MEMORYCD: Benchmarking Long-Context User Memory of LLM Agents for Lifelong Cross-Domain Personalization

Anonymous ACL submission

Abstract

Recent advancements in Large Language Models (LLMs) have expanded context windows to million-token scales, yet benchmarks for evaluating memory remain limited to short-session synthetic dialogues. We introduce MEMORYCD, the first large-scale, user-centric, cross-domain memory benchmark derived from life-long real-world behaviors in the Amazon Review dataset. Unlike existing memory datasets that rely on scripted personas to generate synthetic user data, MEMORYCD tracks authentic user interactions across years and multiple domains. We construct a multi-faceted long-context memory evaluation pipeline of 14 state-of-the-art LLM base models with 6 memory method baselines on 4 distinct personalization tasks over 12 diverse domains to evaluate an agent's ability to simulate real user behaviors in both single and cross-domain settings. Our analysis reveals that existing memory methods are far from user satisfaction in various domains, offering the first testbed for cross-domain life-long personalization evaluation.

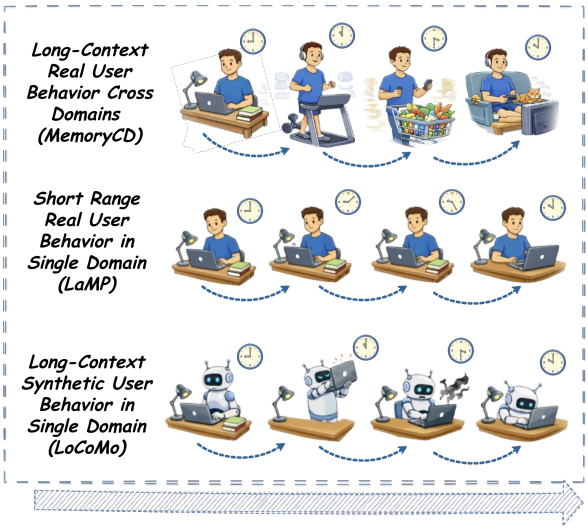

Figure 1: Comparison of user memory benchmarks: **MemoryCD (ours)** captures cross-domain real-user activities over long time horizons. **LaMP** (Salemi et al., 2024) focuses only on the short-term single-domain user behaviors, while **LoCoMo** (Maharana et al., 2024) represents non-authentic LLM-simulated user behaviors.

1 Introduction

The evolution of Large Language Models (LLMs) has ushered in a new era of general intelligence, with LLM-based agents demonstrating remarkable proficiency across standardized reasoning and generation tasks (Naveed et al., 2025; Luo et al., 2025; Zhang et al., 2025b). However, these advancements do not automatically translate into genuine user satisfaction (Zhang et al., 2025e). For these agents to transition from generic tools to genuine personal assistants, they must possess the ability to retain, organize, and utilize the historical context of the user (Cai et al., 2025; Chen et al., 2024). Serving a real user requires more than responding to a single prompt; it requires grounding decisions in a coherent and persistent modeling of the user's history, preferences, and evolving goals (Tan et al., 2025b; Zhang et al., 2025e). Memory provides a principled mechanism for enabling user-specific personalization in LLM agents (Zhang et al., 2025f). As interactions accumulate over time, effective agents must consolidate long-horizon behavioral traces into abstract and durable user memories that support consistent personalization, improved task completion, and more natural user–agent interactions (Sun et al., 2025; Yang et al., 2024; Zhang et al., 2025e). This capability is critical for adapting to individual users, enabling agents to effectively support their work and daily life as lifelong lifelong personal assistant.

Despite its importance, current memory benchmarks fall short of capturing user preferences and evaluating user satisfaction in real-world scenarios (as shown in Figure 1 and Table 1). First, the majority of memory benchmarks (Maharana et al., 2024; Ai et al., 2025; Tan et al., 2025a; Zhao et al., 2025; Wu et al., 2025a; Bei et al., 2026) are synthetic and constructed from LLM simulation that

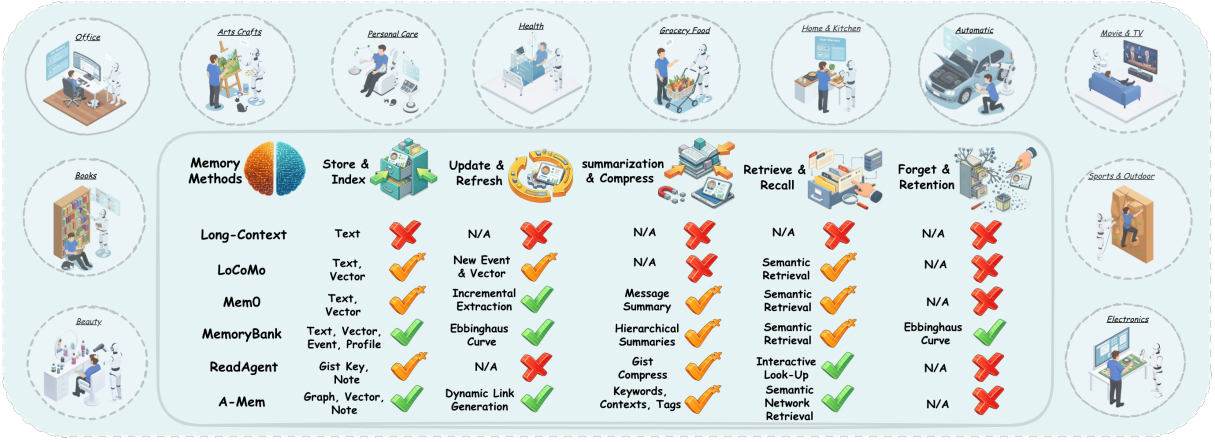

Figure 2: The MemoryCD benchmark spans **12 real-world domains** and evaluates **6 SOTA memory methods**. Different from other memory benchmarks targeting one specific memory stage (mostly retrieval), we design 4 basic tasks with 2 settings to provide end-to-end user satisfaction evaluation based on the user feedback types (Table 2).

fail to provide user feedback from authentic users. Others with real user records, like LaMP (Salemi et al., 2024) and MSC (Xu et al., 2022), are limited in length, offering only short sequences that do not stress the long-context windows available in modern LLMs. Additionally, user-centric datasets only evaluate preferences within isolated domains, ignoring the fact that user behavior is inherently across different domains. Real users express consistent personal tastes across various scenarios, and cross-domain memory can largely increase the cold-start user experiences (Zhang et al., 2025a,c).

From the evaluation perspective, existing memory benchmarks lack end-to-end, user-centric personalization assessments across diverse scenarios. As illustrated in Figure 2, memory systems (Xu et al., 2025; Zhong et al., 2024; Zhu et al., 2025; Zhang et al., 2026b) generally involve five distinct stages; however, prevailing evaluation pipelines often treat memory as a narrow retrieval problem (Maharana et al., 2024). Tasks often reduce to detecting a specific fact in a sentence from a large context input. Therefore, they fail to evaluate the other stages of the memory lifecycle due to a lack of established ground-truth user feedback directly from the synthetic data. In many production-level consumer applications, end-to-end evaluations like A/B testing (Quin et al., 2024) are preferred because they directly reflect user satisfaction. Real personalization requires agents to reason over long-term trajectories, infer latent preferences, and more importantly, act in ways aligned with user goals (Zhang et al., 2025d; Shao et al., 2025). Effective memory evaluation therefore need to go beyond retrieval to user satisfaction assessments of whether stored and retrieved memory meaningfully influence downstream user decisions (Huang et al., 2026).

To address these limitations, we propose **MEMORYCD**, the first benchmark that assesses long-context user-centric memory across domains with real user feedback. Based on data analysis on Amazon-Review (Hou et al., 2024), we find that there is a large overlap in distinct domains, while they demonstrate the user preference covering many real-world applications, ranging from entertainment (*Books* and *Movies*) to healthcare (*Health* and *Personal Care*) and lifestyle (*Sports* and *Outdoors*). In these scenarios, users' preferences expressed in one domain and scenario influence their decisions in others, and an agent must systematically optimize the memory mechanisms to organize and utilize these user histories to provide coherent support across tasks. Furthermore, we implement 6 representative memory systems and find that existing memory mechanisms with long-context LLMs are still far from achieving user satisfaction. Nevertheless, our analysis shows that cross-domain user memory provides a promising pathway toward more comprehensive user understanding, although effective strategies for utilizing such memory remain largely unexplored. In all, our work advances this direction through three key contributions:

- We construct the first cross-domain long-context memory benchmark grounded in real user histories from the user-centric perspective for life-long personalization evaluation.
- We design an end-to-end memory evaluation pipeline with 4 basic tasks in 2 memory settings to test whether agents can leverage mem-

Data Source	Memory Benchmarks	Context Length	# of Sessions	Long Timespan	User Level	Cross-Domain	Real Feedback
Synthetic	LoCoMo (Maharana et al., 2024)	9K	35	✗	✗̸	✗	✗
	PrefEval (Zhao et al., 2025)	100K	–	✗	✗̸	✗̸	✗
	LongMemEval (Wu et al., 2025a)	1.5M	500	✗̸	✗̸	✗	✗
	PersonaMem (Jiang et al., 2025)	1M	60	✓	✗̸	✗̸	✗
	MemBench (Tan et al., 2025a)	100K	0.5K-8K	✗̸	✗̸	✗	✗
	MemoryAgentBench (Hu et al., 2025)	1M	–	✗̸	✗	✗	✗
	HaluMem (Chen et al., 2025)	1M	120	✓	✓	✗	✗̸
Real User	MSC (Xu et al., 2022)	1K	1-5	✗	✗̸	✗	✗̸
	LaMP (Salemi et al., 2024)	1K	–	✗̸	✓	✗	✓
	MEMORYCD (Ours)	400K	1K+	✓	✓	✓	✓

Table 1: Comparison of representative memory benchmarks. ✓: fully covered; ✗̸: partially covered; ✗: not covered. **Synthetic** denotes simulated user behaviors. **Real User** denotes data constructed from real user interaction histories. **# of Sessions** is the number of sessions per individual users.

ory for downstream user satisfaction, rather than merely long-context retrieval.

- We conduct comprehensive evaluations with 14 frontier long-context LLMs and implemented 6 memory systems over 12 diverse domains, providing more insights and conclusions for future memory mechanism design.

2 Task Settings and Problem Formulation

We consider a memory evaluation setting for user-centric life-long personalization, in which the raw memory of each user $u \in U$ includes the interactions histories with various entities $i \in I$ across multiple years and domains $d \in D$. For each user, we construct a long-term behavioral record pool for user memory, defined as

$$\mathcal{M}_u = \{(i, m_i, r_{u,i}, t_{u,i}, x_{u,i})\}_{i \in \mathcal{I}_u},$$

where i represents the item entity along with auxiliary metadata m_i, and $r_{u,i} \in \{1, \ldots, 5\}$ denotes the numerical preference rating. $t_{u,i}$ is the timestamp of the interaction, and $x_{u,i}$ contains the user textual preference opinion, including summarized title $x'_{u,i}$ and full review text $x^*_{u,i}$. This sophisticated and real memory raw data represents the historical context of users over multiple years, during which an LLM agent must infer their long-term preference and linguistic tendencies to generalize to cross-domain behavioral patterns. Each task in MEMORYCD is formalized as a mapping

$$f_\theta : (\mathcal{M}_u, \mathcal{X}) \rightarrow \mathcal{Y},$$

where θ denotes the underlying LLM and memory mechanism, $\mathcal{X}$ specifies the task-dependent instruction input, and $\mathcal{Y}$ denotes the task-specific action output space. These tasks collectively evaluate whether the agent can approximate and align with the real user behavior by conditioning on users' prior behavior patterns in their memory.

In following subsections, we first detail the four basic personalization tasks as in Table 2 and then clarify the two memory evaluation settings with single-domain and cross-domain memory sources.

2.1 Personalized Rating Prediction

The personalized rating prediction task measures an agent's ability to model quantitative user preferences. Given the user memory $\mathcal{M}_u$ and metadata m_i associated with a target item, the objective is to predict the numerical preference rating $\hat{r}_{u,i}$ that user u would assign to item entity i. Formally, the agent implements a regression function

$$\hat{r}_{u,i} = f_\theta(\mathcal{M}_u, m_i),$$

and performance is evaluated using deviation-based metrics: mean absolute error (MAE) and root mean-square error (RMSE) (Hodson, 2022). This task probes the degree of long-term preference modeling and further calibration to numerical predictions.

2.2 Personalized Item Ranking

The personalized ranking task evaluates the agent's capacity to model relative preference strengths of the user. Given the memory $\mathcal{M}_u$ and a set of candidate items $C = \{i_1, \ldots, i_k\}$ in the same domain, the agent predicts a ranking preference $\pi_u(C)$ reflecting the expected ordering of user interest:

$$\pi_u(C) = f_\theta(\mathcal{M}_u, m_{i_1}, \ldots m_{i_k}),$$

Performance is assessed using NDCG@K (Wang et al., 2013), which quantify ranking consistency against the user's true preference signals.

Aspect	★ Rating Prediction	Item Ranking	Review Summarization	Review Generation
Category	User decision-making (quantitative preference modeling)	User decision-making (relative preference modeling)	User text generation (attitude & sentiment reflection)	User text generation (preference & linguistic expression style)
Goal	Predict user's rating (1–5) for a target item.	Rank candidate items by expected user interest.	Generate a concise title reflecting user tone and sentiment.	Generate full review text expressing user preferences and style.
Input	User memory $\mathcal{M}_u$; item metadata m_i	User memory $\mathcal{M}_u$; candidate set C	User memory $\mathcal{M}_u$; item metadata m_i; user review context $x^*_{u,i}$	User memory $\mathcal{M}_u$; item metadata m_i
Output	Rating $\hat{r}_{u,i} \in \{1, \ldots, 5\}$	Ranking $\pi_u(C)$	Short title $x'_{u,i}$	Full review $x^*_{u,i}$
Metrics	MAE, RMSE	NDCG@K	ROUGE-L, BLEU-1	ROUGE-L, BLEU-1

Table 2: Summary of the four personalization tasks in MEMORYCD.

2.3 Personalized Review Summarization

The summarization task measures the agent's ability to produce concise, sentiment-aligned review titles grounded in the user's historical short review title patterns. Given the user memory $\mathcal{M}_u$, metadata m_i for the target item, and the review full text x^*_i, the model generates a short title

$$x'_{u,i} = f_\theta(\mathcal{M}_u, m_i, x^*_{u,i}).$$

Evaluation is conducted using ROUGE-L (Lin, 2004) and BLEU-1 (Papineni et al., 2002) to measure the alignment with the user's tone, sentiment distribution, and lexical tendencies. This tests the capability to correctly integrate long-term preference into concise natural language outputs.

2.4 Personalized Review Generation

The personalized review generation task evaluates the agent's ability to produce full-length review text that reflects both user's preference regarding item attributes and the user's established expression style. Given the user memory $\mathcal{M}_u$ and the metadata m_i of the target item, the agent generates a personalized review

$$x^*_{u,i} = f_\theta(\mathcal{M}_u, m_i).$$

Generated reviews are evaluated using ROUGE-L for sentence-level consistency and BLEU-1 for unigram-level lexical overlap, which collectively assess semantic fidelity, stylistic alignment, and preference-grounded expression. This task provides a comprehensive measure of how effectively an LLM agent can integrate long-horizon preference information into coherent, personalized natural language.

2.5 Benchmark Memory Source Settings

To evaluate how LLM agents utilize long-term user memory to align to the user preference, we design two complementary memory source settings that vary the relationship between the memory source and the target evaluations.

Single-domain memory. In the single-domain setting, the user memory $\mathcal{M}_u$ and the evaluation samples are drawn from the same domain. All historical interactions used for memory construction originate from the target domain. This setting evaluates an agent's ability to perform within-domain long-context personalization of different scenarios.

Cross-domain memory. In the cross-domain setting, the agent is evaluated on a target domain while conditioning on user memory sourced from other domains. The memory formats are close to many real-world (especially cold-start) user histories, where $\mathcal{M}_u$ contains heterogeneous interactions that do not belong to the evaluation domain.

3 Dataset Analysis and Construction

To curate the user-centric memory formulation, we need user-related datasets of long-context tokens with each user of long-range histories. Here, MEMORYCD is constructed from the Amazon Review (Hou et al., 2024) corpus and treats each user as a long-horizon behavioral process rather than a collection of isolated interactions.

To characterize the feasibility of cross-domain personalization, we analyze the degree of user overlap between domains. For each pair of domains (d_i, d_j), we compute the ratio of users who appear

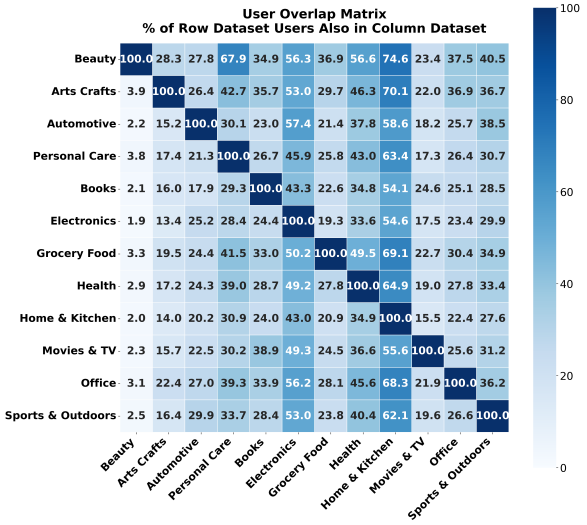

Figure 3: **Ratios of overlapping users across domains**. Each cell shows the percentage of users in one domain that also appear in another, revealing feasible cross-domain personalization settings.

in both domains relative to the total number of users in d_i. Figure 3 visualizes these overlap ratios as a heatmap. While many domain pairs exhibit limited overlap, we observe a non-trivial subset of users who are active across multiple domains, providing a natural foundation for cross-domain evaluation.

We construct MEMORYCD across multiple domains and retain only users with sufficiently long interaction histories to ensure stable preference signals and long context memory lengths. Domain-specific minimum interaction thresholds are applied to filter sparse users, resulting in high-quality user memories with thousands of interactions per user in many domains. In the main experiments, we focus on four representative and complementary domains: *Personal Care*, *Books*, *Electronics*, and *Home & Kitchen*. These domains exhibit diverse interaction patterns, textual styles, and large user overlaps, suitable for evaluating both within-domain and cross-domain user behaviors. More datasets processing details are attached in Appendix C.

4 Experiments

4.1 Experimental Setup

We evaluate 14 frontier long-context LLM backbones in combination with 6 state-of-the-art memory methods. The model suite covers three most advanced LLM families, including the GPT series (Achiam et al., 2023), Claude series (Anthropic, 2024), and Gemini series (Comanici et al., 2025), spanning a wide range of context capacities and agentic behaviors. The memory methods include direct long-context prompting (Liu et al., 2024) and five memory-augmented agents, including LoCoMo (Maharana et al., 2024), Mem0 (Chhikara et al., 2025), ReadAgent (Lee et al., 2024), MemoryBank (Zhong et al., 2024), and A-Mem (Xu et al., 2025), that differ in how user information is stored, compressed, updated, retrieved, and forgotten. More memory methods descriptions are introduced in Appendix D.

For each user, we use the most recent three evaluation instances by default, resulting in hundreds of test cases per domain. Users are filtered using domain-specific minimum interaction thresholds (ranging from 50 to 1,000 historical interactions as in Table 6) to ensure that each input memory constitutes a long-context, preference-revealing user history. To construct a challenging yet well-balanced cross-domain benchmark, we restrict the cross-domain dataset to four domains: *Personal Care*, *Books*, *Electronics*, and *Home & Kitchen*. This selection explicitly balances the trade-off between the number of eligible users and the length of interaction histories per user, ensuring both sufficient population size and rich cross-domain memory. Within single-domain or cross-domain settings, all models and memory methods are evaluated using the same user memories, task prompts, and decoding configurations to ensure fair and controlled comparison. Additional experimental details including prompts are provided in Appendix E and J. Note that, We do not adopt LLM-as-a-judge for text generation tasks due to its scalability limitations, inconsistency, and self-preference bias, as discussed in Appendix G.

4.2 Long-Context Prompting of LLMs in Single-Domain

Although larger and more capable models generally perform better, the performance metrics are still far from perfect alignment with the user preference. In several cases, the gains of more advanced models are uneven across tasks. Table 3 reports the overall performance of state-of-the-art LLM base models across four personalization tasks in MEMORYCD under long-context prompting, where each model directly consumes the full user interaction history without explicit memory abstraction or retrieval. Improvements in rating or ranking scores do not translate into similar gains in personalized text generation quality. This decoupling suggests that personalization under long-context memory

Dataset	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Books	GPT-5-Nano	0.543	0.872	0.135	0.164	0.122	0.055	0.110	0.145
	GPT-5-Mini	0.464	0.816	0.161	0.190	0.138	0.063	0.130	0.174
	GPT-5	**0.330**	**0.624**	0.146	0.177	0.137	0.064	0.132	0.195
	Claude-4 Opus	0.611	0.954	0.142	0.183	0.126	0.070	0.104	0.143
	Claude-4 Sonnet	0.670	1.000	0.109	0.163	0.145	0.089	0.110	0.151
	Gemini-2.5 Flash	0.412	0.750	0.127	0.178	0.182	0.092	0.142	0.159
	Gemini-2.5 Pro	0.401	0.762	**0.206**	**0.229**	**0.208**	**0.128**	**0.162**	**0.207**
Electronics	GPT-5-Nano	0.435	0.789	0.252	0.357	0.173	0.076	0.108	0.124
	GPT-5-Mini	0.354	0.701	0.366	0.488	0.202	0.099	0.132	0.158
	GPT-5	0.289	0.602	**0.443**	**0.604**	0.182	0.082	0.123	0.145
	Claude-4 Opus	**0.281**	**0.581**	0.394	0.527	0.228	0.132	0.138	0.168
	Claude-4 Sonnet	0.374	0.749	0.354	0.511	**0.228**	0.139	0.144	0.176
	Gemini-2.5 Flash	0.358	0.744	0.309	0.447	0.211	0.104	0.146	0.173
	Gemini-2.5 Pro	0.337	0.719	0.378	0.561	0.222	**0.141**	**0.161**	**0.200**
Home & Kitchen	GPT-5-Nano	0.324	0.663	0.258	0.418	0.159	0.054	0.107	0.108
	GPT-5-Mini	0.285	0.618	0.385	0.557	0.188	0.075	0.144	0.161
	GPT-5	**0.261**	**0.551**	**0.446**	**0.610**	0.185	0.077	0.134	0.140
	Claude-4 Opus	0.270	0.595	0.282	0.378	0.213	0.113	0.144	0.157
	Claude-4 Sonnet	0.330	0.630	0.276	0.400	0.189	0.117	0.145	0.158
	Gemini-2.5 Flash	0.275	0.643	0.242	0.381	0.211	0.092	0.163	0.181
	Gemini-2.5 Pro	0.303	0.679	0.315	0.444	**0.222**	**0.133**	**0.178**	**0.207**
Personal Care	GPT-5-Nano	0.432	0.771	0.151	0.243	0.150	0.046	0.104	0.104
	GPT-5-Mini	0.357	0.716	0.253	0.383	0.172	0.075	0.139	0.154
	GPT-5	**0.279**	**0.597**	**0.315**	**0.435**	0.161	0.053	0.131	0.139
	Claude-4 Opus	0.380	0.740	0.234	0.332	0.186	0.103	0.137	0.147
	Claude-4 Sonnet	0.432	0.767	0.203	0.304	0.196	**0.118**	0.138	0.148
	Gemini-2.5 Flash	0.359	0.696	0.242	0.373	0.191	0.072	0.162	0.174
	Gemini-2.5 Pro	0.323	0.665	0.266	0.417	**0.200**	0.110	**0.183**	**0.204**

Table 3: Performance comparison of frontier LLMs in single-domain with long-context memory prompting.

is not merely a naive problem naturally solved by the scaling law and motivates explicit and automatic memory mechanisms that can abstract, compress, and route user information more effectively than raw prompting. Overall, GPT-5 and Gemini-2.5 Pro are the most competitive backbones across tasks, with GPT-5 excelling in rating and ranking tasks, and Gemini-2.5 Pro consistently leading in generation quality. For example, GPT achieves the lowest rating prediction error in the *Books* domain (RMSE 0.624) and delivers the best ranking performance in *Electronics* (NDCG@3 0.604). In contrast, Gemini-2.5 Pro dominates language generation tasks, attaining the highest review generation scores across domains, including ROUGE-L of 0.162 in *Books*.

4.3 Single-Domain Memory Evaluation

Different memory methods show diverse advantages and improvements in different personalization tasks, but no one single method dominates. Table 4 compares representative memory methods on *Books* domain across four personalization tasks under three LLM backbones (GPT-5, Claude-4 Sonnet, Gemini-2.5 Pro). The evaluations on the *Electronics*, *Home & Kitchen*, and *Personal Care* are attached in Appendix Table 9, Table 10, and Table 11, respectively. Concretely, for rating prediction, long-context can be particularly strong (e.g., with GPT-5, long-context attains the best MAE of 0.330 among methods) because the relevant interactions are usually spread across the past user history. Inserting the full interaction history input can avoid the loss of extractions and allow strong backbones to directly aggregate weak but numerous cues. In contrast, more recent memory methods tend to help most when the task is retrieval-sensitive or generation-sensitive. For instance, the Mem0 pipeline based on the Claude-4 Sonnet boosts summarization ROUGE-L from 0.145 to 0.202 by removing redundant turns and gathering user details. A-Mem's structured note construction integrated with dynamic linking can especially benefit text metrics on some backbones by enabling compositional retrieval over organized memories. However, although each mechanism

Method	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Long-Context	GPT-5	**0.330**	**0.624**	0.146	0.177	0.137	0.064	0.132	0.195
	Claude-4 Sonnet	0.670	1.000	0.109	0.163	0.145	0.089	0.110	0.151
	Gemini-2.5 Pro	0.401	0.762	**0.206**	**0.229**	**0.208**	**0.128**	**0.162**	**0.207**
Mem0	GPT-5	**0.419**	**0.698**	**0.184**	0.217	0.173	0.097	0.143	**0.195**
	Claude-4 Sonnet	0.491	0.772	0.172	0.207	0.202	0.126	0.141	0.182
	Gemini-2.5 Pro	0.464	0.803	0.180	**0.221**	**0.215**	**0.135**	**0.162**	0.185
LoCoMo	GPT-5	**0.360**	0.643	**0.232**	**0.253**	0.175	0.107	0.147	**0.201**
	Claude-4 Sonnet	0.399	0.637	0.221	0.241	0.183	0.121	0.144	0.186
	Gemini-2.5 Pro	0.374	**0.632**	0.230	0.248	**0.192**	**0.127**	**0.161**	0.176
ReadAgent	GPT-5	**0.431**	**0.700**	**0.225**	**0.243**	0.174	0.095	0.142	**0.193**
	Claude-4 Sonnet	0.476	0.727	0.202	0.238	**0.188**	**0.120**	0.140	0.186
	Gemini-2.5 Pro	0.479	0.807	0.210	0.235	0.181	0.117	**0.155**	0.173
MemoryBank	GPT-5	**0.356**	**0.627**	0.221	0.237	0.166	0.096	**0.166**	**0.204**
	Claude-4 Sonnet	0.412	0.665	0.210	0.228	0.205	**0.137**	0.145	0.188
	Gemini-2.5 Pro	**0.378**	0.679	**0.228**	**0.242**	**0.216**	0.135	0.160	0.173
A-Mem	GPT-5	**0.337**	**0.624**	0.210	0.232	0.162	0.091	0.143	0.203
	Claude-4 Sonnet	0.371	0.633	0.150	0.188	0.212	**0.150**	0.147	0.197
	Gemini-2.5 Pro	0.390	0.687	**0.217**	**0.237**	**0.221**	0.142	**0.168**	**0.209**

Table 4: Comparison of memory methods on *Books* using representative LLMs.

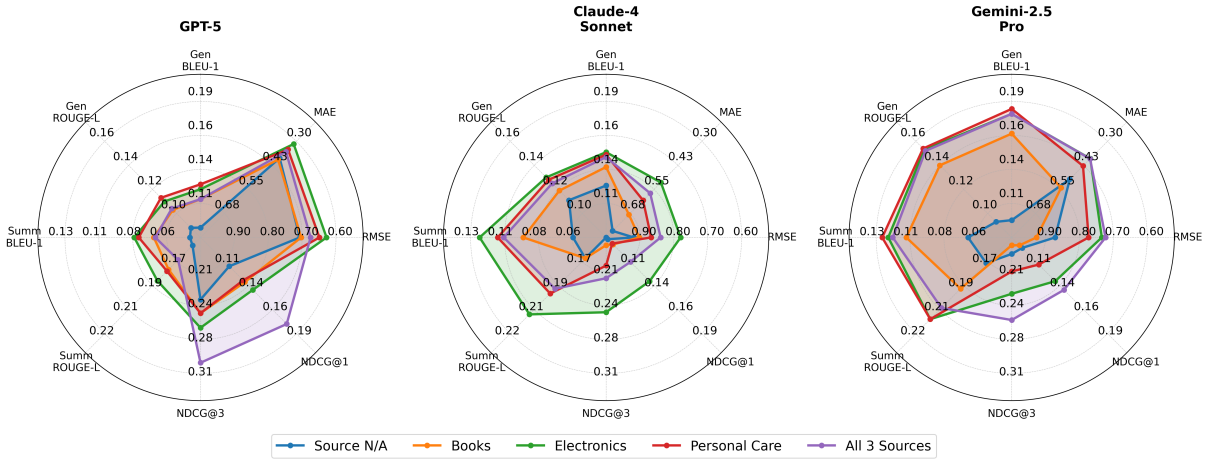

Figure 4: Performance comparisons using different memory sources evaluated on the *Home & Kitchen*. Each radar plot corresponds to a representative frontier long-context LLM (GPT-5, Claude-4 Sonnet, Gemini-2.5 Pro). Curves represents different memory sources: no memory (Source N/A), single-domain memories (*Books*, *Electronics*, *Personal Care*), and aggregated cross-domain memory (All 3 Sources). Scales are unified for fair comparison.

yields different levels of improvements, no method is uniformly dominant across all four tasks and the overall metrics value are still low.

4.4 Cross-Domain Memory Sources Comparison

The effectiveness of cross-domain memory strongly depends on semantic alignment between source and target domains. Figure 4 illustrates that the usefulness of a memory source is highly domain-dependent. Across different backbones, *Electronics* memories yield the strongest personalization on *Home & Kitchen*. This pattern is intuitive that *Electronics* and *Home & Kitchen* share more overlapping product attributes and preference factors (e.g., brand, durability, functionality), making preference cues and evidence more reusable across domains. In contrast, Books memories of the same users show weaker transfer to the *Home & Kitchen* domain and many metrics are closely as those without additional memory sources, representing that the additional transferred memory is not always

Method	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Long-Context	GPT-5	**0.354**	**0.675**	**0.179**	**0.299**	0.174	0.070	0.104	0.111
	Claude-4 Sonnet	0.570	0.839	0.108	0.212	0.192	0.108	0.125	0.145
	Gemini-2.5 Pro	0.391	0.723	0.140	0.255	**0.204**	**0.119**	**0.152**	**0.180**
Mem0	GPT-5	**0.367**	**0.661**	**0.146**	**0.284**	0.189	0.094	0.112	0.122
	Claude-4 Sonnet	0.596	0.823	0.140	0.255	0.197	0.116	0.127	0.146
	Gemini-2.5 Pro	0.458	0.774	0.117	0.230	**0.205**	**0.121**	**0.149**	**0.176**
LoCoMo	GPT-5	**0.348**	**0.653**	**0.214**	**0.355**	0.196	0.094	0.130	0.130
	Claude-4 Sonnet	0.522	0.776	0.155	0.272	0.189	0.106	0.130	0.150
	Gemini-2.5 Pro	0.402	0.759	0.205	0.349	**0.202**	**0.119**	**0.157**	**0.185**
ReadAgent	GPT-5	**0.408**	**0.690**	**0.171**	**0.304**	0.188	0.088	0.114	0.126
	Claude-4 Sonnet	0.590	0.815	0.142	0.244	0.183	0.102	0.127	0.148
	Gemini-2.5 Pro	0.486	0.844	0.150	0.267	**0.192**	**0.111**	**0.152**	**0.180**
MemoryBank	GPT-5	**0.326**	**0.634**	**0.175**	**0.319**	0.206	0.109	0.124	0.139
	Claude-4 Sonnet	0.507	0.758	0.133	0.263	0.201	0.126	0.137	0.159
	Gemini-2.5 Pro	0.397	0.726	0.134	0.260	**0.214**	**0.134**	**0.165**	**0.196**
A-Mem	GPT-5	**0.316**	**0.613**	0.132	0.239	0.194	0.092	0.113	0.122
	Claude-4 Sonnet	0.514	0.766	**0.146**	**0.274**	0.209	0.123	0.130	0.151
	Gemini-2.5 Pro	0.363	0.695	0.117	0.230	**0.221**	**0.134**	**0.155**	**0.184**

Table 5: Cross-domain evaluation of different memory methods on *Home & Kitchen*. Memory includes source domains of *Books, Electronics, and Personal Care*.

helpful but potentially introduces noise.

Notably, aggregating cross-domain memories (**All 3 Sources**) often improves task performance. In all the backbones, **All 3 Sources** perform better than **Source N/A**. For GPT-5, **All 3 Sources** provides the best ranking performance, substantially outperforming **Source N/A**, indicating that broader personal information coverage helps retrieve relevant preference evidence. However, aggregation is not always optimal. The best MAE/RMSE often comes from the most semantically aligned single source (notably *Electronics*), while mixing in less-related domains can introduce conflicting information to hinder the prediction.

4.5 Cross-Domain Memory Evaluation

Cross-domain personalization benefits from memory methods. In Table 5, Home & Kitchen with a memory pool aggregated from Books/Electronics/Personal Care, raw long-context prompting yields only moderate ranking quality, and memory methods mainly help by filtering and compressing heterogeneous histories into task-relevant cues, mitigating distraction from irrelevant evidence in long inputs. For ranking, LoCoMo long-horizon summarization shows the strongest transfer for GPT-5 and Gemini-2.5 Pro (NDCG@3 improves to 0.355 and 0.349, vs. 0.299 and 0.255). This improvement suggests that explicit memory routing is especially important when the model must suppress cross-domain noise. For the summarization and generation task, we observe A-Mem and MemoryBank perform the best, respectively. This shows that either the dynamic and consistent memory refreshment or removing noise can largely support cross-domain personalization on the generation task.

5 Conclusion

This paper introduces MEMORYCD, a large-scale benchmark for evaluating ultra-long-context user memory in LLM agents using real-world, multi-domain interaction histories. By framing personalization as an end-to-end behavioral modeling problem, MEMORYCD moves beyond retrieval-focused evaluation and assesses how memory usage impacts decision-making and personalized generation. Extensive experiments show that increasing model scale alone is insufficient for robust long-context personalization. Instead, memory design plays a critical role, with different mechanisms exhibiting distinct strengths across tasks. These findings underscore the need for principled memory abstractions that can effectively organize and exploit long-term user histories, and position MEMORYCD as a foundation for future research on personalized and memory-driven LLM agents.

Limitations

While MEMORYCD provides a large-scale, real-user benchmark for long-context and cross-domain personalization, it has several limitations. First, although we include 12 domains and a carefully curated cross-domain subset, cross-domain evaluation is restricted to fixed memory operations rather than learned memory mechanisms (Zhang et al., 2026a), leaving broader domain transfer as future work. Second, our evaluation focuses on memory utilization and does not consider user privacy and safety guardrail (Wu et al., 2025b; Huang et al., 2025) over deployment time. Finally, large-scale training-based memory construction approaches (Wang et al., 2025) are not incorporated into the benchmark, as most state-of-the-art long-context LLMs remain closed-source.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. 2025. Memorybench: A benchmark for memory and continual learning in llm systems. *arXiv preprint arXiv:2510.17281*.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1(1):4.

Yuanchen Bei, Tianxin Wei, Xuying Ning, Yanjun Zhao, Zhining Liu, Xiao Lin, Yada Zhu, Hendrik Hamann, Jingrui He, and Hanghang Tong. 2026. Mem-gallery: Benchmarking multimodal long-term conversational memory for mllm agents. *arXiv preprint arXiv:2601.03515*.

Hongru Cai, Yongqi Li, Wenjie Wang, Fengbin Zhu, Xiaoyu Shen, Wenjie Li, and Tat-Seng Chua. 2025. Large language models empowered personalized web agents. In *Proceedings of the ACM on Web Conference 2025*, pages 198–215.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. 2025. Halumem: Evaluating hallucinations in memory systems of agents. *arXiv preprint arXiv:2511.03506*.

Jiangjie Chen, Xintao Wang, Rui Xu, Siyu Yuan, Yikai Zhang, Wei Shi, Jian Xie, Shuang Li, Ruihan Yang, Tinghui Zhu, and 1 others. 2024. From persona to personalization: A survey on role-playing language agents. *Transactions on Machine Learning Research*.

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

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

Timothy O Hodson. 2022. Root mean square error (rmse) or mean absolute error (mae): When to use them or not. *Geoscientific Model Development Discussions*, 2022:1–10.

Yupeng Hou, Jiacheng Li, Zhankui He, An Yan, Xiusi Chen, and Julian McAuley. 2024. Bridging language and items for retrieval and recommendation. *arXiv preprint arXiv:2403.03952*.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025. Evaluating memory in llm agents via incremental multi-turn interactions. *arXiv preprint arXiv:2507.05257*.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, and 1 others. 2026. Rethinking memory mechanisms of foundation agents in the second half. *arXiv preprint arXiv:2602.06052*.

Wei-Chieh Huang, Henry Peng Zou, Yaozu Wu, Dongyuan Li, Yankai Chen, Weizhi Zhang, Yangning Li, Angelo Zangari, Jizhou Guo, Chunyu Miao, and 1 others. 2025. Deepresearchguard: Deep research with open-domain evaluation and multi-stage guardrails for safety. *arXiv preprint arXiv:2510.10994*.

Bowen Jiang, Zhuoqun Hao, Young-Min Cho, Bryan Li, Yuan Yuan, Sihao Chen, Lyle Ungar, Camillo J Taylor, and Dan Roth. 2025. Know me, respond to me: Benchmarking llms for dynamic user profiling and personalized responses at scale. *arXiv preprint arXiv:2504.14225*.

Kuang-Huei Lee, Xinyun Chen, Hiroki Furuta, John Canny, and Ian Fischer. 2024. A human-inspired reading agent with gist memory of very long contexts. In *International Conference on Machine Learning*, pages 26396–26415. PMLR.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Junyu Luo, Weizhi Zhang, Ye Yuan, Yusheng Zhao, Junwei Yang, Yiyang Gu, Bohan Wu, Binqi Chen, Ziyue Qiao, Qingqing Long, and 1 others. 2025. Large language model agent: A survey on methodology, applications and challenges. *arXiv preprint arXiv:2503.21460*.

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

Humza Naveed, Asad Ullah Khan, Shi Qiu, Muhammad Saqib, Saeed Anwar, Muhammad Usman, Naveed Akhtar, Nick Barnes, and Ajmal Mian. 2025. A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

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

Federico Quin, Danny Weyns, Matthias Galster, and Camila Costa Silva. 2024. A/b testing: A systematic literature review. *Journal of Systems and Software*, 211:112011.

Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2024. Lamp: When large language models meet personalization. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7370–7392.

Yifan Shao, Peilin Zhou, Shoujin Wang, Weizhi Zhang, Xu Cai, and Sunghun Kim. 2025. Dtrec: Learning dynamic reasoning trajectories for sequential recommendation. *arXiv preprint arXiv:2512.14036*.

Weiwei Sun, Miao Lu, Zhan Ling, Kang Liu, Xuesong Yao, Yiming Yang, and Jiecao Chen. 2025. Scaling long-horizon llm agent via context-folding. *arXiv preprint arXiv:2510.11967*.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025a. Membench: Towards more comprehensive evaluation on the memory of llm-based agents. *arXiv preprint arXiv:2506.21605*.

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, and 1 others. 2025b. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439.

Yining Wang, Liwei Wang, Yuanzhi Li, Di He, and Tie-Yan Liu. 2013. A theoretical analysis of ndcg type ranking measures. In *Conference on learning theory*, pages 25–54. PMLR.

Yu Wang, Ryuichi Takanobu, Zhiqi Liang, Yuzhen Mao, Yuanzhe Hu, Julian McAuley, and Xiaojian Wu. 2025. Mem-{\alpha}: Learning memory construction via reinforcement learning.

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

Yaozu Wu, Jizhou Guo, Dongyuan Li, Henry Peng Zou, Wei-Chieh Huang, Yankai Chen, Zhen Wang, Weizhi Zhang, Yangning Li, Meng Zhang, and 1 others. 2025b. Psg-agent: Personality-aware safety guardrail for llm-based agents. *arXiv preprint arXiv:2509.23614*.

Jing Xu, Arthur Szlam, and Jason Weston. 2022. Beyond goldfish memory: Long-term open-domain conversation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 5180–5197.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.

John Yang, Carlos E Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems*, 37:50528–50652.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026a. Memskill: Learning and evolving memory skills for self-evolving agents. *arXiv preprint arXiv:2602.02474*.

Haozhen Zhang, Haodong Yue, Tao Feng, Quanyu Long, Jianzhu Bao, Bowen Jin, Weizhi Zhang, Xiao Li, Jiaxuan You, Chengwei Qin, and 1 others. 2026b. Learning query-aware budget-tier routing for runtime agent memory. *arXiv preprint arXiv:2602.06025*.

Weizhi Zhang, Yuanchen Bei, Liangwei Yang, Henry Peng Zou, Peilin Zhou, Aiwei Liu, Yinghui Li, Hao Chen, Jianling Wang, Yu Wang, and 1 others. 2025a. Cold-start recommendation towards the era of large language models (llms): A comprehensive survey and roadmap. *arXiv preprint arXiv:2501.01945*.

Weizhi Zhang, Yangning Li, Yuanchen Bei, Junyu Luo, Guancheng Wan, Liangwei Yang, Chenxuan Xie, Yuyao Yang, Wei-Chieh Huang, Chunyu Miao, and 1 others. 2025b. From web search towards agentic deep research: Incentivizing search with reasoning agents. *arXiv preprint arXiv:2506.18959*.

Weizhi Zhang, Liangwei Yang, Wooseong Yang, Henry Peng Zou, Yuqing Liu, Ke Xu, Sourav Medya, and Philip S Yu. 2025c. Llminit: A free lunch from large language models for selective initialization of recommendation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 2016–2024.

Weizhi Zhang, Wooseong Yang, Yuxin Cui, Zhaohui Guo, Hins Hu, Liangwei Yang, Henry Peng Zou, Qifei Wang, Hanqing Zeng, Jiayi Liu, and 1 others. 2025d. Toward user preference alignment in llm recommendation via explicit context feedback. In *2025 IEEE 7th International Conference on Cognitive Machine Intelligence (CogMI)*, pages 381–388. IEEE.

Weizhi Zhang, Xinyang Zhang, Chenwei Zhang, Liangwei Yang, Jingbo Shang, Zhepei Wei, Henry Peng Zou, Zijie Huang, Zhengyang Wang, Yifan Gao, and 1 others. 2025e. Personaagent: When large language model agents meet personalization at test time. *arXiv preprint arXiv:2506.06254*.

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

Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. Do llms recognize your preferences? evaluating personalized preference following in llms. *arXiv preprint arXiv:2502.09597*.

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

Yuanjie Zhu, Liangwei Yang, Ke Xu, Weizhi Zhang, Zihe Song, Jindong Wang, and Philip S Yu. 2025. Llm-memcluster: Empowering large language models with dynamic memory for text clustering. *arXiv preprint arXiv:2511.15424*.

A Related Works

As summarized in Table 1, existing memory benchmarks for LLM agents differ substantially in data source, temporal coverage, personalization granularity, and evaluation scope. The majority of prior benchmarks are built on synthetic user behaviors, including LoCoMo (Maharana et al., 2024), PrefEval (Zhao et al., 2025), LongMemEval (Wu et al., 2025a), PersonaMem (Jiang et al., 2025), MemBench (Tan et al., 2025a), and MemoryAgentBench (Hu et al., 2025). These benchmarks progressively extend context length and session counts, with some supporting million-token contexts or hundreds of simulated sessions. However, as shown in Table 1, most synthetic benchmarks only partially cover long-term temporal span and individual-level personalization, and they largely focus on single-domain settings without explicit user feedback. Consequently, evaluation often reduces memory to fact retrieval or isolated reasoning tasks, making it difficult to assess whether memory mechanisms genuinely support personalized decision-making or improve user satisfaction.

A smaller line of work evaluates memory using real user interaction histories, including MSC (Xu et al., 2022) and LaMP (Salemi et al., 2024). These datasets introduce authentic user behaviors and, in the case of LaMP, individual-level preference modeling with real feedback signals. However, Table 1 shows that existing real-user benchmarks are limited by short context lengths, few interaction sessions, and a lack of cross-domain coverage, which prevents stress-testing long-context memory and cross-domain preference transfer in modern LLMs. In contrast, MEMORYCD uniquely integrates long-context inputs, long-term interaction histories, individual-level personalization, cross-domain behaviors, and explicit user feedback within a unified evaluation framework. This combination enables end-to-end assessment of whether memory mechanisms meaningfully influence downstream tasks across domains, addressing critical gaps left by prior synthetic and real-user benchmarks.

B Data Collection and Usage Compliance

MEMORYCD is constructed exclusively from the publicly released Amazon Reviews 2023 dataset (Hou et al., 2024), following all data usage rules and licensing conditions specified by the dataset maintainers.[1] The dataset consists of anonymized, user-generated product reviews and ratings, collected and released for non-commercial academic research. No private, sensitive, or personally identifiable information is included or inferred at any stage of benchmark construction.

We strictly adhere to the original dataset schema and semantics. All user identifiers are treated as anonymous tokens, and MEMORYCD does not attempt to de-anonymize users or link identities across external sources. Reviews, ratings, timestamps, and item metadata are used only as provided, without modification or augmentation beyond task-oriented filtering and aggregation.

MEMORYCD is designed purely for evaluation purposes. The benchmark does not train or fine-tune language models on Amazon data, nor does it enable product-level inference or commercial recommendation use. All reported results measure an agent's ability to condition on historical context of users at inference time, consistent with the intended academic use of the Amazon Reviews 2023 dataset.

C Datasets Details

Table 6 summarizes the domain-level statistics of the proposed ultra long-context, user-centric memory benchmark constructed from 12 product domains. The dataset exhibits extremely long user interaction histories, with context lengths ranging from 11K tokens (*Beauty*) to over 314K tokens (*Movies & TV*). Highly active domains such as *Books*, *Electronics*, *Personal Care*, and *Movies & TV* contain thousands of sessions per user (e.g., Books averages 1,552 sessions with a maximum of 3,888), reflecting rich long-term behavioral traces. To ensure meaningful personalization signals, each domain applies a minimum interaction threshold (50–1000 sessions depending on activity level), retaining only sufficiently active users for evaluation. Across domains, the benchmark around 100 users per domain, with each user contributing the most recent three samples for evaluation. The aggregated cross-domain setting further combines interactions across various scenarios, yielding the largest context length (387K tokens) and 5,862 sessions, covering 323 users and 969 samples, which provides a challenging testbed for evaluating memory mechanisms under ultra long-context and cross-domain personalization scenarios.

[1] https://amazon-reviews-2023.github.io/

Domain	Context Length	Max # Sessions	Avg # Sessions	Min # Sessions	Filtering Thresholds	# of Users	# of Samples
Beauty	11,137	165	81	54	50	57	171
Arts	49,101	801	338	204	200	50	150
Automotive	27,104	512	258	203	200	64	192
Personal Care	202,096	2,973	839	504	500	128	384
Books	272,372	3,888	1,552	1,002	1,000	89	267
Electronics	184,084	2,602	781	502	500	94	282
Grocery Food	124,204	1,524	760	510	500	45	135
Health	180,757	2,640	831	501	500	75	225
Home & Kitchen	175,039	2,754	768	501	500	110	330
Movies & TV	313,857	4,369	1,376	804	800	49	147
Office	71,497	921	331	203	200	25	75
Sports & Outdoors	64,527	858	296	203	200	66	198
Cross-Domain	386,736	5,862	973	243	200	323	969

Table 6: Domain-level statistics for the proposed ultra long-context user-centric memory benchmark. Only users exceeding the minimum interaction threshold are retained; each user contributes three evaluation samples.

D Memory Methods and Baselines

We benchmark a diverse set of memory baselines that differ in how user information is stored, retrieved, and integrated into the model. These baselines span from naive full-history prompting to structured or compressed memory representations, allowing us to analyze how different memory designs trade off coverage, relevance, and robustness under ultra-long contexts.

Long-Context. The Long-Context (Liu et al., 2024) method treats memory as a purely implicit mechanism by directly concatenating all historical user interactions into the prompt, relying on the expanded context window of modern LLMs rather than any explicit storage, indexing, or retrieval module. As shown in Figure 2, this approach lacks dedicated components for memory update, summarization, retrieval, or forgetting, and therefore serves as a minimal baseline that tests the raw capacity of long-context modeling. However, as the LLMs training paradigm naturally extend the context window size to the million-token level, this method serve as a simple yet strong baseline.

LoCoMo. LoCoMo (Maharana et al., 2024) implements a retrieval-augmented generation (RAG) paradigm that couples explicit memory storage with vector indexing and on-the-fly retrieval to supplement the base LLM's context window, addressing the limitations of purely concatenated long context. During inference, LoCoMo performs semantic retrieval to bring relevant memories into the prompt, improving response accuracy compared to the Long-Context baseline by reducing reliance on position biases and effectively foregrounding past events even when distant in the session. However, LoCoMo's performance is still bounded by the quality of retrieved vectors and lacks sophisticated forget / retention mechanisms, which can lead to noise accumulation in long-running dialogues.

Mem0. Mem0 (Chhikara et al., 2025) introduces a scalable, modular, and production-level long-term memory architecture that goes beyond simple RAG by dynamically extracting, consolidating, and maintaining salient information from past interactions to support persistent conversational memory across sessions. Rather than relying on concatenated context or raw chunk retrieval, Mem0 uses an LLM to identify key facts and preferences from recent dialogue and then stores these as structured memory entries.

ReadAgent. ReadAgent (Lee et al., 2024) is a human-inspired LLM agent designed to efficiently handle very long contexts by mimicking how people read and remember lengthy documents: it segments text into coherent "episodes," compresses each segment into a gist memory, and selectively looks up detailed passages when needed to complete a task. Unlike simple RAG methods, ReadAgent explicitly performs store and index by grouping content into episodic memory and conducts summarization and compress via gist extraction, creating a compact representation of the original text. During inference, it uses an interactive look-up mechanism to retrieve relevant raw text based on the stored gists, enabling effective utilization of both global context and local details for comprehension tasks.

A-Mem A-Mem (Xu et al., 2025) introduces an agentic memory framework that dynamically organizes and evolves as knowledge-graph notes using principles inspired by the Zettelkasten method, enabling LLM agents to automatically generate structured notes with keywords, contextual descriptions, and semantic tags. When storing a new memory, A-Mem not only indexes it but also analyzes past memories to identify and instantiate dynamic links, thereby forming an interconnected knowledge network that continuously refines itself as new information arrives. During inference, this semantic network retrieval allows the agent to traverse linked memories and surface relevant historical information with greater contextual coherence than flat retrieval baselines.

E Experimental Details

E.1 Evaluation Pipeline

For each user u, we first construct the user memory $\mathcal{M}_u$ by aggregating all historical interactions. Across all domains, this process yields different long-context token length (details in Table 6). Memory-augmented methods are restricted to operating on $\mathcal{M}_u$ only, while all LLMs receive the same information through direct long-context prompting.

E.2 Single-Domain and Cross-Domain Settings Datasets

In the single-domain setting, both the user memory and evaluation queries are drawn from the same domain. We construct 12 single-domain benchmarks, each containing around 100 users with long interaction histories, depending on the domain. In the cross-domain setting, we identify 323 users who are simultaneously active in *Books*, *Electronics*, *Home & Kitchen*, and *Personal Care*. For each such user, the memory aggregates interactions from *Books*, *Electronics*, and *Personal Care* domains, while evaluating to the target domain *Home & Kitchen* by default.

E.3 Implementation and Platforms

All experiments are implemented in a unified evaluation framework. LLM backbones are accessed through official inference APIs, and all memory methods are implemented as modular components that interface with the same prompting templates and evaluation scripts. For each task, decoding hyperparameters are fixed across all models and methods. All evaluations are executed in a single run on the standardized cloud-based compute environment to avoid hardware-induced variability.

E.4 Reproducibility and Fairness

We fix random seeds for all stochastic components and evaluate each model and method configuration over the full set of test instances in each domain. No memory method has access to additional user profile, external corpora, or domain-specific information beyond the provided user memory. All metrics are computed based on identical implementations across settings to ensure that observed differences arise solely from model capacity and memory design rather than evaluation pipelines.

F Evaluation of LLMs with long-context prompting in All Single Domains

As presented in Table 7, across the eight remaining single domains, we observe a clear backbone specialization: GPT-5 is strongest on decision tasks (rating and ranking), and Gemini-2.5 Pro is strongest on generated-related tasks (summarization/generation). For rating prediction, GPT-5 achieves the best MAE in most of the domains, with particularly large margins in *Arts & Crafts*. For ranking, GPT-5 leads NDCG@3 in most of the domains (e.g., *Automotive* 0.606; *Toys & Games* 0.557). In contrast, Gemini-2.5 Pro consistently leads all domains on summarization and generation. This trend shows a systematic advantage of Gemini-2.5 Pro in producing higher-overlap text outputs.

In addition, achieving user satisfaction remains challenging and strongly task-dependent. For example, *Beauty* consistently proves more difficult than structured product domains such as *Arts & Crafts*, suggesting that preference signals in certain domains are noisier or less directly predictive. Moreover, model rankings vary across task types. The backbone that performs best at predicting preferences (GPT-5) is not the one that best verbalizes preferences (Gemini-2.5 Pro), and the performance gap becomes particularly pronounced in generation-intensive domains. Overall, Table 7 reinforces that there is no universally best backbone across all tasks. Choosing the model and the memory method should be conditioned on whether the target objective is decision-centric (rating/ranking) or language-centric (summarization/generation).

Dataset	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Beauty	GPT-5	0.544	0.946	0.175	0.290	0.203	0.085	0.231	0.118
	Claude-4 Sonnet	**0.544**	**0.889**	**0.246**	0.323	0.195	0.117	0.231	0.118
	Gemini-2.5 Pro	0.719	1.221	0.228	**0.349**	**0.230**	**0.129**	**0.284**	**0.159**
Arts Crafts	GPT-5	**0.167**	**0.440**	**0.367**	**0.531**	0.209	0.092	0.246	0.131
	Claude-4 Sonnet	0.347	0.739	0.267	0.430	0.250	0.170	0.230	0.134
	Gemini-2.5 Pro	0.327	0.787	0.327	0.486	**0.292**	**0.201**	**0.307**	**0.182**
Automotive	GPT-5	**0.318**	**0.851**	**0.443**	**0.606**	0.226	0.117	0.159	0.078
	Claude-4 Sonnet	0.427	0.919	0.333	0.453	0.202	0.127	0.181	0.094
	Gemini-2.5 Pro	0.427	1.021	0.328	0.477	**0.291**	**0.188**	**0.228**	**0.118**
Grocery Food	GPT-5	**0.370**	**0.678**	**0.193**	**0.324**	0.192	0.083	0.259	0.141
	Claude-4 Sonnet	0.578	0.919	0.148	0.256	0.151	0.101	0.197	0.116
	Gemini-2.5 Pro	0.533	0.927	0.178	0.293	**0.224**	**0.155**	**0.314**	**0.192**
Health	GPT-5	**0.382**	**0.712**	**0.302**	**0.428**	0.149	0.060	0.250	0.132
	Claude-4 Sonnet	0.627	0.955	0.196	0.324	0.141	0.089	0.207	0.118
	Gemini-2.5 Pro	0.431	0.803	0.173	0.322	**0.180**	**0.106**	**0.310**	**0.188**
Movies & TV	GPT-5	**0.524**	**0.869**	**0.367**	**0.474**	0.131	0.063	0.243	0.142
	Claude-4 Sonnet	0.667	1.030	0.252	0.350	0.129	0.087	0.205	0.119
	Gemini-2.5 Pro	0.612	1.010	0.265	0.391	**0.152**	**0.105**	**0.287**	**0.170**
Office	GPT-5	**0.427**	0.980	**0.320**	**0.498**	0.221	0.097	0.278	0.158
	Claude-4 Sonnet	0.533	0.980	0.227	0.351	0.181	0.113	0.273	0.151
	Gemini-2.5 Pro	0.467	**0.917**	0.320	0.441	**0.252**	**0.169**	**0.341**	**0.206**
Toys & Games	GPT-5	**0.266**	**0.556**	**0.416**	**0.557**	0.199	0.103	0.257	0.142
	Claude-4 Sonnet	0.519	0.892	0.385	0.508	0.229	0.149	0.206	0.119
	Gemini-2.5 Pro	0.403	0.809	0.383	0.547	**0.247**	**0.164**	**0.323**	**0.203**

Table 7: Overall performance of representative LLMs across four tasks in the rest eight single domains of MemoryCD with long-context memory prompting.

G Bias of LLM-as-a-judge for Frontier Models

We focus on the two language-generation tasks: personalized review summarization and personalized review generation. The text outputs are scored independently by three judge models of the frontier long-context LLMs: GPT-5, Claude 4 Sonnet, and Gemini-2.5 Pro. The judge reports scalar preference scores on a 1–10 scale upon comare the LLM results with the ground-truth real user drafted texts.

The results in Table 8 reveal two key patterns. First, individual LLM judges exhibit consistent self-preference bias, assigning higher scores to outputs generated by their own model. This leads to noticeable discrepancies in per-judge rankings despite small absolute score differences. Second, when aggregating scores across judges, the final ranking aligns with ROUGE-L and BLEU-1 based evaluation. In particular, Gemini-2.5 Pro ranks first in both tasks, while the relative ordering between GPT-5 and Claude 4 Sonnet varies by task. These findings suggest that although LLM-as-a-judge is biased at the individual level, simple aggregation can effectively mitigate this bias, albeit at the cost of increased computational overhead. Moreover, LLM-based evaluation is not fully stable or consistent: even under low-temperature settings, the same input can yield varying scores across runs. Therefore, we adopt ROUGE-L and BLEU-1 as scalable and reproducible metrics for evaluating text generation tasks.

H Evaluation of All LLMs with Representative Memory Methods in Single Domains Setting

Figure 5, Figure 6, Figure 7, and Figure 8 compare 14 frontier long-context LLMs on Books and Home & Kitchen under two representative memory methods (MemoryBank and A-Mem). Across both domains, we observe a stable task-wise pattern: (i) performance varies substantially across task types and models that excel at decision-oriented tasks (rating and ranking) are often different from those leading language generation tasks (summarization and generation); (ii) language tasks are relatively consistent across memory choices, where Gemini-

Task	Model	GPT-5	Claude	Gemini	Avg.	Final Rank	ROUGE/BLEU Rank
Review Summarization	GPT-5	**5.9**	5.3	5.1	5.43	3	3
	Claude 4 Sonnet	5.4	**5.8**	5.6	5.60	2	2
	Gemini-2.5 Pro	5.7	5.6	**6.1**	5.80	1	1
Review Generation	GPT-5	**4.6**	4.0	3.8	4.13	2	2
	Claude 4 Sonnet	4.0	**4.3**	4.0	4.10	3	3
	Gemini-2.5 Pro	4.5	4.3	**4.8**	4.53	1	1

Table 8: LLM-as-a-judge evaluation with scores averaged across three judge models. Individual LLM judges exhibit self-preference bias, tending to favor outputs from their own model family. The final ranking is determined by the averaged scores, which align with ROUGE-L and BLEU-1 based evaluation in the Books domain.

family models tend to stay strong on generation and Claude-family models remain competitive on summarization, and GPT-family perform the best on decision tasks (rating/ranking); and (iii) domain difficulty differs: *Home & Kitchen* typically yields higher and more separated ranking scores than Books, suggesting stronger item-attribute regularities and cleaner preference signals in the former, whereas Books appears more ambiguous and thus compresses model gaps.

A second key observation is that the memory method changes which backbones benefit most, especially for rating/ranking: MemoryBank tends to amplify backbones that are already strong at holistic language modeling and preference aggregation, thereby improving decision-oriented tasks such as rating prediction and item ranking. In contrast, A-Mem more frequently shifts the advantage toward models that can effectively leverage structured and explicitly linked memory notes, which proves beneficial for generation-related tasks. Mechanically, the difference arises from how the two memory systems organize and expose user information. MemoryBank emphasizes continuous memory synthesis into a user-centric profile with retrieval for personalization. Therefore, the joint memory can support the decision-making task, such as rating and ranking. Whereas A-Mem organizes memories as structured "notes" with dynamic indexing and linking, enabling compositional traversal over an evolving memory graph, which better supports generative tasks.

I Evaluation of Different Methods in Single Domain Settings

Table 9, Table 10, and Table 11 are jointly show that memory methods improve personalization in a strongly task- and domain-dependent way, with no universally dominant design. Across *Electronics*, *Home & Kitchen*, and *Personal Care*, retrieval and selection-style methods (e.g., Mem0/LoCoMo) most consistently boost ranking (often producing large NDCG gains over long-context prompting). In contrast, structured and profile-style memories (e.g., MemoryBank, A-Mem) more often yield balanced improvements. These highlight that cross-task gains depend not only on memory content but also on how each backbone consumes retrieved and condensed evidence.

On *Home & Kitchen*, ranking benefits the most from retrieval/selection, while rating and generation favor more structured memories and backbone strengths. Compared to long-context prompting, Mem0, LoCoMo, and MemoryBank substantially increase ranking scores for GPT-5 and Claude-4 Sonnet, whereas A-Mem is more competitive on rating (e.g., lowest MAE for GPT-5) but does not match the best ranking variants, suggesting a trade-off between decision-time evidence selection and preference calibration via structured memory. For generation tasks, Gemini-2.5 Pro remains the strongest overall generator in this domain and shows its best generation under MemoryBank-like profile memory, while GPT-5 stays comparatively stronger on decision metrics when paired with the right memory routing. Overall, for generation tasks, Gemini-2.5 Pro remains the strongest generator in this domain, achieving its best performance when paired with MemoryBank-style profile memories, while GPT-5 continues to dominate decision-oriented metrics when combined with appropriate memory strategies.

J Prompt Templates for Personalized Tasks

We present the exact prompt templates used for all four personalized tasks: rating prediction, item ranking, review summarization, and review generation. Each prompt follows a unified structure that conditions on the user's recent interaction history

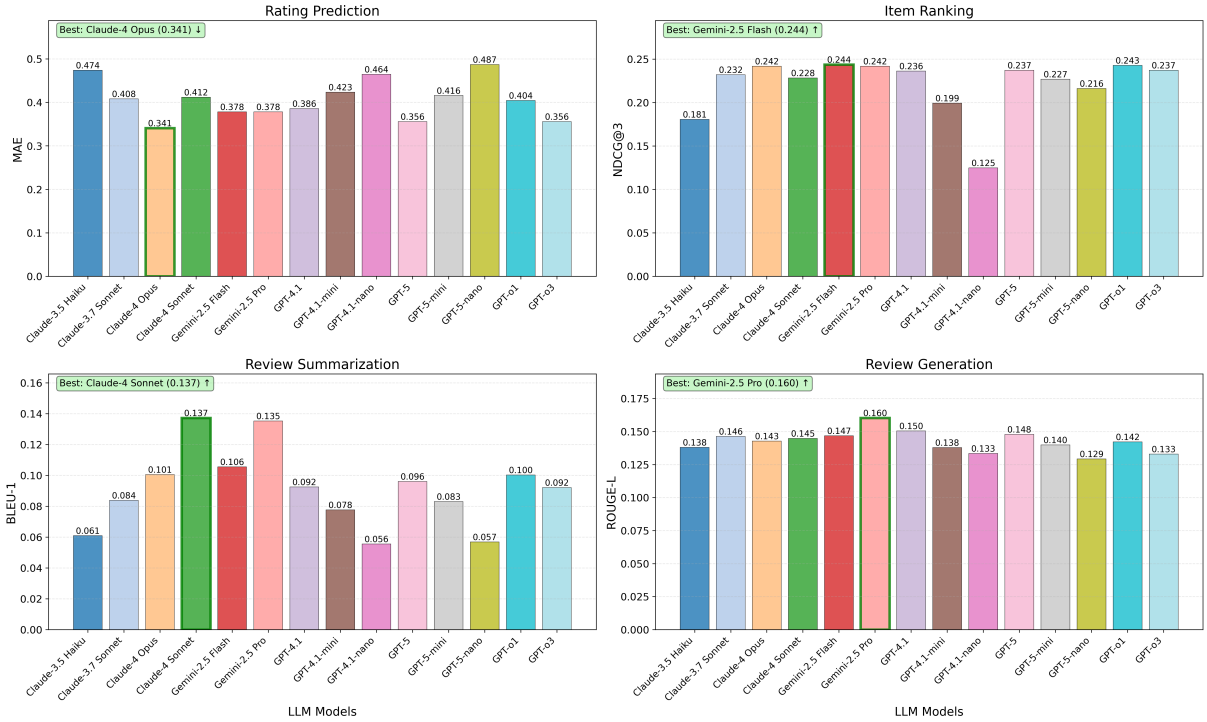

Figure 5: Performance comparisons of 14 frontier long-context LLMs on Books based on MemoryBank.

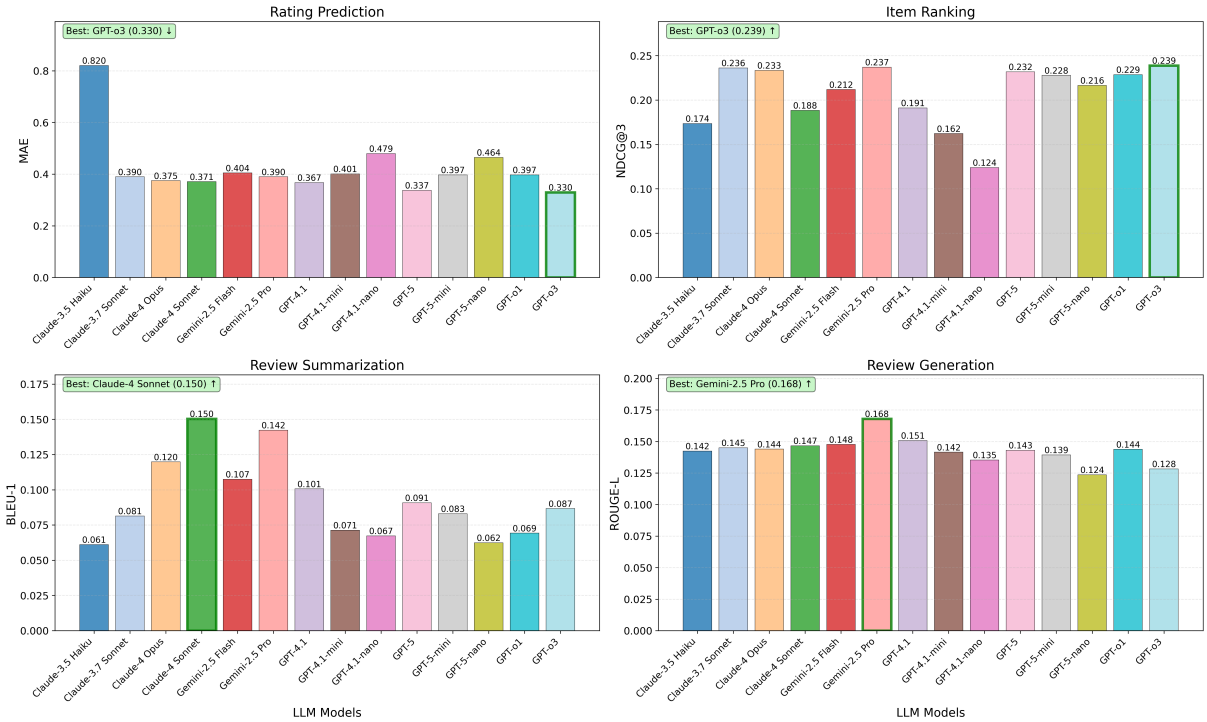

Figure 6: Performance comparisons of 14 frontier long-context LLMs on Books based on A-Mem.

(serving as an implicit memory of preferences and writing style) and task-specific inputs (e.g., target product, candidate set, or review text). We enforce strict output formats to ensure consistency and enable reliable automatic evaluation across tasks.

Rating Prediction Prompt

```
You are a personalized rating prediction
system. Based on a user's purchase and
review history, predict what rating (1-5)
they would give to a new product.
```

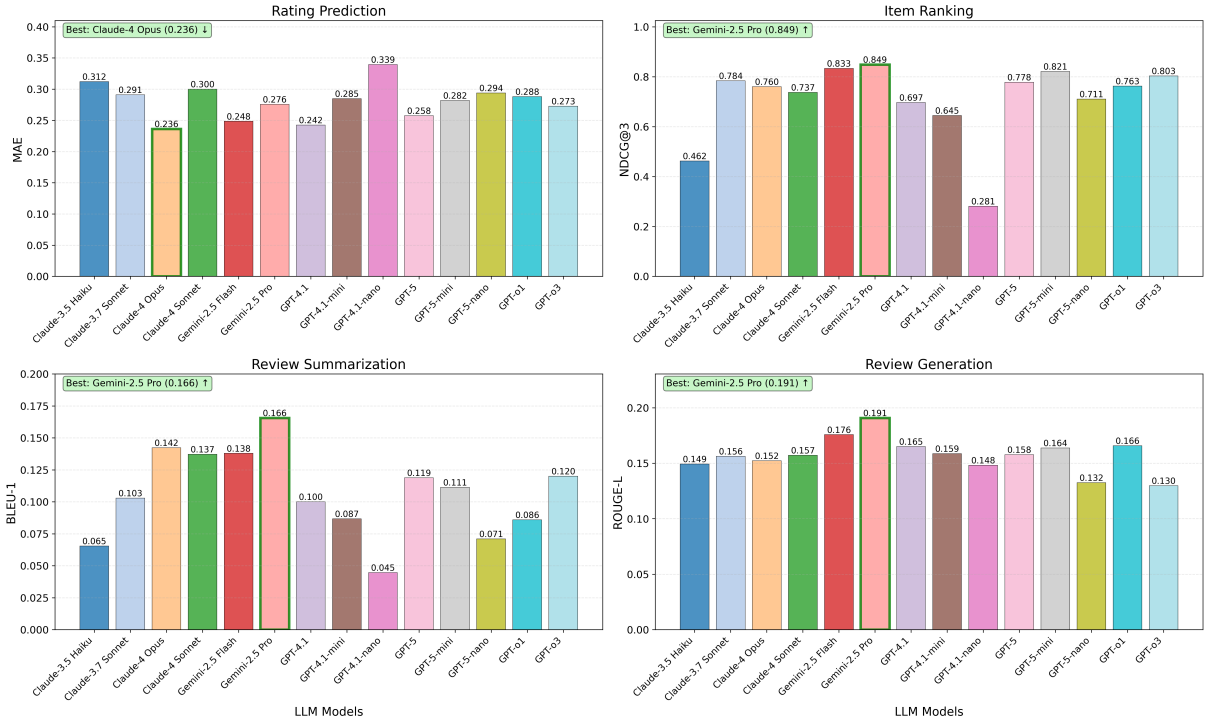

Figure 7: Performance comparisons of 14 frontier long-context LLMs on Home & Kitchen based on MemoryBank.

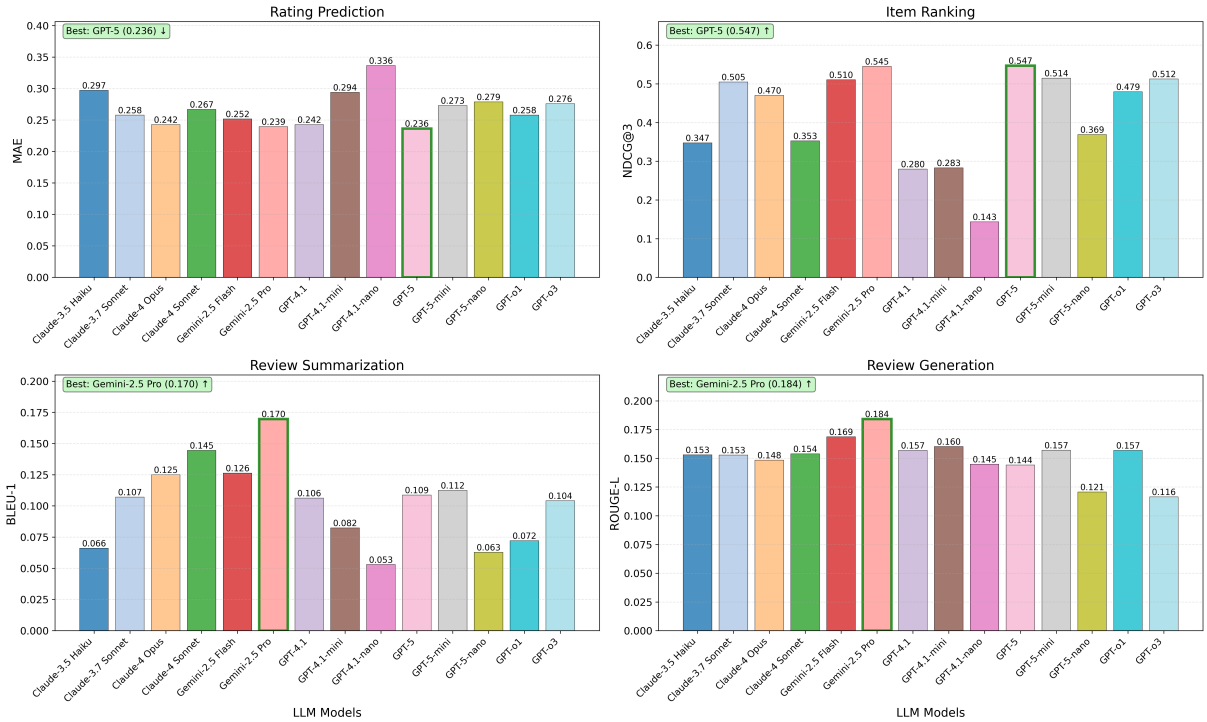

Figure 8: Performance comparisons of 14 frontier long-context LLMs on Home & Kitchen based on A-Mem.

```
User's Recent Review History:
1. [Formatted Interaction 1]
2. [Formatted Interaction 2]
. . .
n. [Formatted Interaction n]
```

```
Target Product to Rate:
[Formatted Target Item without Rating and
Review Text]

Based on this user's history and
preferences, predict the rating (1-5) they
would give to the target product.
```

Method	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Long-Context	GPT-5	**0.289**	**0.602**	**0.443**	**0.604**	0.182	0.082	0.123	0.145
	Claude-4 Sonnet	0.374	0.749	0.354	0.511	**0.228**	0.139	0.144	0.176
	Gemini-2.5 Pro	0.337	0.719	0.378	0.561	0.222	**0.141**	**0.161**	**0.200**
Mem0	GPT-5	**0.354**	**0.735**	**0.646**	**0.744**	0.209	0.115	0.133	0.153
	Claude-4 Sonnet	0.411	0.741	0.565	0.661	**0.238**	**0.145**	0.144	0.174
	Gemini-2.5 Pro	0.382	0.760	0.618	0.718	0.223	0.135	**0.160**	**0.192**
LoCoMo	GPT-5	**0.327**	0.716	**0.797**	**0.855**	0.188	0.098	0.141	0.171
	Claude-4 Sonnet	0.358	**0.636**	0.740	0.803	0.216	**0.141**	0.149	0.183
	Gemini-2.5 Pro	0.430	0.839	0.758	0.824	**0.219**	0.140	**0.167**	**0.203**
ReadAgent	GPT-5	**0.378**	0.735	**0.545**	**0.662**	0.203	0.102	0.132	0.157
	Claude-4 Sonnet	0.390	**0.704**	0.472	0.594	0.216	0.128	0.142	0.173
	Gemini-2.5 Pro	0.419	0.794	0.520	0.645	**0.224**	**0.131**	**0.158**	**0.189**
MemoryBank	GPT-5	**0.341**	0.716	0.663	0.763	0.201	0.114	0.143	0.172
	Claude-4 Sonnet	0.362	**0.660**	0.593	0.684	0.232	**0.153**	0.152	0.183
	Gemini-2.5 Pro	0.415	0.727	**0.703**	**0.801**	**0.236**	0.143	**0.170**	**0.202**
A-Mem	GPT-5	0.325	0.681	**0.504**	**0.650**	0.198	0.105	0.132	0.153
	Claude-4 Sonnet	**0.321**	**0.628**	0.435	0.560	**0.246**	**0.155**	0.149	0.183
	Gemini-2.5 Pro	0.362	0.762	0.492	0.643	0.235	**0.155**	**0.166**	**0.202**

Table 9: Comparison of memory methods on *Electronics* using representative LLMs.

```
Respond with ONLY a single integer between
1 and 5. Output only the number.
```

Item Ranking Prompt

```
You are a personalized item recommendation
system. Based on a user's purchase and
review history, rank the candidate items in
order of relevance (most likely to purchase
first).

User's Recent Purchase and Review History:
1. [Formatted Interaction 1]
2. [Formatted Interaction 2]
. . .
n. [Formatted Interaction n]

Candidate Items to Rank:
1. [Candidate Item 1]
2. [Candidate Item 2]
. . .
k. [Candidate Item k]

Based on this user's history and
preferences, rank the candidate items from
most to least relevant.
Respond with ONLY the ranked list of item
indices (e.g., 3, 1, 2, ...).
```

Review Summarization Prompt

```
You are a personalized review summarization
system. Based on a user's review history
and writing style, generate a concise
review title that summarizes their full
review.

User's Recent Review History (showing their
writing style and preferences):
1. [Formatted Interaction 1]
2. [Formatted Interaction 2]
. . .
n. [Formatted Interaction n]

Target Product and Review to Summarize:
[Formatted Target Item without Rating and
Review Text]
User's Rating: [Target Rating]/5
User's Full Review: [Target Review Text]

Based on this user's writing style and
the full review, generate a concise review
title.
The title should capture the main sentiment
and key points in the user's style.
Respond with ONLY the review title (3-10
words).
```

Review Generation Prompt

```
You are a personalized review generation
system. Based on a user's review history
and writing style, generate a detailed
review text for a product they rated.

User's Recent Review History (showing their
writing style and preferences):
1. [Formatted Interaction 1]
2. [Formatted Interaction 2]
. . .
n. [Formatted Interaction n]
```

Method	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Long-Context	GPT-5	**0.261**	**0.551**	**0.446**	**0.610**	0.185	0.077	0.134	0.140
	Claude-4 Sonnet	0.330	0.630	0.276	0.400	0.189	0.117	0.145	0.158
	Gemini-2.5 Pro	0.303	0.679	0.315	0.444	**0.222**	**0.133**	**0.178**	**0.207**
Mem0	GPT-5	**0.261**	**0.561**	0.606	0.707	0.221	0.122	0.151	0.163
	Claude-4 Sonnet	0.339	0.623	0.521	0.613	0.227	0.152	0.152	0.167
	Gemini-2.5 Pro	0.306	0.654	**0.652**	**0.746**	**0.231**	**0.155**	**0.184**	**0.215**
LoCoMo	GPT-5	**0.268**	**0.558**	**0.824**	**0.889**	0.219	0.121	0.157	0.176
	Claude-4 Sonnet	0.324	0.583	0.755	0.839	0.204	0.126	0.151	0.169
	Gemini-2.5 Pro	0.310	0.632	0.798	0.855	**0.241**	**0.164**	**0.191**	**0.215**
ReadAgent	GPT-5	**0.255**	**0.534**	0.409	**0.557**	0.184	0.091	0.149	0.163
	Claude-4 Sonnet	0.345	0.661	0.412	0.547	0.193	0.117	0.152	0.171
	Gemini-2.5 Pro	0.339	0.726	**0.430**	0.573	**0.215**	**0.132**	**0.180**	**0.211**
MemoryBank	GPT-5	**0.258**	**0.564**	0.688	0.778	0.210	0.119	0.158	0.177
	Claude-4 Sonnet	0.300	0.564	0.645	0.737	0.219	0.137	0.157	0.176
	Gemini-2.5 Pro	0.276	0.575	**0.764**	**0.849**	**0.235**	**0.166**	**0.191**	**0.219**
A-Mem	GPT-5	**0.236**	**0.505**	**0.412**	**0.547**	0.209	0.109	0.144	0.153
	Claude-4 Sonnet	0.267	0.539	0.239	0.353	0.220	0.145	0.154	0.168
	Gemini-2.5 Pro	0.239	0.531	0.403	0.545	**0.249**	**0.170**	**0.184**	**0.214**

Table 10: Comparison of memory methods on *Home & Kitchen* using representative LLMs.

```
Target Product to Review:
[Formatted Target Item without Rating and
Review Text]
User's Rating: [Target Rating]/5

Based on this user's writing style,
preferences, and the rating they gave,
generate a detailed review text.
The review should reflect the user's
typical style, length, and sentiment.
Respond with ONLY the review text.
```

Method	Model	Rating ↓		Ranking ↑		Summarization ↑		Generation ↑	
		MAE	RMSE	NDCG@1	NDCG@3	ROUGE-L	BLEU-1	ROUGE-L	BLEU-1
Long-Context	GPT-5	**0.279**	**0.597**	**0.315**	**0.435**	0.161	0.053	0.131	0.139
	Claude-4 Sonnet	0.432	0.767	0.203	0.304	0.196	**0.118**	0.138	0.148
	Gemini-2.5 Pro	0.323	0.665	0.266	0.417	**0.200**	0.110	**0.183**	**0.204**
Mem0	GPT-5	**0.331**	**0.635**	0.578	0.685	0.175	0.082	0.150	0.164
	Claude-4 Sonnet	0.361	0.672	0.583	0.683	0.197	0.111	0.151	0.163
	Gemini-2.5 Pro	0.353	0.704	**0.612**	**0.725**	**0.214**	**0.131**	**0.182**	**0.201**
LoCoMo	GPT-5	**0.315**	**0.633**	**0.823**	**0.897**	0.192	0.095	0.160	0.177
	Claude-4 Sonnet	0.362	0.635	0.802	0.885	0.198	0.115	0.158	0.172
	Gemini-2.5 Pro	0.375	0.670	0.814	0.889	**0.208**	**0.133**	**0.189**	**0.211**
ReadAgent	GPT-5	**0.339**	**0.633**	0.404	0.535	0.170	0.069	0.148	0.164
	Claude-4 Sonnet	0.427	0.703	0.380	0.525	**0.208**	0.107	0.149	0.164
	Gemini-2.5 Pro	0.365	0.688	**0.417**	**0.567**	0.206	**0.113**	**0.172**	**0.195**
MemoryBank	GPT-5	**0.276**	**0.582**	0.742	0.823	0.195	0.095	0.163	0.175
	Claude-4 Sonnet	0.318	0.629	0.664	0.771	0.219	0.135	0.161	0.176
	Gemini-2.5 Pro	0.349	0.688	**0.768**	**0.859**	**0.221**	**0.138**	**0.192**	**0.213**
A-Mem	GPT-5	**0.299**	**0.589**	**0.365**	**0.496**	0.186	0.082	0.145	0.156
	Claude-4 Sonnet	0.341	0.627	0.336	0.466	**0.230**	**0.142**	0.157	0.170
	Gemini-2.5 Pro	0.305	0.619	0.339	0.477	0.218	0.128	**0.188**	**0.213**

Table 11: Comparison of memory methods on *Personal Care* using representative LLMs.