently falls victim to **Entity Attribution Errors**—mistakenly attributing a preference expressed by Character A to Character B. This highlights that "User Prefer-

[1]During our experiments, we found that Mem0's memory extraction code is deeply coupled with openai models. Thus, when using other base models, Mem0's performance was not as good as expected.

Table 2: Forgetting Measure (F_n) on Personamem-128k. **High Forgetting** in later stages highlights system fragility.

Method	$F_{0.25}$	$F_{0.5}$	$F_{0.75}$	$F_{1.0}$
Model: GPT-4o-mini				
Full Memory	0.1273	0.1708	0.2236	0.3727
Rag Memory	0.1180	0.1646	0.2050	0.2174
Memp	0.1273	0.1677	0.2050	0.2422
Model: Gemini-2.5-flash				
Full Memory	0.1273	0.1957	0.1863	0.2391
Rag Memory	0.1304	0.2143	0.1988	0.2609
Memp	0.1522	0.1957	0.1832	0.2267
Model: deepseek-v3.2				
Full Memory	0.1584	0.2081	0.2267	0.2640
Rag Memory	0.1429	0.2174	0.2298	0.2578
Memp	0.1553	0.1957	0.2143	0.3137
Model: deepseek-v3.2-think				
Full Memory	0.1335	0.1770	0.2857	0.3323
Rag Memory	0.1553	0.2236	0.2267	0.2578
Memp	0.1522	0.2081	0.2484	0.2516

Table 3: Forward Transfer (FWT_n) on Personamem-128k. **Low Transfer** in later stages highlights system fragility.

Method	$FWT_{0.25}$	$FWT_{0.5}$	$FWT_{0.75}$	$FWT_{1.0}$
Model: GPT-4o-mini				
Full Memory	0.2112	0.0373	0.0248	0.0155
Rag Memory	0.2236	0.0466	0.0186	0.0186
Memp	0.2329	0.0559	0.0404	0.0404
Model: Gemini-2.5-flash				
Full Memory	0.3230	0.0590	0.0342	0.0124
Rag Memory	0.3137	0.0404	0.0217	0.0124
Memp	0.2888	0.0528	0.0280	0.0217
Model: deepseek-v3.2				
Full Memory	0.2298	0.0373	0.0373	0.0093
Rag Memory	0.1708	0.0683	0.0404	0.0155
Memp	0.2143	0.0652	0.0373	0.0248
Model: deepseek-v3.2-think				
Full Memory	0.2826	0.0497	0.0280	0.0093
Rag Memory	0.2329	0.0466	0.0186	0.0311
Memp	0.2298	0.0528	0.0248	0.0373

ence" memory is not isomorphic to "Multi-Party" memory; the latter requires rigorous entity tracking that current graph/vector stores lack.

DeepSeek's Context Limitation: We note that DeepSeek-V3.2 and DeepSeek-V3.2-think (context window $\approx$ 128k) show a sharp accuracy drop-off at Checkpoint 1.0 in Dialsim and Personamem-128k (which approaches their context limit). This reinforces the necessity of memory compression for infinite-context agents, as even strong models may fail when context saturation is reached.

4.5 Analysis III: The Role of Reasoning in Memory Coherence

A unique dimension of CoMem is the comparison between standard instruction-tuned models (**DeepSeek-V3.2**) and reasoning-enhanced models (**DeepSeek-V3.2-think**). We isolate the impact of Chain-of-Thought on preference consistency by analyzing the *All-Correct Ratio*.

Reasoning Mitigates Retrieval Noise. As shown in Table 4 (for more details, refer to Tables 13-15 in Appendix A.1), the reasoning model consistently achieves higher stability compared to its standard counterpart when using RAG. On Dialsim, DeepSeek-V3.2-think improves the RAG stability from **44.1%** to **50.4%**. We hypothesize that explicit reasoning acts as a **"Denoising Filter"**. RAG retrievers often return top-k chunks containing conflicting temporal information. While a standard model might hallucinate based on semantic overlap, the reasoning model uses CoT to analyze temporal markers within the chunks, logically resolving conflicts before generating an answer.

Table 4: Impact of Reasoning on Stability (All-Correct Ratio). DeepSeek-V3.2-think consistently acts as a **"Denoising Filter"** for RAG systems, improving consistency. However, it suffers under high cognitive load with Full Memory in complex and long-context scenarios (Dialsim and Personamem-128k).

Dataset	**System**	**DS-V3.2**	**DS-V3.2-thinking**	Δ
PM-32k	RAG	0.480	**0.540**	+6.0%
	Full	0.580	**0.660**	+8.0%
PM-128k	RAG	0.391	**0.422**	+3.1%
	Full	**0.422**	0.376	-4.6%
Dialsim	RAG	0.441	**0.504**	+6.3%
	Full	**0.448**	0.415	-3.3%

The Cognitive Load Threshold. However, reasoning is not a panacea. Table 4 also reveals a critical limitation: with Full Memory, DeepSeek-V3.2-think's performance drops by **3.3%** on Dialsim, and **4.6%** for Personamem-128k. This suggests a **Cognitive Load Threshold**. When the context becomes extremely long and complex (as in Dialsim's multi-party scripts), the computational overhead of performing deep reasoning over the *entire* context exceeds the model's effective attention span. This implies that for infinite-context agents and complex tasks, **Reasoning + Compressed Memory** (e.g., RAG/Memp) is a more viable path than Reasoning + Full Context.

Table 5: Stability Ratio (All-Correct) on Personamem-128k, acting as a stable anchor against retrieval jitter.

Model	Full Mem	RAG Mem	Memp
GPT-4o-mini	0.255	0.367	0.413
Gemini-2.5-flash	0.509	0.453	0.469
DS-V3.2	0.422	0.391	0.385
DS-V3.2-thinking	0.376	0.422	0.438

4.6 Analysis IV: Stability vs. Accuracy — The "Oscillation" Phenomenon

Standard benchmarks typically report average accuracy at specific snapshots. However, for a user-facing agent, *consistency* is paramount. We analyze this using the **Stability Ratio** (proportion of questions consistently answered correctly across all checkpoints).

The Illusion of High Accuracy. Comparing Figure 3 (Accuracy) with Table 5, we reveal a hidden instability. For example, while **GPT-4o-mini + Memp** maintains a respectable average accuracy of $\sim 60\%$ on Personamem-128k, its *All-Correct Ratio* is only **41.3%**, where the same pattern also holds for other datasets. This discrepancy indicates a severe **"Oscillation Phenomenon"**: the model is flipping its answers between checkpoints. A question answered correctly at Checkpoint 0.5 might be answered incorrectly at 0.75 due to index pollution, and then correctly again at 1.0 due to a lucky retrieval.

4.7 Summary of Insights

Synthesizing the experiments above, we derive four critical insights for designing next-generation memory systems:

- **Native vs. External:** For high-capacity models (e.g., Gemini-2.5-flash), full context remains superior to current retrieval methods. RAG is primarily beneficial for weaker or constrained models.
- **The Forgetting Trap:** Preference evolution is non-monotonic. Current systems lack a robust mechanism to "overwrite" outdated keys, leading to destructive interference as dialogue length increases.
- **Reasoning as a Buffer:** Enhanced reasoning capabilities can partially compensate for noisy retrieval by logically resolving conflicting memory chunks, but they cannot fix fundamental retrieval failures.
- **Stability over Accuracy:** Evaluating average accuracy masks memory volatility. Current retrieval or memory systems still exist with a relatively serious oscillation issue that urgently needs to be addressed.

5 Conclusion

In this paper, we introduced **CoMem**, a comprehensive benchmark designed to evaluate the capability of Long-Context Agents to maintain continual memory and adapt to dynamic preference evolution. Unlike prior benchmarks that focus on static retrieval fidelity, CoMem rigorously assesses an agent's ability to navigate non-monotonic data streams, distinguishing valid user updates from temporal noise across sequential dialogue checkpoints.

Our extensive experiments across diverse LLMs and memory architectures reveal a complex landscape for next-generation agents. We observe that native long-context models (e.g., Gemini-2.5-Flash) utilizing full history consistently outperform current external memory mechanisms (RAG, Mem0), suggesting that internal attention mechanisms are currently superior at capturing distributed preference cues than retrieval-based compression. However, a pervasive challenge remains: most systems exhibit significant catastrophic forgetting and "Forward Transfer Saturation" as dialogues progress, failing to effectively overwrite outdated beliefs with new information. Furthermore, our stability analysis exposes a critical "Oscillation Phenomenon," where agents inconsistently flip between correct and incorrect answers, proving that average accuracy masks underlying volatility in user modeling. While reasoning-enhanced models act as a partial denoising filter for retrieval systems, they face cognitive load thresholds in ultra-long contexts, indicating that reasoning alone is not a panacea for memory coherence.

Ultimately, CoMem demonstrates that context capacity does not equate to effective personalization. As we move from stateless query engines to lifelong companions, future research should pivot from simple information retention to dynamic memory consolidation—developing systems that can robustly update, forget, and stabilize user models over time. We hope the CoMem dataset and evaluation toolkit serve as a foundational standard for measuring progress toward these truly adaptive agents.

Limitations and Potential Societal Impacts

5.1 Limitations

While CoMem provides a robust framework, we acknowledge current limitations, including the reliance on specific backbone models for memory extraction and the exclusion of multi-step complex reasoning tasks. Future iterations will address these by integrating more diverse memory backbones and expanding the complexity of user queries.

5.2 Potential Societal Impacts

The development of agents capable of continual memory and personalized evolution introduces significant ethical considerations.

Privacy and Data Persistence: Unlike stateless models, long-context agents accumulate a granular history of user attributes. The "Oscillation Phenomenon" we observed suggests that deleted or updated preferences are not always reliably overwritten. This raises concerns about the *right to be forgotten*—if an agent cannot reliably forget outdated information, it may inadvertently leak sensitive past user states (e.g., medical conditions or political views) that the user intended to erase.

Echo Chambers and Manipulation: Agents that perfectly optimize for user preference evolution risk creating feedback loops. By consistently reinforcing a user's existing beliefs or preferences to maximize alignment, such systems could inadvertently construct "echo chambers," narrowing the user's information diversity or potentially being exploited for personalized persuasion. Future research must investigate guardrails that balance personalization with objectivity.

References

2024. Gpt-4o mini: advancing cost-efficient intelligence.

Deepseek AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, and some other authors. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Devesh Batra, Conor B. Hamill, John Hartley, Ramin Okhrati, Dale Seddon, Harvey Miller, Raad Khraishi, and Greig A. Cowan. 2025. A review of llm agent applications in finance and banking.

David Castillo-Bolado, Joseph Davidson, Finlay Gray, and Marek Rosa. 2024. Beyond prompts: Dynamic conversational benchmarking of large language models. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, Vancouver, Canada.

Arslan Chaudhry, Puneet K Dokania, Thalaiyasingam Ajanthan, and Philip HS Torr. 2018. Riemannian walk for incremental learning: Understanding forgetting and intransigence. In *Proceedings of the European conference on computer vision (ECCV)*, pages 532–547.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2023. Benchmarking large language models in retrieval-augmented generation. *Preprint*, arXiv:2309.01431v2.

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

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, and some other authors. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, page 150–158, St. Julians, Malta. Association for Computational Linguistics.

Runnan Fang, Yuan Liang, Xiaobin Wang, Jialong Wu, Shuofei Qiao, Pengjun Xie, Fei Huang, Huajun Chen, and Ningyu Zhang. 2025. Memp: Exploring agent procedural memory. *Preprint*, arXiv:2508.06433v2.

Xian Gao, Zongyun Zhang, Ting Liu, and Yuzhuo Fu. 2025. Onlinemate: An llm-based multi-agent companion system for cognitive support in online learning. *Preprint*, arXiv:2509.14803v2.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, page 13584–13606, Singapore. Association for Computational Linguistics.

Bowen Jiang, Zhuoqun Hao, Young Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo Jose

Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. In *Proceedings of the 2nd Conference on Language Modeling (COLM 2025)*, Montreal, Canada.

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

Dong-Ho Lee, Adyasha Maharana, Jay Pujara, Xiang Ren, and Francesco Barbieri. 2025. Realtalk: A 21-day real-world dataset for long-term conversation. *Preprint*, arXiv:2502.13270.

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

Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A benchmark for llms robustness against external counterfactual knowledge. *Preprint*, arXiv:2311.08147.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. *Advances in neural information processing systems*, 30.

Yuanjie Lyu, Zhiyu Li, Simin Niu, Feiyu Xiong, Bo Tang, Wenjin Wang, Hao Wu, Huanyong Liu, Tong Xu, and Enhong Chen. 2024. Crud-rag: A comprehensive chinese benchmark for retrieval-augmented generation of large language models. *Preprint*, arXiv:2401.17043v3.

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

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. Mirage: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, page 2883–2900, Albuquerque, New Mexico. Association for Computational Linguistics.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Processings of Neural Information Processing Systems 37 (NeurIPS 2024)*.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, page 338–354, Mexico City, Mexico. Association for Computational Linguistics.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 19336–19352, Vienna, Austria. Association for Computational Linguistics.

Yixuan Tang and Yi Yang. 2024. Multihop-rag: Benchmarking retrieval-augmented generation for multi-hop queries. *Preprint*, arXiv:2401.15391.

Wenxuan Wang, Zizhan Ma, Zheng Wang, Chenghan Wu, Jiaming Ji, Wenting Chen, Xiang Li, and Yixuan Yuan. 2025. A survey of llm-based agents in medicine: How far are we from baymax? *Preprint*, arXiv:2502.11211v2.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. 2025. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory. *Preprint*, arXiv:2511.20857.

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

Zhikun Xu, Yinghui Li, Ruixue Ding, Xinyu Wang, Boli Chen, Yong Jiang, Haitao Zheng, Wenlian Lu, Pengjun Xie, and Fei Huang. 2025. Let llms take on the latest challenges! a chinese dynamic question answering benchmark. In *Proceedings of the 31st International Conference on Computational Linguistics*, page 10435–10448, Abu Dhabi, UAE. Association for Computational Linguistics.

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do llms recognize your preferences? evaluating personalized preference following in llms. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, Singapore.

A Appendix

A.1 Part 1: Additional Experiment Results

Method	ckpt 0.25	ckpt 0.5	ckpt 0.75	ckpt 1
Model: GPT-4o-mini				
Full Memory	0.1100	0.1500	0.1450	0.1750
Rag Memory	0.1300	0.1450	0.1400	0.1550
Memp	0.1350	0.1900	0.1700	0.1650
Mem0	0.0900	0.1650	0.1700	0.1850
Model: Gemini-2.5-flash				
Full Memory	0.1150	0.1350	0.1450	0.1300
Rag Memory	0.1150	0.1700	0.1650	0.1750
Memp	0.1750	0.1800	0.1850	0.1500
Mem0	0.1100	0.1650	0.2600	0.2450
Model: deepseek-v3.2				
Full Memory	0.1150	0.1400	0.1000	0.1500
Rag Memory	0.1450	0.1650	0.1400	0.1650
Memp	0.1800	0.2050	0.1550	0.1400
Mem0	0.0750	0.1550	0.1450	0.1800
Model: deepseek-v3.2-think				
Full Memory	0.0850	0.1100	0.1550	0.1350
Rag Memory	0.0850	0.1650	0.1500	0.1750
Memp	0.1450	0.1800	0.2050	0.1900
Mem0	0.0950	0.1850	0.2100	0.2450

Table 6: Personamem 32k: Forgetting Measure Results.

Method	ckpt 0.25	ckpt 0.5	ckpt 0.75	ckpt 1
Model: GPT-4o-mini				
Full Memory	0.1550	0.0250	0.0150	0.0150
Rag Memory	0.1800	0.0200	0.0250	0.0100
Memp	0.1900	0.0400	0.0200	0.0100
Mem0	0.1950	0.0400	0.0500	0.0100
Model: Gemini-2.5-flash				
Full Memory	0.3200	0.0450	0.0000	0.0050
Rag Memory	0.2550	0.0700	0.0050	0.0000
Memp	0.2150	0.0450	0.0100	0.0100
Mem0	0.3050	0.0800	0.0300	0.0150
Model: deepseek-v3.2				
Full Memory	0.2250	0.0300	0.0000	0.0050
Rag Memory	0.1550	0.0650	0.0200	0.0200
Memp	0.1750	0.0500	0.0200	0.0100
Mem0	0.2950	0.1000	0.0150	0.0100
Model: deepseek-v3.2-think				
Full Memory	0.1900	0.0450	0.0100	0.0100
Rag Memory	0.2150	0.0450	0.0250	0.0100
Memp	0.1650	0.0800	0.0250	0.0000
Mem0	0.3500	0.0800	0.0250	0.0250

Table 7: Personamem 32k: Forward Transfer Results.

Method	ckpt 0.25	ckpt 0.5	ckpt 0.75	ckpt 1
Model: GPT-4o-mini				
Full Memory	0.0705	0.1618	0.2365	0.2725
Rag Memory	0.0512	0.1065	0.1425	0.1687
Memp	0.0913	0.1231	0.1508	0.1757
Mem0	0.0387	0.0830	0.1397	0.1784
Model: Gemini-2.5-flash				
Full Memory	0.0816	0.1065	0.1300	0.1632
Rag Memory	0.0871	0.1134	0.1508	0.1881
Memp	0.1286	0.1604	0.2019	0.2213
Mem0	0.0097	0.0609	0.1024	0.1314
Model: deepseek-v3.2				
Full Memory	0.0456	0.0719	0.1383	0.2531
Rag Memory	0.0581	0.1189	0.1466	0.1867
Memp	0.0871	0.1591	0.1964	0.2434
Mem0	0.0138	0.0705	0.1134	0.1411
Model: deepseek-v3.2-think				
Full Memory	0.1079	0.1162	0.2448	0.4053
Rag Memory	0.1231	0.1535	0.2102	0.2310
Memp	0.1425	0.2089	0.2296	0.2296
Mem0	0.0111	0.0664	0.1093	0.1411

Table 8: Dialsim: Forgetting Measure Results.

Method	ckpt 0.25	ckpt 0.5	ckpt 0.75	ckpt 1
Model: GPT-4o-mini				
Full Memory	0.3154	0.0526	0.0318	0.0207
Rag Memory	0.3181	0.0526	0.0318	0.0083
Memp	0.2877	0.0553	0.0415	0.0249
Mem0	0.1342	0.0263	0.0152	0.0180
Model: Gemini-2.5-flash				
Full Memory	0.3790	0.0526	0.0263	0.0194
Rag Memory	0.2988	0.0456	0.0277	0.0221
Memp	0.2365	0.0512	0.0235	0.0124
Mem0	0.1245	0.0055	0.0014	0.0000
Model: deepseek-v3.2				
Full Memory	0.4025	0.0401	0.0304	0.0097
Rag Memory	0.3430	0.0553	0.0360	0.0111
Memp	0.3402	0.0802	0.0512	0.0166
Mem0	0.1314	0.0083	0.0028	0.0041
Model: deepseek-v3.2-think				
Full Memory	0.3458	0.0387	0.0111	0.0055
Rag Memory	0.3043	0.0484	0.0194	0.0097
Memp	0.2393	0.0346	0.0207	0.0180
Mem0	0.1286	0.0083	0.0028	0.0041

Table 9: Dialsim: Forward Transfer Results.

	ckpt 0.25		ckpt 0.5		ckpt 0.75		ckpt 1	
Method	Forget	FWT	Forget	FWT	Forget	FWT	Forget	FWT
Personamem 32k, GPT-4o-mini								
Full Memory	0.1100	0.1550	0.1500	0.0250	0.1450	0.0150	0.1750	0.0150
Rag Memory	0.1300	0.1800	0.1450	0.0200	0.1400	0.0250	0.1550	0.0100
Memp	0.1350	0.1900	0.1900	0.0400	0.1700	0.0200	0.1650	0.0100
Mem0	0.0900	0.1950	0.1650	0.0400	0.1700	0.0500	0.1850	0.0100
Personamem 32k, Gemini-2.5-flash								
Full Memory	0.1150	0.3200	0.1350	0.0450	0.1450	0.0000	0.1300	0.0050
Rag Memory	0.1150	0.2550	0.1700	0.0700	0.1650	0.0050	0.1750	0.0000
Memp	0.1750	0.2150	0.1800	0.0450	0.1850	0.0100	0.1500	0.0100
Mem0	0.1100	0.3050	0.1650	0.0800	0.2600	0.0300	0.2450	0.0150
Personamem 32k, deepseek-v3.2								
Full Memory	0.1150	0.2250	0.1400	0.0300	0.1000	0.0000	0.1500	0.0050
Rag Memory	0.1450	0.1550	0.1650	0.0650	0.1400	0.0200	0.1650	0.0200
Memp	0.1800	0.1750	0.2050	0.0500	0.1550	0.0200	0.1400	0.0100
Mem0	0.0750	0.2950	0.1550	0.1000	0.1450	0.0150	0.1800	0.0100
Personamem 32k, deepseek-v3.2-think								
Full Memory	0.0850	0.1900	0.1100	0.0450	0.1550	0.0100	0.1350	0.0100
Rag Memory	0.0850	0.2150	0.1650	0.0450	0.1500	0.0250	0.1750	0.0100
Memp	0.1450	0.1650	0.1800	0.0800	0.2050	0.0250	0.1900	0.0000
Mem0	0.0950	0.3500	0.1850	0.0800	0.2100	0.0250	0.2450	0.0250

Table 10: Personamem 32k Forgetting Measure and Forward Transfer. This is forgetting measure AND forward transfer for all memory systems Full, RAG, Memp and Mem0 and all 4 models GPT-4o-mini,Gemini-2.5-flash,Deepseek-v3.2 and Deepseek-v3.2-think ON DATASET Personamem 32k over ALL 4 checkpoints.

	ckpt 0.25		ckpt 0.5		ckpt 0.75		ckpt 1	
Method	Forget	FWT	Forget	FWT	Forget	FWT	Forget	FWT
Personamem 128k, GPT-4o-mini								
Full Memory	0.1273	0.2112	0.1708	0.0373	0.2236	0.0248	0.3727	0.0155
Rag Memory	0.1180	0.2236	0.1646	0.0466	0.2050	0.0186	0.2174	0.0186
Memp	0.1273	0.2329	0.1677	0.0559	0.2050	0.0404	0.2422	0.0404
Mem0	–	–	–	–	–	–	–	–
Personamem 128k, Gemini-2.5-flash								
Full Memory	0.1273	0.3230	0.1957	0.0590	0.1863	0.0342	0.2391	0.0124
Rag Memory	0.1304	0.3137	0.2143	0.0404	0.1988	0.0217	0.2609	0.0124
Memp	0.1522	0.2888	0.1957	0.0528	0.1832	0.0280	0.2267	0.0217
Mem0	–	–	–	–	–	–	–	–
Personamem 128k, deepseek-v3.2								
Full Memory	0.1584	0.2298	0.2081	0.0373	0.2267	0.0373	0.2640	0.0093
Rag Memory	0.1429	0.1708	0.2174	0.0683	0.2298	0.0404	0.2578	0.0155
Memp	0.1553	0.2143	0.1957	0.0652	0.2143	0.0373	0.3137	0.0248
Mem0	–	–	–	–	–	–	–	–
Personamem 128k, deepseek-v3.2-think								
Full Memory	0.1335	0.2826	0.1770	0.0497	0.2857	0.0280	0.3323	0.0093
Rag Memory	0.1553	0.2329	0.2236	0.0466	0.2267	0.0186	0.2578	0.0311
Memp	0.1522	0.2298	0.2081	0.0528	0.2484	0.0248	0.2516	0.0373
Mem0	–	–	–	–	–	–	–	–

Table 11: Personamem 128k Forgetting Measure and Forward Transfer. This is forgetting measure AND forward transfer for 3 memory systems Full, RAG and Memp and all 4 models GPT-4o-mini,Gemini-2.5-flash,Deepseek-v3.2 and Deepseek-v3.2-think ON DATASET Personamem 128k over ALL 4 checkpoints.

	ckpt 0.25		ckpt 0.5		ckpt 0.75		ckpt 1	
Method	Forget	FWT	Forget	FWT	Forget	FWT	Forget	FWT
Dialsim, GPT-4o-mini								
Full Memory	0.0705	0.3154	0.1618	0.0526	0.2365	0.0318	0.2725	0.0207
Rag Memory	0.0512	0.3181	0.1065	0.0526	0.1425	0.0318	0.1687	0.0083
Memp	0.0913	0.2877	0.1231	0.0553	0.1508	0.0415	0.1757	0.0249
Mem0	0.0387	0.1342	0.0830	0.0263	0.1397	0.0152	0.1784	0.0180
Dialsim, Gemini-2.5-flash								
Full Memory	0.0816	0.3790	0.1065	0.0526	0.1300	0.0263	0.1632	0.0194
Rag Memory	0.0871	0.2988	0.1134	0.0456	0.1508	0.0277	0.1881	0.0221
Memp	0.1286	0.2365	0.1604	0.0512	0.2019	0.0235	0.2213	0.0124
Mem0	0.0097	0.1245	0.0609	0.0055	0.1024	0.0014	0.1314	0.0000
Dialsim, deepseek-v3.2								
Full Memory	0.0456	0.4025	0.0719	0.0401	0.1383	0.0304	0.2531	0.0097
Rag Memory	0.0581	0.3430	0.1189	0.0553	0.1466	0.0360	0.1867	0.0111
Memp	0.0871	0.3402	0.1591	0.0802	0.1964	0.0512	0.2434	0.0166
Mem0	0.0138	0.1314	0.0705	0.0083	0.1134	0.0028	0.1411	0.0041
Dialsim, deepseek-v3.2-think								
Full Memory	0.1079	0.3458	0.1162	0.0387	0.2448	0.0111	0.4053	0.0055
Rag Memory	0.1231	0.3043	0.1535	0.0484	0.2102	0.0194	0.2310	0.0097
Memp	0.1425	0.2393	0.2089	0.0346	0.2296	0.0207	0.2296	0.0180
Mem0	0.0111	0.1286	0.0664	0.0083	0.1093	0.0028	0.1411	0.0041

Table 12: Dialsim Forgetting Measure and Forward Transfer. This is forgetting measure AND forward transfer for all memory systems Full, RAG, Memp and Mem0 and all 4 models GPT-4o-mini,Gemini-2.5-flash,Deepseek-v3.2 and Deepseek-v3.2-think ON DATASET Dialsim over ALL 4 checkpoints.

Table 13: All Correct/Wrong ratio of questions over 4 checkpoints on Personamem 32k dataset. Results for all 4 LLMs GPT-4o-mini, Gemini-2.5-flash, Deepseek-v3.2 and Deepseek-v3.2-think and 4 memory systems Full, RAG, Mem0 and Memp.

Method	**GPT-4o-mini**		**Gemini-2.5-flash**		**DeepSeek-V3.2**		**DeepSeek-V3.2-think**	
	All Correct	All Wrong	All Correct	All Wrong	All Correct	All Wrong	All Correct	All Wrong
Full Memory	0.430	0.285	0.635	0.075	0.580	0.155	0.660	0.110
RAG Memory	0.440	0.270	0.585	0.130	0.480	0.175	0.540	0.120
Memp	0.405	0.255	0.500	0.180	0.425	0.170	0.510	0.145
Mem0	0.420	0.245	0.385	0.215	0.470	0.190	0.430	0.180

Table 14: All Correct/Wrong ratio of questions over 4 checkpoints on Personamem 128k dataset. Results for all 4 LLMs GPT-4o-mini, Gemini-2.5-flash, Deepseek-v3.2 and Deepseek-v3.2-think and 3 memory systems Full, RAG and Memp.

Method	**GPT-4o-mini**		**Gemini-2.5-flash**		**DeepSeek-V3.2**		**DeepSeek-V3.2-think**	
	All Correct	All Wrong	All Correct	All Wrong	All Correct	All Wrong	All Correct	All Wrong
Full Memory	0.2547	0.2857	0.5093	0.0714	0.4224	0.1366	0.3758	0.0807
RAG Memory	0.3665	0.2422	0.4534	0.1460	0.3913	0.1553	0.4224	0.1149
Memp	0.4130	0.1677	0.4689	0.1087	0.3851	0.1211	0.4379	0.1429
Mem0	–	–	–	–	–	–	–	–

Table 15: All Correct/Wrong ratio of questions over 4 checkpoints on Dialsim dataset. Results for all 4 LLMs GPT-4o-mini, Gemini-2.5-flash, Deepseek-v3.2 and Deepseek-v3.2-think and 4 memory systems Full, RAG, Mem0 and Memp.

Method	**GPT-4o-mini**		**Gemini-2.5-flash**		**DeepSeek-V3.2**		**DeepSeek-V3.2-think**	
	All Correct	All Wrong	All Correct	All Wrong	All Correct	All Wrong	All Correct	All Wrong
Full Memory	0.3900	0.2448	0.5602	0.0830	0.4481	0.1770	0.4149	0.0526
RAG Memory	0.4952	0.2282	0.4993	0.1770	0.4412	0.2310	0.5035	0.0899
Memp	0.4260	0.2614	0.4205	0.2434	0.3541	0.2047	0.4329	0.1881
Mem0	0.2863	0.4689	0.2683	0.5726	0.2683	0.5560	0.2683	0.5602

A.2 Part 2: Full Explanation of Benchmarks and Dynamics of User Preferences

Current benchmarks related to user preference span a spectrum: some long-term memory evaluations do not involve user preference at all, such as Evo-Memory (Wei et al., 2025) and Conversation Chronicles (Jang et al., 2023), whereas others test whether memory systems or agents can represent user- or persona-level likes, dislikes, and traits, including MemBench(Tan et al., 2025), PerLTQA (Du et al., 2024), LTM Benchmark (Castillo-Bolado et al., 2024), LoCoMo (Maharana et al., 2024), Dialsim (Kim et al., 2024), LongMemEval (Wu et al., 2024), REALTALK (Lee et al., 2025), PrefEval (Zhao et al., 2025), and Personamem (Jiang et al., 2025). Among these, only PrefEval (Zhao et al., 2025) and Personamem (Jiang et al., 2025) directly evaluate user preference following, i.e., whether system outputs comply with stated or inferred preferences over time. By contrast, RAG-oriented benchmarks such as RGB (Chen et al., 2023), MultiHop-RAG (Tang and Yang, 2024), RECALL (Liu et al., 2023), MIRAGE(Park et al., 2025) , CDQA (Xu et al., 2025), CRUD-RAG (Lyu et al., 2024), RAGChecker (Ru et al., 2024), ARES (Saad-Falcon et al., 2024), and RAGAS (Es et al., 2024) focus on retrieval quality, robustness to noisy or counterfactual context, and answer faithfulness but NOT explicit evaluation of user preference modeling or preference-following behavior.

A.3 Part 3: Detailed Steps and Important Notices for dataset construction

A.3.1 Detailed Steps

Context Construction We extract our context from Personamem or Dialsim dataset. We directly use contexts from Personamem, but for Dialsim, we group scripts from the same TV series and season, then sort them by date and scene order to construct our Dialsim context files.

1st Round of Correction For each context, we identify the indices corresponding to the 0.25, 0.5, 0.75, and 1.0 checkpoints and create truncated versions of the context at each of these indices. For each question and each checkpoint, we input the question, corresponding choices, original answer, and the truncated context corresponding to the chosen checkpoint to the GPT-5 model, and ask the model to find the correct answer of the question , check if original answer is correct or not and provide reasons at the given checkpoint based on given information.

2nd Round of Correction For each context we repeat the truncation process above. For each question and each checkpoint, we input the question, corresponding choices, correct answer generated in 1st round, and the truncated context corresponding to the chosen checkpoint to the Claude Sonnet 4.5 thinking model, and ask the model to find the correct answer of the question, and check if the 1st round correct answer is correct or not and provide reasons at the given checkpoint based on given information.

Filtering Questions For questions from Personamem 32k, Personamem 128k, and Dialsim Dataset, we choose the ones with same correct answers across both correction rounds above and all 4 checkpoints. We collect these questions, their adjusted answers after 2 correction rounds at different checkpoints, and the corresponding contexts and the reasons behind choosing the correct answers and the truncation indexes at each checkpoint to construct our dataset.

Finally, for checkpoint 0 and for each question, we adopt the final corrected answer for checkpoint 1. Now for all 5 checkpoints, we have same group of questions but with possibly different correct answers.

A.3.2 Important Notices

About the chunk size for dialogue records For the chunks or messages in each of dialogue records, we shall control the chunk size of messages. For Personamem, the chunk size of 250-350 tokens is fine, but for Dialsim, with scripts up to thousands of tokens, we need to further split the scripts in each season of each TV series into chunks of approximately 300 tokens. When it comes to Dialsim dataset, notice that for Full Memory system, we do not need to further split chunks, for RAG memory system, we need OVERLAPPING chunks of token size 300 to avoid truncating important information. Also, when it comes to Personamem dataset, we need to group neighboring user and assistant messages together for RAG memory system. For the rest of memory systems, however, such post processing steps are not necessary, and we only need to make sure that we fed the non overlapping 300 token size Dialsim chunks to them.

Notices for Data Construction During data cleaning and construction process above, the data

fed into the cleaning LLM prompt should be identical to the information used by previous researchers when testing on the same dataset. For example, when cleaning Personamem datasets as above, researchers shall input the explicit PERSONA information into the cleaning LLM prompt as well as other messages from contexts, since most researchers keep the PERSONA information in their tests. Ignoring such information might result in less accurate answer and repeated effort to clean the data, resulting in lots of extra expenses in time and money. Because of the aforementioned reasons, the cost of data construction for our benchmark is as high as around 20000 dollars.

A.4 Part 4: Detailed discussion of Experiment Main Results

As we can see from the Figure 3, the combination of Gemini-2.5-flash model and Full Memory has best performance across all 3 datasets, with a range of accuracy between 0.72 and 0.77 over the Personamem 128k dataset, which has highest average context length. Also, the accuracy of the same combination is considerably higher than almost any other combinations of memory systems and models over Personamem 32k and Dialsim dataset. When we take into account the tables of forward transfer (Table 7,3,9) and forgetting measures (Table 6,2,8), we find that the combination of Gemini-2.5-flash model and Full Memory still has best performance. For most other combinations of models and memory systems, their forgetting measure at checkpoint 1.0, especially over long context datasets Personamem 128k, as shown in Table 2, are large and over 25 percent (Personamem 128k), but for Gemini-2.5-flash model and Full Memory, this value is 23.91 percent, only higher than combination of Rag memory and GPT-4o-mini model, and combination of Memp and Gemini-2.5-flash. Also, Gemini-2.5-flash model has best performance over 2 Personamem datasets. The overall range of accuracy between 0.56 and 0.82 for Personamem 32k dataset and range of accuracy between 0.62 and 0.77 for Personamem 128k dataset, which are significantly higher than the performance of any other models over the 2 datasets. When it comes to ratio of all correctly answered/ all incorrectly answered questions over 4 checkpoints (0.25 to 1), Table 13,14,15, the combination of Gemini-2.5-flash model and Full Memory system has almost the highest ratio of all correctly answered questions and lowest ratio of all incorrectly answered questions. In summary, best combination of LLM and memory system is Gemini-2.5-flash and Full Memory, and best LLM is Gemini-2.5-flash. The memory system with worst performance is Mem0 as shown in Figure 3 and table 13,14,15.

Also, notice that when it comes to Personamem 32k, when the fundamental capability of the backbone model is strong enough, the full memory system has highest accuracy for Gemini-2.5-flash, Deepseek v3.2 and Deepseek v3.2-think model over the Personamem 32k dataset.

One noticeable fact is that as the context added to memory system + model combination increases in length, the corresponding forgetting measure also increases, as shown in Table 2 and Table 8. For example, in Table 2, for combination of Gemini-2.5-flash model and Memp, forgetting measure increases from 0.1522 to 0.2267. The possible reason behind this phenomenon is that current memory systems can only handle dialogues at or below a certain length, and as the input context exceeds this length, these memory systems might start to forget previously added dialogue information. One evidence is that the aforementioned increasing trend for forgetting measure does not appear in Table 6, forgetting measure and forward transfer over Personamem 32k dataset.

Also, another notable fact is that across the three datasets, as shown in Table 7,3,9, significant forward transfer is observed only at the 0.25 checkpoint, where most value are greater than 0.2, with many even exceeding 0.2. After the 0.25 checkpoint, most forward transfer values drop below 0.07. This suggests that most memory systems do not learn much knowledge about users after checkpoint 0.25. The reason might be similar to above, which is that memory systems have upper limits for the amount of dialogue information they can extract and summarize accurately.

As suggested by some notable facts above, memory systems might have an upper limit for the amount of information they can extract, summarize and memorize accurately.