

MEMORYREWARDBENCH: Benchmarking Reward Models for Long-Term Memory Management in Large Language Models

Zecheng Tang^{1,2*}, Baibei Ji^{1,2*}, Ruoxi Sun^{1,2}, Haitian Wang^{1,2}, Juntao Li^{1,2†}, Min Zhang¹

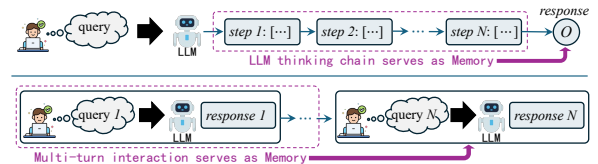
¹Soochow University ²LCM Laboratory
{zctang, bbji}@stu.suda.edu.cn {ljt, minzhang}@suda.edu.cn

Abstract

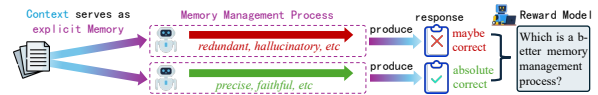
With the surge of AI-powered real-world applications, large language models (LLMs) increasingly adopt memory-centric mechanisms to handle multi-turn interactive tasks. Under this paradigm, intermediate processing trajectories, i.e., memory, are central to ensuring LLMs’ faithfulness and effectiveness. Scalable evaluation of the memory management process serves as the cornerstone for further enhancing model performance. Yet, appropriate evaluators for the assessment of memory are currently lacking. Thus, in this work, we seek to employ reward models (RMs) as evaluators and introduce **MemoryRewardBench**, the first benchmark to systematically study the capability of RMs in assessing LLM memory management processes. MemoryRewardBench covers both long-form comprehension and generation tasks that simulate real-world multi-turn interaction scenarios, featuring 10 distinct settings with different memory management patterns. Evaluations on 13 cutting-edge RMs indicate a diminishing performance gap between open-source and proprietary models, with newer-generation models consistently outperforming their predecessors regardless of parameter count. Additionally, we reveal the fundamental limitations of RMs in evaluating LLMs’ memory.¹

1 Introduction

Large Language Models (LLMs) have shown exceptional capabilities in comprehending contextual information (Xu et al., 2025a). In long-input scenarios, chunk-based *segmented processing* offers superior multi-turn scalability compared to window-dependent *holistic processing*. When engaging in multi-turn interactions with users or processing queries in multiple stages, LLMs adopt memory-centric strategies, i.e., maintaining and



(a) Illustration of treating context as LLM’s explicit memory.



(b) Demonstration of how RMs evaluate the LLM memory management process.

Figure 1: As LLM utilizes memory to propagate contextual information (Figure 1a), employing RMs to evaluate the LLM memory management process (Figure 1b) is crucial for providing dense supervision signals for further LLM optimization.

updating intermediate processed results (Luo et al., 2025) until completion (see Figure 1a). Under this paradigm, particularly when context accumulates to excessive length, **memory serves as the critical bridge between the past and present information**. Therefore, effective management of memory and its corresponding optimization is paramount to the model’s success (Xu et al., 2025b).

However, current training supervision signals that focus solely on model outputs while ignoring memory suffer from sparsity (Zheng et al., 2025), leaving a critical gap in scalable memory evaluation. Reward models (RMs) present a potentially effective solution, as we can employ them to directly assess LLM memory. Yet this is non-trivial without understanding their limitations. Thus, a fundamental question arises: *In the context of assessing LLM memories, how can we evaluate the capability of RMs, and what are their current boundaries?*

In this work, we introduce **MemoryRewardBench**, the first benchmark that is specifically designed to assess how effectively RMs judge the quality of long-term intermediate memories in LLMs. Unlike prior efforts that evaluate memory reten-

*Equal Contribution

†Corresponding author.

¹Code & data are available at: <https://github.com/LCM-Lab/MemoryRewardBench>.

Benchmark	Evaluation Target	Process Evaluation	Static vs. Dynamic	Context Length	Memory Abilities				
					DU	MR	KU	TR	GEN
LongBench (Bai et al., 2024)	LLM	✗	Static	0 ~ 64K	✗	✓	✗	✗	✗
RULER (Hsieh et al., 2024)	LLM	✗	Static	4K ~ 128K	✗	✓	✗	✗	✗
LongMemEval (Wu et al., 2024b)	LLM	✓	Dynamic	4K ~ 115K	✓	✓	✓	✓	✗
MemoryBank (Zhong et al., 2023)	LLM	✗	Static	0 ~ 5K	✓	✗	✗	✓	✗
LoCoMo (Maharana et al., 2024)	LLM	✗	Dynamic	4K ~ 16K	✓	✓	✗	✓	✓
MemBench (Tan et al., 2025)	LLM	✓	Dynamic	0 ~ 100K	✓	✓	✓	✓	✗
PerLTQA (Du et al., 2024)	LLM	✓	Static	1M	✓	✗	✗	✗	✗
MemoryRewardBench (ours)	RM	✓	Static & Dynamic	8K ~ 128K	✓	✓	✓	✓	✓

Table 1: Comparison of our benchmark with existing memory benchmarks, where **DU** denotes Dialogue Understanding, **MR** denotes Multi-hop Reasoning, **KU** denotes Knowledge Update, **TR** denotes Temporal Reasoning, **GEN** denotes Generation. More details and explanations are shown in Appendix A.

tion in LLMs directly or focus solely on either static or dynamic context evaluation(see Table 2), we are the first to shift the focus toward benchmarking RMs themselves while simultaneously covering both static and dynamic context evaluation². MemoryRewardBench includes both comprehension and generation tasks, encompassing 10 diverse memory management configurations across three representative tasks: long-context reasoning, multi-turn dialogue, and long-form generation. For each evaluation, RM is provided with the original context (from 8K to 128K tokens), two candidate memory management trajectories, and their respective outcomes. The RM’s task is to select the superior sample according to the criteria specified for each task, while also providing a justifying explanation (see Figure 1b).

To encourage RMs to *prioritize memory management quality over mere outcome correctness*, we design 2 evaluation criteria to decouple the quality of the memory management process from the correctness of the outcome:

Type 1 Outcome-based: the RM should prefer a memory management trajectory that leads to a correct outcome over one that results in an incorrect outcome.

Type 2 Process-based: both memory management trajectories yield correct final outcomes, but RM should prefer the one that demonstrates more accurate, concise, and logically coherent memory updates.

We select 13 cutting-edge and widely-used LLMs as RMs, comprising 3 proprietary models and 10 open-source models. Our results show

²**Static** denotes context that contains only a single-turn interaction, while **Dynamic** denotes context involving multi-turn interactions.

that the performance gap between open-source and proprietary models has further narrowed. Surprisingly, we also find that model performance does not monotonically scale with model size. Instead, we observe a pronounced *generational advantage*, whereby newer-generation models consistently outperform their predecessors regardless of parameter count, e.g., Qwen3-4B (Yang et al., 2025) surpasses the substantially larger Qwen2.5-7B-Instruct (Yang et al., 2024). Furthermore, we uncover several critical behavioral patterns in RMs, revealing both their capabilities and limitations in evaluating LLM memory management across diverse settings.

2 Related Work

2.1 Memory Management Evaluation

Existing memory management benchmarks for LLMs primarily evaluate the memories they produce and can be broadly classified into two categories. The first category directly examines how models retain, update, and regulate memory over time (Wu et al., 2024b; Deshpande et al., 2025; Tan et al., 2025). This includes narrative- or domain-specific variants like StoryBench (Wan and Ma, 2025), LoCoBench-Agent (Qiu et al., 2025), and PerLTQA (Du et al., 2024), which adopt structured external memory to assess model robustness or evolving memory dynamics, such as MemoryBank (Zhong et al., 2023), StuLife (Cai et al., 2025), StreamBench (Wu et al., 2024a), and EvoMemory (Wei et al., 2025). The second category evaluates memory through *final outcomes*, measuring long-term consistency in user modeling (Maharana et al., 2024; Tavakoli et al., 2025), persona tracking (Jiang et al., 2025), or preference evolution (Zhao et al., 2025). Extensions like LongMT-Bench+(Pan et al., 2025) probe long-range di-

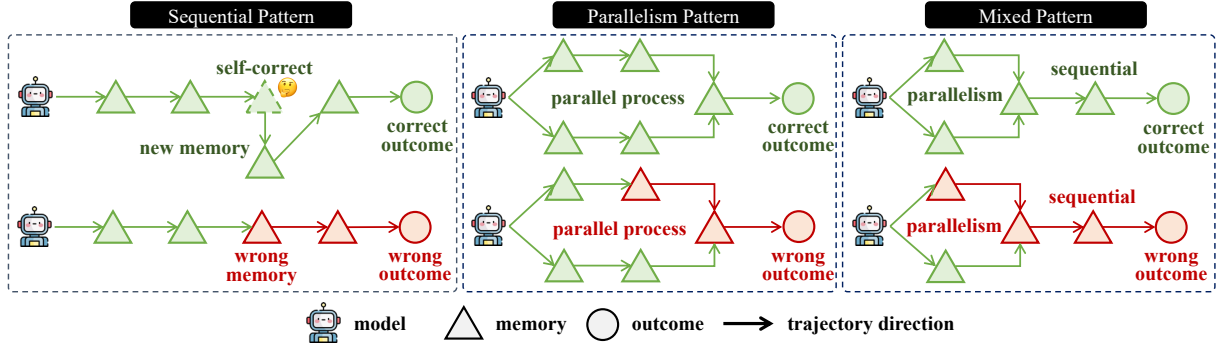


Figure 2: Illustrations of three memory management patterns. From left to right: Sequential pattern, Parallelism pattern, and Mixed pattern. Each pattern depicts both correct and incorrect memory update trajectories. For clarity, context chunks are omitted, and only intermediate memory states are shown. More details are shown in Appendix B.

alogue recall, while Minerva (Xia et al., 2025), MemoryAgentBench (Hu et al., 2025a), MemoryBench (Ai et al., 2025), and MeetingQA (Zhang et al., 2025) introduce memory-stress scenarios, stepwise fact accumulation, and large-scale context fidelity under realistic interaction settings. Despite extensive efforts, existing approaches heavily rely on rule-based heuristics or manual annotation. Yet, automated and scalable memory assessment paradigms based on RMs remain largely unexplored.

2.2 Reward Model

Reward models (RMs) serve as proxies for human-derived preferences, providing training signals that align language models with desired values and behaviors (Bai et al., 2022; Dubois et al., 2023; Li et al., 2023). Following the taxonomy introduced in previous work (Liu et al., 2024), RM paradigms can be broadly categorized into discriminative rewards (Dubois et al., 2023; Yuan et al., 2024; Dou et al., 2025), generative rewards (Zheng et al., 2023; Li et al., 2024; Tang et al., 2025a), and implicit rewards (Rafailov et al., 2024; Xu et al., 2025c). Among these, generative RMs directly leverage the generalization capabilities of LLMs to generate preference judgments, thereby enabling flexible and general-purpose reinforcement learning (Zhong et al., 2025; Yu et al., 2025). In this work, we primarily focus on generative RMs, as this paradigm is the only one that potentially supports memory evaluation.

3 Introduce MemoryRewardBench

Notably, different memory management strategies are adopted across tasks, and each task involves its distinct memory management patterns. There-

fore, we first identify three memory management patterns in § 3.1 and then define three task formulations in § 3.2. Finally, we outline the data collection and benchmark construction process in § 3.3.

3.1 Memory Management Pattern

As shown in Figure 2, given a model Φ and a sequence that is divided into chunks $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$, the intermediate memories $\mathcal{M} = \{m_1, m_2, \dots, m_n\}$ are managed according to one of the following two atomic patterns:

- **Sequential Pattern:** the memory state evolves step-by-step along the chunks, where $m_1 = \Phi(c_1)$, $m_t = \Phi(m_{t-1}, c_t)$ for $t = 2, \dots, n$, and the final outcome can be obtained from the final memory m_n .
- **Parallelism Pattern:** the input context is partitioned into k independent groups: $\mathcal{C} = \{\mathcal{G}_1, \dots, \mathcal{G}_k\}$, and where each group $\mathcal{G}_j = \{c_{j,1}, \dots, c_{j,n_j}\}$ is processed by Φ in parallel. Within each group, memory states are updated sequentially according to the **Sequential Pattern**, yielding each group’s final memory state $m^{(j)}$. The final outcome is obtained by aggregating all $m^{(j)}$ through a fusion operation g : $o = g(m^{(1)}, \dots, m^{(k)})$.

Notably, any memory management can be categorized as either an instance of the above two patterns, or a composition of both, i.e., the **Mixed Pattern**.

3.2 Task Overview

Overall, the goal of MemoryRewardBench is to evaluate how well RMs can assess and explain the quality of an LLM’s long-term memory management according to predefined criteria. In this section, we outline *how memory management manifests across different task settings*, and formalize *the criteria*

Task Type	Setting	Source	Length Distribution					Total
			8k	16k	32k	64k	128k	
Long-context Reasoning	Sequential-Noise	BABILong (Kuratov et al., 2024), LongMiT (Chen et al., 2025)	101	44	43	36	31	255
	Sequential-Drop		35	22	22	40	15	134
	Mixed-Noise		22	33	49	46	34	184
	Mixed-Drop		19	65	72	43	28	227
Multi-turn Dialogue Understanding	Mem0-Out	LoCoMo (Maharana et al., 2024), MemoryAgentBench (Hu et al., 2025a)	27	27	42	48	23	167
	Mem0-Mem		25	25	41	47	21	159
	A-Mem-Out		42	42	48	50	47	229
	A-Mem-Mem		48	45	49	53	50	245
Long-form Generation	Parallel	LongEval (Wu et al., 2025), LongGenBench (Wu et al., 2024c)	49	152	147	67	42	457
	Sequential	LongProc (Ye et al., 2025)	51	48	53	133	58	343
Statistic	10 settings	-	419	503	566	563	349	2400

Table 2: Distribution and statistics of tasks in MemoryRewardBench, where the settings (the “Setting” column) are named and defined according to the benchmark construction process described in § 3.3.

that distinguish correct and incorrect memory management behaviors. We focus on 3 representative tasks that require memory management:

- (1) **Long-context reasoning:** the model processes a sequence of chunks $\mathcal{C}$ to extract question-relevant evidence, incrementally updating its memory, and finally produces the outcome;
- (2) **Multi-turn dialogue understanding:** given an extremely long conversation, e.g., hundreds of turns, the model maintains a persistent memory to record the dialogue and finally retrieves relevant dialogue turns to answer queries about a specific point in the dialogue;
- (3) **Long-form generation:** Given an instruction with explicit constraints, the model generates structured content over multiple steps, where intermediate generations serve as memory that must adhere to the specified constraints.

RM Evaluation Criteria The goal of RMs is to select the better trajectory from two candidates. The inferior trajectory may suffer from various defects, including redundancy, incoherence, and inaccuracy. For comprehension-oriented tasks (1) and (2), the evaluation criteria for RMs are: (i) *outcome-based*—whether the final outcome is accurate; and (ii) *process-based*—whether the intermediate memory is concise and relevant to the outcome (An example is provided in Figure 8). For the generation-oriented task (3), the key criterion is whether the intermediate memory complies with the constraints specified in the instruction. In short, even when two samples produce equally correct outcomes, one may still exhibit a superior memory management trajectory.

3.3 Benchmark Construction

Table 2 summarizes all task types, settings, data sources, and length distributions in our benchmark. For each task type, we outline how we construct pairs that exhibit chosen and rejected memory management below. Due to space limitations, we outline the core benchmark construction process below and provide further details in Appendix B.

Long-context Reasoning For a long input sequence, we employ the Sequential and the Mixed memory management patterns described in § 3.1 to construct memory management trajectories. We select instances with correct final outcomes as the *chosen* samples. Then, we obtain a *rejected* counterpart by introducing one of two error-inducing perturbations on the chosen sample: (1) **NOISE**: injecting redundant, irrelevant information into the memory trajectory, or (2) **DROP**: dropping partial critical information from the input sequence. Both perturbations interfere with the memory management process, potentially compromising the final outcome. We regard all such memory management trajectories as rejected samples.

Multi-turn Dialogue Understanding Strong inter-turn dependencies in this task require both comprehensive information preservation and robust temporal tracking to retrieve query-relevant memories. Accordingly, only the sequential memory-management pattern is applicable. To construct controlled preference pairs, we focus on methods whose persistent memory states can be directly manipulated. Conventional RAG systems instead rely on external retrieval and would require

Models	Long-context Reasoning					Multi-turn Dialogue Understanding					Long-form Generation			Avg.
	S-Noise	S-Drop	M-Noise	M-Drop	Avg.	MO	MM	AO	AM	Avg.	S	P	Avg.	
🏢 Proprietary Models														
Claude-Opus-4.5	47.84	91.05	52.72	92.51	68.88	64.07	45.91	70.61	84.28	68.25	89.50	83.97	87.13	74.75
Gemini3-Pro	54.51	88.81	48.91	88.99	68.75	61.08	45.28	71.43	82.10	67.13	75.06	84.26	79.00	71.63
Qwen3-Max	42.75	87.31	42.94	86.78	62.75	48.50	42.14	62.86	74.24	59.00	85.12	76.97	81.63	67.79
🏠 Open-source Models														
GLM4.5-106A12B	54.65±0.30	86.64±0.58	49.05±0.39	90.29±0.32	68.83±0.23	52.40±0.28	42.82±0.30	59.69±0.32	66.34±0.39	56.85±0.23	80.26±0.51	77.92±0.28	79.26±0.41	68.31±0.12
Qwen3-235A22B	39.15±0.57	92.08±0.28	41.05±0.41	89.04±0.07	62.61±0.33	57.65±0.45	40.51±0.36	71.85±0.49	58.39±0.16	58.53±0.14	85.09±0.26	71.66±0.27	79.33±0.08	66.83±0.18
Qwen2.5-72B	38.36±0.26	74.65±0.44	52.72±0.43	88.58±0.37	61.99±0.30	37.25±0.35	26.94±0.29	37.71±0.53	44.89±0.35	37.67±0.24	57.01±0.29	54.96±0.38	56.13±0.19	51.93±0.23
Qwen3-32B	46.97±0.43	82.09±0.33	55.88±0.41	81.90±0.30	64.81±0.35	48.24±0.37	43.88±0.25	58.73±0.54	56.16±0.24	52.74±0.19	71.25±0.43	70.49±0.51	70.92±0.46	62.83±0.31
Qwen3-14B	49.78±0.35	86.25±0.33	57.44±0.65	82.10±0.47	66.82±0.29	44.92±0.44	40.61±0.37	50.99±0.29	49.42±0.38	47.18±0.31	70.78±0.63	63.54±0.42	67.68±0.46	60.56±0.35
Qwen3-8B	53.32±0.33	68.74±0.36	51.04±0.07	72.10±0.22	60.71±0.21	28.02±0.17	26.95±0.13	47.98±0.13	55.43±0.33	41.92±0.18	66.00±0.11	72.80±0.35	68.92±0.21	57.18±0.19
Qwen2.5-7B	28.46±0.40	38.58±0.33	38.88±0.35	48.67±0.33	38.23±0.35	43.60±0.34	32.82±0.45	30.18±0.35	23.34±0.34	31.41±0.35	47.36±0.41	40.24±0.36	44.30±0.38	38.00±0.36
Qwen3-4B	52.92±0.58	70.57±0.47	45.83±0.52	68.15±0.81	58.57±0.61	37.62±0.57	29.87±0.67	42.75±0.55	49.12±0.57	41.07±0.56	55.89±0.50	56.63±0.73	56.20±0.60	51.95±0.57
Llama3.3-70B	46.87±0.34	70.67±0.33	43.63±0.50	76.13±0.15	58.41±0.31	52.30±0.28	41.84±0.25	52.95±0.50	54.89±0.29	51.20±0.32	64.18±0.51	63.17±0.19	63.76±0.36	57.79±0.22
Llama3.1-8B	37.38±0.21	53.51±0.21	42.88±0.18	61.64±0.05	48.23±0.11	41.25±0.09	43.87±0.22	35.41±0.14	35.88±0.10	38.46±0.07	46.24±0.09	43.08±0.09	44.88±0.01	43.86±0.05

Table 3: Results on MemoryRewardBench, where “S” and “M” refer to “Sequential” and “Mixed” respectively. “MO”, “MM”, “AO” and “AM” refer to “Mem0-OUT”, “Mem0-MEM”, “A-Mem-OUT” and “A-Mem-MEM”, respectively. For each metric, the best-performing result is **bolded**, and the second-best is underlined.

a separate failure-injection protocol; we therefore exclude RAG-based methods from this task, and adopt two dialogue memory-management methods: A-Mem (Xu et al., 2025b) and Mem0 (Chhikara et al., 2025). Both methods dynamically update a summary of the dialogue history after each turn; however, A-Mem additionally annotates each summary with semantic tags (e.g., “personal-communication”) to enable efficient memory call-back. To construct preference pairs, we select the *chosen* sample based on the correctness of the final outcome. The *rejected* sample is constructed by skipping memory updates for more than one turn of dialogue. Notably, even when a rejected sample produces a correct final response, its intermediate memory is suboptimal due to missing or delayed updates. For clarity, we categorize such examples into two types: samples with correct final outcomes but flawed memory management are labeled as MEM, while those with incorrect final outcomes are labeled as OUT.

Long-form Generation Unlike the aforementioned comprehension tasks that provide both a question and a reference context, long-form generation supplies only an instruction with embedded constraints, requiring the model to generate content that satisfies all specified sub-constraints. The generation process can follow either a Sequential or a Parallel memory management pattern. In both cases, the question is decomposed into a sequence of step-wise constraints, and the model generates content at each step to satisfy the corresponding constraint. The intermediate outputs are maintained as memory, where each generation is incorporated into the historical memory and conditions subsequent steps. After all constraints are

processed, the accumulated memory states are concatenated to produce the final *chosen* response. To construct a *rejected* generation, we perturb the instruction, such as dropping key constraints or injecting interference content, to cause LLMs to generate incorrect intermediate memory.

4 Evaluation

4.1 Settings

As there are currently no RMs specifically designed for evaluating memory management processes, we experiment with 13 cutting-edge LLMs as proxy RMs, including 3 proprietary models: Claude-Opus-4.5 (Anthropic, 2025), Gemini-3.0-Pro (Google DeepMind, 2025), and Qwen3-Max (Qwen, 2025) and 10 open-source models spanning the Qwen2.5 series (Yang et al., 2024), Qwen3 series (Yang et al., 2025), and Llama3 series (Dubey et al., 2024), as well as GLM4.5-Air (GLM4.5-106A12B) (Zeng et al., 2025). All RMs support a context window of at least 128K tokens. We calculate the *judgment accuracy* for each RM. Notably, the theoretical accuracy of random guessing is 50%. Yet, in practice, some RM outputs cannot be parsed, and we treat such cases as incorrect, resulting in observed accuracies falling below 50%. Due to resource constraints, all closed-source models were evaluated once with seed 42, whereas each open-source model is evaluated three times using seeds 42, 43, and 44, with the mean and standard deviation reported. The evaluation implementation details are provided in Appendix C.

4.2 Overall Observation

We report RMs’ judgment accuracy in Table 3.

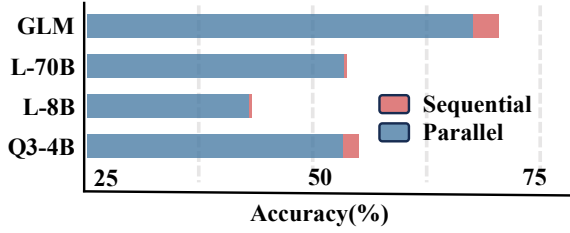


Figure 3: Performance comparison between Sequential and Parallel memory management patterns on the long-context reasoning and long-form generation tasks.

Proprietary vs. Open-source RMs Overall, *proprietary models maintain a performance advantage*, with Claude-Opus-4.5 achieving the highest average score of 74.75, followed closely by Gemini3-Pro at 71.63. However, the performance gap between proprietary and open-source models has narrowed. GLM4.5-106A12B emerges as the strongest open-source model in our evaluation with an average score of 68.31, even outperforming the proprietary Qwen3-Max (67.79). Specifically, we find that proprietary models remain superior in handling complex temporal dependencies and enforcing long-term constraints, as evidenced by their dominance in multi-turn dialogue and long-form generation tasks.

Open-source RMs Analysis The performance of open-source RMs reveals a pronounced *decoupling between parameter count and practical capability*, underscoring the impact of more efficient training data curation and increasingly effective post-training strategies in the latest generation of models. This trend is particularly evident in the Qwen3 series, where Qwen3-32B (62.83) outstrips much larger models such as Llama3.3-70B (57.79). From another perspective, the newer-generation models significantly outperform their predecessors, e.g., Qwen3-8B (57.18) achieves a substantial performance gain over the previous-generation Qwen2.5-7B (38.00). This improvement is likely attributable to advances in context-scaling training and post-training strategies adopted in newer models, which may foster more robust reasoning processes that align more closely with the judgment-and-explanation paradigm required by RM evaluation.

Cross-Task Capability Characterization The comparison across task categories reveals differences in task difficulty and model strengths. Multi-turn dialogue is the most challenging task, consis-

tently yielding lower RM scores due to the need for RMs to accurately perceive conversational state transitions in order to assess the correctness of intermediate memory. Long-form generation is moderately difficult, as it requires RMs to assess whether memory updating process exhibits sustained adherence to global constraints throughout the generation process. On both tasks, proprietary RMs maintain a performance lead, and cutting-edge open-source RMs demonstrate competitive results. In contrast, Long-form Generation appears to be the most tractable task, achieving the highest overall scores across most RMs. This suggests that retrieving and reasoning over static information has become a relatively mature capability for current LLMs, and *effective management of dynamic memory and long-range constraints* remains the key factor distinguishing top-performing RMs. Further details regarding the sources of task difficulty are provided in Appendix E.6

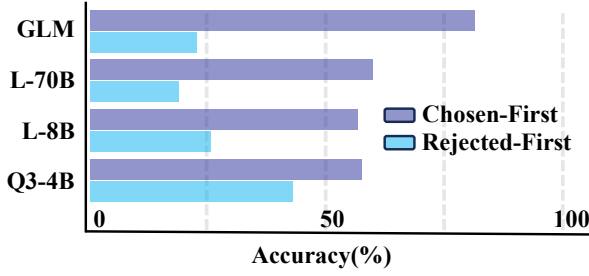
5 Ablation Study

In this section, we analyze RM behavior from four perspectives: (1) LLM memory management patterns (§ 5.1); (2) RM evaluation criteria: distinguishing between outcome-based and process-based judgments as well as examining RM robustness to global constraints (§ 5.2); (3) RM sensitivity to memory management trajectory length (§ 5.3); and (4) the impact of memory-enhancement strategies on RM performance (§ 5.4). We hope the above experiments provide insights for applying and improving RMs for evaluating LLM memory.

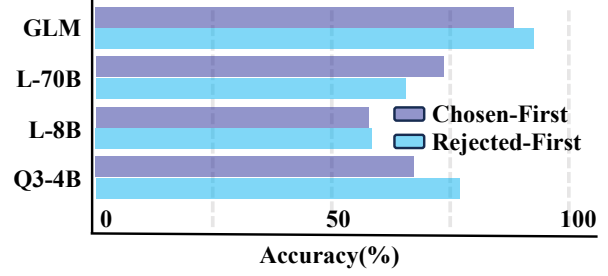
RM Selection and Notation We primarily select the following RMs for analysis and introduce shorthand notations for clarity: GLM-4.5-106A12B (GLM), Llama-3.3-70B-Instruct (L-70B), Llama-3.1-8B-Instruct (L-8B), and various sizes of Qwen3 (Q3), e.g., Q3-4B denotes Qwen3-4B.

5.1 Effect of Memory Management Patterns

We compare the performance of RMs under Sequential and Parallelism memory management patterns on long-context reasoning and long-form generation tasks. As shown in Figure 3, results show that RMs achieve significantly higher accuracy under the Sequential pattern. This suggests that *current RMs exhibit a stronger preference for progressive, step-by-step reasoning processes*, which aligns more closely with the causal structures commonly present in models’ training data and lan-



(a) Both chosen and rejected samples have correct outcome, but rejected one has a redundant memory management trajectory.



(b) Chosen sample has correct outcome and rejected sample has wrong outcome.

Figure 4: Comparison between process-based and outcome-based reward criteria. Chosen-First indicates that the chosen sample is presented before the rejected sample in the input context fed to the RM, and vice versa.

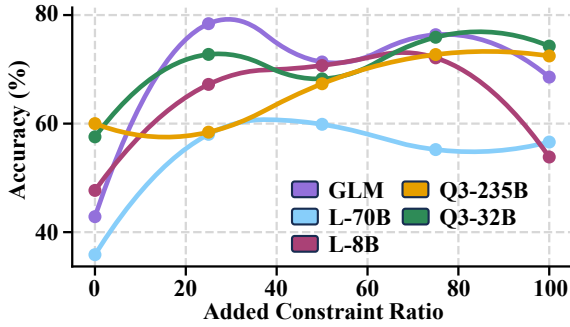


Figure 5: Performance trends of RMs with increasing constraint density in long-form generation instructions.

guage modeling (Jiao et al., 2025). In contrast, RMs struggle to effectively evaluate outputs generated through parallel processing and subsequent merging, highlighting a notable limitation and a promising direction for future improvement. More results are provided in Appendix E.1.

5.2 Effect of RM Evaluation Criteria

Outcome- vs. Process-based Criterion Evaluating memory management requires RMs to assess not only outcomes but also the quality of intermediate memory states. To investigate whether RMs prioritize outcome-based or process-based signals when evaluating memory management, we compare their behavior under outcome-based and process-based reward modeling paradigms. We adopt **RM evaluation consistency** as the primary metric. Specifically, we swap the positions of the chosen and rejected samples in the RM’s input context and evaluate the RM twice: once with the original ordering and once with the reversed ordering, to assess whether the RM’s preference remains stable under this positional perturbation. We perform the above evaluation under two settings: (a) *process-based*, where both outcomes are correct

but differ in the quality of their memory management trajectories; and (b) *outcome-based*, where only one sample produces a correct outcome. As shown in Figure 4, RMs exhibit inconsistency in the process-based setting, displaying a positional bias that favors samples appearing earlier in the input context. In contrast, under the outcome-based setting, RMs show robust and consistent preferences aligned with the ground truth³. The position sensitivity observed in this setting should be distinguished from the conventional candidate-order bias reported in prior LLM-as-a-judge work. In MT-Bench, a judge may favor a response simply because it is presented in a particular candidate slot, and its preference can change when the two candidates are swapped (Zheng et al., 2023). In contrast, MemoryRewardBench evaluates long memory-management trajectories, where the relevant positional factor is the location of critical factual evidence within the trajectory rather than the presentation order of candidate responses. The RM must locate and integrate distributed evidence across long-horizon memory state. Therefore, although both settings involve positional sensitivity, our finding concerns context retrieval and factual reasoning rather than a pure order preference.

Adherence to Global Constraint RMs are expected to rigorously evaluate outcomes based on all constraints specified in the instruction. To assess RMs’ global constraint coherence, we evaluate performance under progressively increasing constraint densities, ranging from topic-only prompts to fully specified, multi-constraint instructions. Implementation details are provided in Appendix E.4. As shown in Figure 5, RM performance initially

³In our main experiments, the chosen and rejected samples are randomly ordered, which mitigates the impact of positional bias and ensures the overall evaluation results remain reliable.

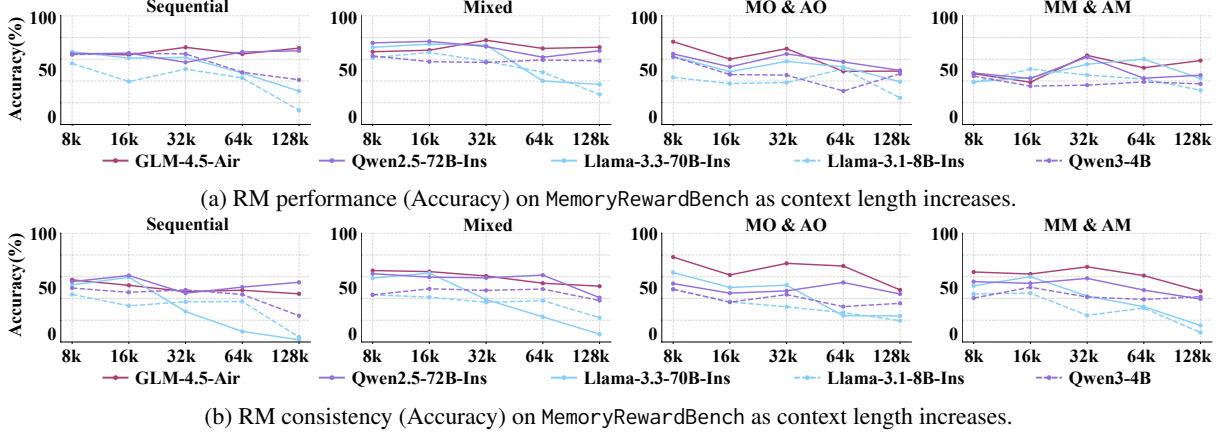


Figure 6: Trends in RM performance and consistency with respect to memory management trajectory length. The 1st and 2nd columns correspond to the *Long-context Reasoning* task, the 3rd column to the *Multi-turn Dialogue Understanding* task (average score), and the 4th column to the *Long-form Generation* task (average score).

improves as more instructional detail is incorporated, peaking at approximately 25% constraint density. At this level, RMs benefit from sufficient explicit criteria to ground their judgments in concrete instruction-following signals. However, further increases in constraint density do not yield continued gains; instead, performance plateaus or even declines. This suggests that current RMs are only partially capable of leveraging dense, multi-faceted constraints to assess memory fidelity.

5.3 Effect of Memory Management Trajectory Length

We first plot the trend of RM performance as context length increases in Figure 6a. Within a context length of up to 64K tokens, most RMs can maintain accuracy above 50%. Then, we evaluate RM consistency across different context length intervals using the position-swapping protocol described in § 5.2. Our results show that only GLM-4.5-Air and Qwen2.5-72B-Instruct maintain stable performance, achieving above 50% accuracy in most context lengths. In contrast, other models fail to maintain accuracy above 50%, exhibiting pronounced inconsistency once the context length exceeds 32K tokens. While this degradation can be attributed to limited parameter size, e.g., Qwen3-4B, models from the Llama family, particularly Llama-3.3-70B-Instruct, exhibit severe performance collapse at 64K and 128K context length, despite their substantially larger parameter count. We provide a detailed case analysis of such abnormal behavior and report comprehensive model performance across all tasks in Appendix E.2.

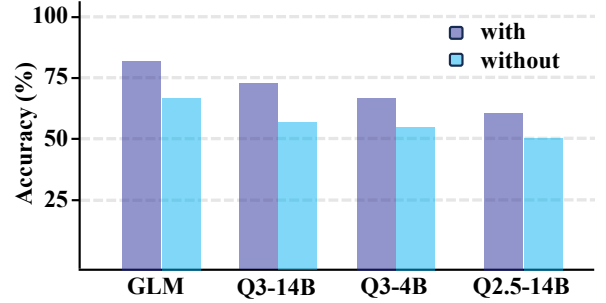


Figure 7: Comparison on multi-turn dialogue understanding task with and without auxiliary signals.

5.4 Effect of Memory Augmentation Strategy

Finally, we evaluate RMs under augmented memory management settings, aligning with the core focus of recent work that seeks to improve performance by introducing additional constraints or enhancements (Le et al., 2025). We conduct experiments in the challenging multi-turn dialogue understanding task. One way to enhance memory management in multi-turn dialogue understanding is to annotate each memory update with a semantic tag, e.g., personal-communication, that characterizes the contextual nature of the dialogue segment (Xu et al., 2025b). To investigate the impact of such auxiliary signals on RMs' evaluation accuracy, we compare RM performance under two settings: memory updates with explicit tags and memory updates without tags. We show the above data structure in Appendix E.5. As shown in Figure 7, incorporating auxiliary signals consistently improves the accuracy of RMs in evaluating memory management quality. Combined with the analysis in § 5.3, semantic tags provide RMs with

concise, high-level summaries of dialogue context, thereby enabling more reliable judgment without requiring the model to parse potentially redundant or verbose memory update trajectories.

6 Conclusion

Automatically evaluating the long-term memory management process of LLMs is essential. In this work, we introduce MemoryRewardBench, the first benchmark to systematically assess how effectively current RMs evaluate memory management. Our evaluation shows that open-source RMs have largely closed the gap with proprietary models on long-context reasoning. Our analysis further highlights both the strengths and fundamental limitations of current RMs in evaluating LLM memory management. We hope MemoryRewardBench provides a valuable benchmark and offers practical guidance for improving reward modeling and advancing memory-centric LLMs.

Limitations

MemoryRewardBench is the first benchmark designed to assess RM performance in evaluating long-term memory management in LLMs. Several limitations remain and motivate future work.

- (1) **Covering VLM Memory Evaluation:** While our benchmark primarily targets memory evaluation for language models, memory management is equally crucial for vision-language models (VLMs) (Lin et al., 2025). In settings that require reasoning over multiple images (Zhou et al., 2025), where input sequences can become extremely long, explicit memory mechanisms can be leveraged to support intermediate reasoning and information aggregation. Extending RM-based memory evaluation to such multimodal scenarios remains an open and challenging problem, which we leave for future work.
- (2) **Involving realistic tasks:** Our evaluation is based on synthetic tasks designed to systematically cover different memory management patterns. In real-world applications, however, memory behaviors are often more complex and heterogeneous, e.g., interaction with the environment (Hu et al., 2025b), potentially requiring task-specific designs and even manual annotation of memory states. Constructing

such realistic benchmarks would entail substantial (human) effort, and we also leave this to future work.

- (3) **Architectural coverage:** Our benchmark construction does not cover RAG-based retrieval pipelines, explicit graph-memory systems, or episodic/event-memory architectures. Although A-Mem maintains dynamically linked memory notes, this does not constitute an evaluation of explicit entity-relation graph memories. Moreover, conventional RAG pipelines may require additional instrumentation or a different failure-injection protocol to study persistent memory-write and memory-update failures in a manner comparable to our setting. Therefore, our findings should be interpreted within the evaluated systems; extending the benchmark to these alternative memory architectures is an important direction for future work.
- (4) **Limited Downstream Validation:** MemoryRewardBench evaluates whether reward models can distinguish high-quality from flawed memory-management trajectories. However, we do not yet establish whether stronger benchmark performance translates into improved end-to-end agent behavior or downstream task success. Thus, the scores should be interpreted as indicators of memory-evaluation capability rather than direct evidence of downstream utility. Future work should validate this relationship across diverse agent tasks and memory architectures.

Acknowledgments

We want to thank all the anonymous reviewers for their valuable comments. This work was supported by the National Natural Science Foundation of China (NSFC No. 62576232) and Key Laboratory of Data Intelligence and Advanced Computing, Soochow University.

References

- Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. 2025. Memory-bench: A benchmark for memory and continual learning in llm systems. *arXiv preprint arXiv:2510.17281*.
- Andrei Alexandru, Antonia Calvi, Henry Broomfield, Jackson Golden, Kyle Dai, Mathias Leys, Maurice

- Burger, Max Bartolo, Roman Engeler, Sashank Pisupati, and 1 others. 2025. Atla selene mini: A general purpose evaluation model. *arXiv preprint arXiv:2501.17195*.
- Anthropic. 2025. Introducing claude opus 4.5. <https://www.anthropic.com/news/claude-opus-4-5>.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, and 1 others. 2024. Longbench: A bilingual, multitask benchmark for long context understanding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3119–3137.
- Yuxuan Cai, Yipeng Hao, Jie Zhou, Hang Yan, Zhikai Lei, Rui Zhen, Zhenhua Han, Yutao Yang, Junsong Li, Qianjun Pan, Tianyu Huai, Qin Chen, Xin Li, Kai Chen, Bo Zhang, Xipeng Qiu, and Liang He. 2025. [Building self-evolving agents via experience-driven lifelong learning: A framework and benchmark](#). *Preprint*, arXiv:2508.19005.
- Zhi Chen, Qiguang Chen, Libo Qin, Qipeng Guo, Haijun Lv, Yicheng Zou, Hang Yan, Kai Chen, and Dahua Lin. 2025. What are the essential factors in crafting effective long context multi-hop instruction datasets? insights and best practices. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27129–27151.
- 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*.
- Darshan Girish Deshpande, Varun Prashant Gangal, Hersh Mehta, Jędrzej Rostańiec, Anand Kannappan, Rebecca Qian, and Peng Wang. 2025. [MEM-TRACK: Evaluating long-term memory and state tracking in multi-platform dynamic agent environments](#). In *Workshop on Scaling Environments for Agents*.
- Shihan Dou, Shichun Liu, Yuming Yang, Yicheng Zou, Yunhua Zhou, Shuhao Xing, Chenhao Huang, Qiming Ge, Demin Song, Haijun Lv, and 1 others. 2025. Pre-trained policy discriminators are general reward models. *arXiv preprint arXiv:2507.05197*.
- Yiming Du, Hongru Wang, Zhengyi Zhao, Bin Liang, Baojun Wang, Wanjuan Zhong, Zezhong Wang, and Kam-Fai Wong. 2024. [Perltqa: A personal long-term memory dataset for memory classification, retrieval, and synthesis in question answering](#). *CoRR*, abs/2402.16288.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.
- Google DeepMind. 2025. Gemini 3 pro. <https://deepmind.google/models/gemini/pro/>.
- Cheng-Ping Hsieh, Simeng Sun, Samuel Krizan, Shantanu Acharya, Dima Rekesch, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. Ruler: What’s the real context size of your long-context language models? *CoRR*.
- Yuanzhe Hu, Yu Wang, and Julian McAuley. 2025a. [Evaluating memory in LLM agents via incremental multi-turn interactions](#). In *ICML 2025 Workshop on Long-Context Foundation Models*.
- Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, and 1 others. 2025b. Memory in the age of ai agents. *arXiv preprint arXiv:2512.13564*.
- Bowen Jiang, Yuan Yuan, Maohao Shen, Zhuoqun Hao, Zhangchen Xu, Zichen Chen, Ziyi Liu, Anvesh Rao Vijjini, Jiashu He, Hanchao Yu, Radha Poovendran, Gregory Wornell, Lyle Ungar, Dan Roth, Sihao Chen, and Camillo Jose Taylor. 2025. [Personamem-v2: Towards personalized intelligence via learning implicit user personas and agentic memory](#). *Preprint*, arXiv:2512.06688.
- Yizhu Jiao, Jiaqi Zeng, Julien Veron Vialard, Oleksii Kuchaiev, Jiawei Han, and Olivier Delalleau. 2025. Think twice: Branch-and-rethink reasoning reward model. *arXiv preprint arXiv:2510.23596*.
- Yury Kuratov, Aydar Bulatov, Petr Anokhin, Ivan Rodkin, Dmitry Sorokin, Artyom Sorokin, and Mikhail Burtsev. 2024. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554.
- Hung Le, Dung Nguyen, Kien Do, Sunil Gupta, and Svetha Venkatesh. 2025. Stable hadamard memory: Revitalizing memory-augmented agents for reinforcement learning. In *The Thirteenth International Conference on Learning Representations*.
- Junlong Li, Shichao Sun, Weizhe Yuan, Run-Ze Fan, Pengfei Liu, and 1 others. 2024. Generative judge for evaluating alignment. In *The Twelfth International Conference on Learning Representations*.

- Zihao Li, Zhuoran Yang, and Mengdi Wang. 2023. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*.
- Zeteng Lin, Xingxing Li, Wen You, Xiaoyang Li, Zehan Lu, Yujun Cai, and Jing Tang. 2025. Iut-plugin: A plug-in tool for interleaved image-text generation. *arXiv preprint arXiv:2510.10969*.
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.
- Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiakai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, and 1 others. 2026. Skywork-reward-v2: Scaling preference data curation via human-ai synergy. In *International Conference on Learning Representations*, volume 2026, pages 133805–133838.
- 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.
- Zhuoshi Pan, Qianhui Wu, Huiqiang Jiang, Xufang Luo, Hao Cheng, Dongsheng Li, Yuqing Yang, Chin-Yew Lin, H. Vicky Zhao, Lili Qiu, and Jianfeng Gao. 2025. On memory construction and retrieval for personalized conversational agents. *Preprint*, arXiv:2502.05589.
- Jielin Qiu, Zuxin Liu, Zhiwei Liu, Rithesh Murthy, Jianguo Zhang, Haolin Chen, Shiyu Wang, Ming Zhu, Liangwei Yang, Juntao Tan, Roshan Ram, Akshara Prabhakar, Tulika Awalgaonkar, Zixiang Chen, Zhepeng Cen, Cheng Qian, Shelby Heinecke, Weiran Yao, Silvio Savarese, and 2 others. 2025. Locobench-agent: An interactive benchmark for llm agents in long-context software engineering. *Preprint*, arXiv:2511.13998.
- Qwen. 2025. Qwen3-max. <https://qwen.ai/blog?id=qwen3-max>.
- Rafael Rafailov, Joey Hejna, Ryan Park, and Chelsea Finn. 2024. From r to q^* : Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*.
- 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 *ACL (Findings)*, pages 19336–19352.
- Zecheng Tang, Baibei Ji, Quantong Qiu, Haitian Wang, Xiaobo Liang, Juntao Li, and Min Zhang. 2025a. Longrm: Revealing and unlocking the context boundary of reward modeling. *arXiv preprint arXiv:2510.06915*.
- Zecheng Tang, Haitian Wang, Quantong Qiu, Baibei Ji, Ruoxi Sun, Keyan Zhou, Juntao Li, and Min Zhang. 2025b. Loom-scope: a comprehensive and efficient long-context model evaluation framework. *arXiv preprint arXiv:2507.04723*.
- Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J Ross Mitchell. 2025. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *arXiv preprint arXiv:2510.27246*.
- Luanbo Wan and Weizhi Ma. 2025. Storybench: A dynamic benchmark for evaluating long-term memory with multi turns. *Preprint*, arXiv:2506.13356.
- Tianxin Wei, Naveen 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.
- Cheng-Kuang Wu, Zhi Rui Tam, Chieh-Yen Lin, Yun-Nung Chen, and Hung yi Lee. 2024a. Streambench: Towards benchmarking continuous improvement of language agents. *Preprint*, arXiv:2406.08747.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2024b. Longmemeval: Benchmarking chat assistants on long-term interactive memory. *CoRR*, abs/2410.10813.
- Siwei Wu, Yizhi Li, Xingwei Qu, Rishi Ravikumar, Yucheng Li, Tyler Loakman, Shanghaoran Quan, Xiaoyong Wei, Riza Batista-Navarro, and Chenghua Lin. 2025. Longeval: A comprehensive analysis of long-text generation through a plan-based paradigm. *arXiv preprint arXiv:2502.19103*.
- Yuhao Wu, Ming Shan Hee, Zhiqing Hu, and Roy Ka-Wei Lee. 2024c. Longgenbench: Benchmarking long-form generation in long context llms. *arXiv preprint arXiv:2409.02076*.
- Menglin Xia, Victor Ruehle, Saravan Rajmohan, and Reza Shokri. 2025. Minerva: A programmable memory test benchmark for language models. *arXiv preprint arXiv:2502.03358*.
- Fengli Xu, Qianyu Hao, Zefang Zong, Jingwei Wang, Yunke Zhang, Jingyi Wang, Xiaochong Lan, Jiahui Gong, Tianjian Ouyang, Fanjin Meng, and 1 others. 2025a. Towards large reasoning models: A survey of reinforced reasoning with large language models. *arXiv preprint arXiv:2501.09686*.

- Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025b. A-mem: Agentic memory for llm agents. *arXiv preprint arXiv:2502.12110*.
- Zaiyan Xu, Sushil Vemuri, Kishan Panaganti, Dileep Kalathil, Rahul Jain, and Deepak Ramachandran. 2025c. Distributionally robust direct preference optimization. *arXiv e-prints*, pages arXiv–2502.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 technical report. *arXiv e-prints*, pages arXiv–2412.
- Xi Ye, Fangcong Yin, Yinghui He, Joie Zhang, Howard Yen, Tianyu Gao, Greg Durrett, and Danqi Chen. 2025. Longproc: Benchmarking long-context language models on long procedural generation. *arXiv preprint arXiv:2501.05414*.
- Hongli Yu, Tinghong Chen, Jiangtao Feng, Jiangjie Chen, Weinan Dai, Qiying Yu, Ya-Qin Zhang, Wei-Ying Ma, Jingjing Liu, Mingxuan Wang, and Hao Zhou. 2026. [Memagent: Reshaping long-context LLM with multi-conv RL-based memory agent](#). In *The Fourteenth International Conference on Learning Representations*.
- Rui Yu, Shenghua Wan, Yucen Wang, Chen-Xiao Gao, Le Gan, Zongzhang Zhang, and De-Chuan Zhan. 2025. Reward models in deep reinforcement learning: A survey. *arXiv preprint arXiv:2506.15421*.
- Lifan Yuan, Wendi Li, Huayu Chen, Ganqu Cui, Ning Ding, Kaiyan Zhang, Bowen Zhou, Zhiyuan Liu, and Hao Peng. 2024. Free process rewards without process labels. *arXiv preprint arXiv:2412.01981*.
- Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, and 1 others. 2025. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.
- Kai Zhang, Xinyuan Zhang, Ejaz Ahmed, Hongda Jiang, Caleb Kumar, Kai Sun, Zhaojiang Lin, Sanat Sharma, Shereen Oraby, Aaron Colak, Ahmed Aly, Anuj Kumar, Xiaozhong Liu, and Xin Luna Dong. 2025. [As-somem: Scalable memory qa with multi-signal associative retrieval](#). *Preprint*, arXiv:2510.10397.
- Siyan Zhao, Mingyi Hong, Yang Liu, Devamanyu Hazarika, and Kaixiang Lin. 2025. [Do llms recognize your preferences? evaluating personalized preference following in llms](#). *Preprint*, arXiv:2502.09597.
- Congming Zheng, Jiachen Zhu, Zhuoying Ou, Yuxiang Chen, Kangning Zhang, Rong Shan, Zeyu Zheng, Mengyue Yang, Jianghao Lin, Yong Yu, and 1 others. 2025. A survey of process reward models: From outcome signals to process supervisions for large language models. *arXiv preprint arXiv:2510.08049*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Jialun Zhong, Wei Shen, Yanzeng Li, Songyang Gao, Hua Lu, Yicheng Chen, Yang Zhang, Wei Zhou, Jinjie Gu, and Lei Zou. 2025. A comprehensive survey of reward models: Taxonomy, applications, challenges, and future. *arXiv preprint arXiv:2504.12328*.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. [Memorybank: Enhancing large language models with long-term memory](#). *CoRR*, abs/2305.10250.
- Keyan Zhou, Zecheng Tang, Lingfeng Ming, Guanghao Zhou, Qiguang Chen, Dan Qiao, Zheming Yang, Libo Qin, Minghui Qiu, Juntao Li, and 1 others. 2025. Mmlongcite: A benchmark for evaluating fidelity of long-context vision-language models. *arXiv preprint arXiv:2510.13276*.

A Comparison between MemoryRewardBench and Existing Memory Benchmarks

Table 1 provides an overall comparison between MemoryRewardBench and representative existing memory benchmarks in terms of evaluation targets, evaluation paradigms, task settings, and the dimensions of memory ability they cover. Below, we will clarify key terminology and provide a more detailed explanation of the comparison criteria.

Evaluation Target refers to the primary object of assessment. Most existing memory benchmarks take LLMs as the direct evaluation target and measure memory capability based on the correctness or consistency of model outputs. In contrast, MemoryRewardBench focuses on RMs, evaluating how effectively RMs judge the quality of long-term intermediate memories in LLMs.

Process Evaluation distinguishes whether a benchmark explicitly evaluates intermediate states during task execution. Benchmarks marked with “✗” focus solely on outcome’s correctness and do not assess intermediate reasoning or memory update processes, whereas those marked with “✓” evaluate intermediate steps. MemoryRewardBench emphasizes the assessment of long-term intermediate memory trajectories and therefore supports process-level evaluation.

Static vs. Dynamic characterizes whether task-relevant information changes over time. Static settings assume a fixed context or memory state, requiring the model to reason over unchanging information. In contrast, Dynamic settings involve continual information updates during interaction or generation, where the model must maintain, revise, or overwrite existing memory. MemoryRewardBench incorporates both static and dynamic scenarios: Multi-turn Dialogue Understanding is inherently dynamic due to the continuous integration of temporal signals and new interactions, whereas Long-context Reasoning and Long-form Generation are treated as static. This design better reflects realistic memory-management requirements in practical deployments.

Context Length specifies the range of context lengths involved in each benchmark, reflecting the extent to which long-term memory capability is evaluated. Compared with benchmarks constrained to shorter context windows, MemoryRewardBench

covers a broader range of long contexts, enabling a more systematic analysis of RMs’ long-horizon memory evaluation capabilities.

In the **Memory Abilities** column, we map the task types covered by different benchmarks to five core memory capabilities:

- **DU (Dialogue Understanding)**, which evaluates a model’s ability to comprehend multi-turn dialogue histories and maintain consistency across interactions;
- **MR (Multi-hop Reasoning)**, which assesses the model’s capacity to perform reasoning and information integration across multiple inference steps;
- **KU (Knowledge Update)**, which focuses on the model’s ability to update existing knowledge or internal memory upon the introduction of new information;
- **TR (Temporal Reasoning)**, which examines the model’s ability to model and reason about temporal order, event sequencing, and evolutionary processes;
- **GEN (Generation)**, which evaluates the model’s ability to maintain content coherence and satisfy multiple constraints in long-form or multi-stage generation tasks.

Existing memory benchmarks typically cover only a subset of the aforementioned memory abilities, whereas MemoryRewardBench integrates multiple task formulations to achieve systematic coverage of diverse memory capabilities within a unified evaluation framework. Specifically, DU and GEN are directly assessed through the Multi-turn Dialogue Understanding and Long-form Generation tasks, respectively. In the Long-context Reasoning task, the incorporation of multi-hop reasoning data enables the evaluation of both MR and TR, while the use of the MemAgent (Yu et al., 2026) paradigm in Long-context Reasoning further equips the task with the ability to assess KU. Through these designs, MemoryRewardBench provides a more comprehensive and fine-grained characterization of RMs’ memory evaluation capabilities compared to other existing memory benchmarks.

B Benchmark Construction

Overview of MemoryRewardBench construction statistics is shown in Table 4. We provide the construction details below.

Task	LLM Memory Management	Type	Preference Construction	Dataset	Construction Description
Long-context Reasoning	MemAgent (Yu et al., 2026)	Sequential	Self-Correct	BABILong LongMIT	Process context sequentially, Self-Correct denotes real-time correction of memory. Drop-Info denotes masking key information.
			Drop-Info	LongMIT	
		Parallel	Self-Correct	BABILong LongMIT	Process context in parallel, then aggregate memories of each line.
			Drop-Info	LongMIT	
Multi-turn Dialogue Understanding	Mem0 (Chhikara et al., 2025)	Sequential	OUT	LoCoMo MemoryAgentBench	Classify the samples based on whether the answer of the rejected is correct or not. OUT is 0 as easy. MEM is classified as hard.
	A-Mem (Xu et al., 2025b)		MEM	LoCoMo MemoryAgentBench	
Long-form Generation	-	Sequential	Direct-Generate	LongProc	Plan routes based on the given mode of transportation.
			Prompt-Modify	LongEval LongGenBench	Generate a complete article based on the given outline. Generate long text based on given constraints.
		Parallel	Prompt-Modify	LongEval	Generate a complete article based on the given outline.
				LongGenBench	Generate long text based on given constraints.

Table 4: Overview of construction details of MemoryRewardBench. Due to space limitation, we provide partial citations here: BABILong (Kuratov et al., 2024), LongMIT (Chen et al., 2025), LoCoMo (Maharana et al., 2024), MemoryAgentBench (Hu et al., 2025a), LongProc (Ye et al., 2025), LongEval (Wu et al., 2025), and LongGenBench (Wu et al., 2024c).

B.1 Long-context Reasoning

Prototype Description We first define the key information as the content to which the model must attend in order to correctly answer the question, and define the contexts that include key information as key contexts. As described in § 3.3, we introduce two error-inducing perturbations: NOISE and DROP. We apply these two methods by chunking a long context into equal segments, iteratively updating its memory after each segment to condense key information into a fixed-size buffer for response generation. We construct “Sequential” pattern data based on the MemAgent’s sequential processing mechanism, and the “Mixed” pattern is built upon the “Sequential” pattern.

Dataset Description We build up our Long-Context Reasoning task on Babilong (Kuratov et al., 2024) and LongMiT (Chen et al., 2025), for the convenience of extracting clue documents and background documents. For LongMiT, we take the clue documents as the key contexts. For babilong, we take the needles as the key contexts.

NOISE Perturbation We define noise as any *redundant information* introduced by the MemAgent during the memory updating process. In this work, we consider two types of redundancy: (1) incorrect memory updates and (2) repeated memory updates. To induce incorrect memory updates, we employ weaker models (Llama-3.1-8B-Instruct/Qwen2.5-7B-Instruct) as the MemAgent engine. After each memory update, we

adopt an LLM-as-judge paradigm, using a stronger model (Qwen3-235A22B) to assess the quality of the updated memory. If the update is judged correct, it is passed to the next chunk; otherwise, the engine model is instructed to revise the memory according to the feedback. This correction process is repeated whenever erroneous updates persist. Samples that either never fail or fail to be corrected after 10 attempts are discarded. Upon completion of the above process, we cache the full inference trajectory, extract the complete trajectory as the chosen sample, and remove the incorrect updates from it to construct the corresponding rejected sample. Then, to induce redundancy caused by repetition, we reuse the previously discarded never-failed samples as chosen samples. For each such trajectory, we randomly insert an additional memory segment extracted from the same trajectory, thereby introducing redundant repetition to form the rejected sample. For more details, see the case shown in Figure 16.

DROP Perturbation Another approach to inducing errors is dropping key information from the context. Specifically, we remove all key information while retaining the full background context, ensuring that the resulting memories lack crucial evidence. This setup simulates scenarios in which the agent ignores essential clues; the corresponding trajectory is used to construct the rejected sample. In contrast, we apply the standard MemAgent process to the complete context to obtain the chosen

sample. Throughout this procedure, we ensure that key information is always contained within a single, complete chunk rather than split across adjacent chunks, guaranteeing that its removal leads to an incorrect response. We further filter out samples whose chosen trajectories fail to produce correct answers, thereby avoiding cases where both chosen and rejected samples omit critical information and exhibit insufficient preference separation. For more details, see the case shown in Figure 13.

Implementation Details of Mixed Pattern Following the same procedure as before, we first divide the long context into segments of equal length. As illustrated on the right group of Figure 2, each segment is initially processed independently under a parallel pattern. For the parallel part of the mixed pattern, let p be the number of chunks divided, referring to the parallel size, we set $p \in \{2, 3\}$. We then introduce an aggregation mechanism to integrate the memories from all segments. After aggregation, the agent continues to update its memory sequentially to generate the final answer, adhering to the sequential pattern throughout this stage. Since this process combines both parallel and sequential memory management patterns, we categorize the resulting preference pairs as Mixed-Noise or Mixed-Drop, respectively. For more details, see the case shown in Figure 14, 15.

B.2 Multi-turn Dialogue Understanding

As described in § 5.4, we evaluate RM performance on the multi-turn dialogue understanding task with and without auxiliary signals. We focus on two representative memory systems, A-Mem (Xu et al., 2025b) and Mem0 (Chhikara et al., 2025), both of which are designed for long-dialogue scenarios and operate by dynamically storing and iteratively updating memories across conversation turns. However, the two systems differ substantially in how memory is organized and maintained.

A-Mem assigns semantic tags to different segments of a conversation at each round to summarize their content. When a new conversation round begins, the system retrieves and updates the top- k most relevant memories based on the current dialogue context. In contrast, Mem0 does not employ a tagging mechanism. Instead, it maintains a global memory summary, into which new information is directly incorporated at each update step (As illustrated by the examples in Fig. 9).

Pair-data Construction The conversational data used in the experiment comes from the Lo-CoMo (Maharana et al., 2024) dataset and the Conflict_Resolution dataset from MemoryAgent-Bench (Hu et al., 2025a), which focuses on relationship information and key statements within the conversations.

For sample generation, we produce both positive and negative samples from the conversation data. Positive samples are created by processing all rounds of the conversation and ensuring that the final memory chain leads to a correct answer. Only memory chains that result in accurate answers are retained as positive samples.

Negative samples are generated by manipulating the frequency of memory update triggers, intentionally leaving some updates incomplete (e.g., skipping updates in certain rounds). This results in memory chains that are less complete and less well-organized compared to the corresponding positive samples. These incomplete memory chains are further classified into two categories:

(Mem) Negative Samples: If the memory chain retains the key information required to answer the question but has flaws in memory management (such as reduced efficiency in information retrieval), and the system still outputs the correct answer, the memory chain is marked as Mem, indicating "correct result but with memory management defects."

(Out) Negative Samples: If the memory chain is so incomplete that key information is lost or difficult to retrieve, leading to an incorrect answer, it is marked as Out, indicating "incorrect result due to failure in key information retrieval."

B.3 Long-form Generation

Prototype Description As described in Section 3.3, Long-form Generation is modeled as a constraint-driven task, where the model iteratively generates content based on a series of progressive constraints. The preference settings for long text generation tasks are based on the path-traversal subset of LongProc (Ye et al., 2025), as well as the LongGenBench (Wu et al., 2024c) and LongEval (Wu et al., 2025) datasets.

(1) Longproc: The pathtraversal subtask of LongProc provides the model with all known routes along with specified start and end points, and requires it to generate the correct sequence of steps leading from the source to the destination. This task emphasizes procedural reasoning, as each interme-

diated state is explicitly dependent on the preceding generation step, rendering the process inherently sequential. Accordingly, parallel generation is not applicable to LongProc, and only the sequential generation scheme is employed for this dataset.

(2) LongGenBench and LongEval: In contrast, LongGenBench takes a set of clearly specified constraints as input, including detailed content requirements and overall structural conditions. Similarly, LongEval uses instructions containing paragraph-level constraints to define the generation requirements for different parts of the article. For both datasets, the model needs to generate a long text output that satisfies all the given constraints. Due to the decomposable input and segmented output characteristics of these two datasets, we applied both sequential and parallel generation schemes.

Benchmark Construction Based on LongProc

For the path traversal subset of LongProc datasets, only employ a sequential generation scheme. A representative example is presented in Figure 10.

(1) Chosen-sample Construction: The provided reference answers are utilized as the chosen samples, representing the gold-standard outputs that fully satisfy all specified constraints.

(2) Rejected-sample Construction: Rejected samples are those where the model generates results that satisfy the initial and final constraints but contain errors in the intermediate steps. The output length of these samples is controlled to be consistent with the length of the selected samples.

Benchmark Construction Based on LongGenBench and LongEval The construction of the benchmark based on LongGenBench and LongEval involves two principal components: the generation schemes used to produce long-form outputs and the construction of chosen and rejected samples.

(1) Sequential and Parallel generation schemes: In the sequential mode, each input instruction is decomposed into an ordered sequence of step-wise constraints, and the model incrementally generates intermediate outputs conditioned on the current constraint and accumulated memory. As shown in Figure 11 and Figure 12, in the sequential generation scheme, both chosen and rejected samples indicate that each paragraph should be generated conditioned on the previously generated text. After processing all constraints, the memory states are concatenated to produce the final response.

In the parallel generation setting, the original long instruction is first decomposed into multiple

sub-instruction segments. Each segment is then processed in parallel and independently, yielding its corresponding generated output and memory state. Finally, the memory states and outputs produced by all sub-generation processes are aggregated to form a complete generation trajectory.

(2) Chosen and Rejected Sample Construction: To select Chosen samples, the long-form outputs are evaluated in a block-wise manner using a sufficiently capable model, with each segment checked against all input constraints. Only outputs that fully satisfy every specified constraint are designated as correct samples. As illustrated in Figure 11 and Figure 12, the bolded segments of the chosen samples align with the bolded elements in the question. For the rejected sample, LongEval induces memory errors by perturbing step-wise constraints, whereas LongGenBench simulates memory loss by dropping a subset of constraints. In Figure 11, the bolded content of the rejected sample fails to match the bolded elements in the question, and in Figure 12, the bolded portions of the question are omitted from the rejected sample. In both cases, an auxiliary model is used to modify the original constraints, leading to hallucinated or missing intermediate memories.

C Evaluation Settings

C.1 Prompts for each Task

As shown in Figure 17, we provide the prompt used for evaluation. We use different system prompts to adapt the task format. For Long-context Reasoning and Multi-turn Dialogue Understanding, we use the “System Prompt of Understanding Tasks”. For Long-form Generation, we use the “System Prompt of Generation Tasks”. When constructing the evaluation samples, we first implement the user template: The chosen and rejected trajectories in the evaluation preference pair will be randomly shuffled and put into {Response A} and {Response B}, respectively. Then, we concatenate the system prompt and the implemented user template as the final prompt for every sample.

C.2 Evaluation Framework

For proprietary models, we conduct evaluations via the official APIs. For open-source models, all evaluations are performed within the LOOM-Scope (Tang et al., 2025b) framework. Across all models, we apply the same sampling hyperparameters: we set the top- p value to 0.95, the temperature

to 0.7, and the maximum number of generation tokens to 16,384, ensuring that model outputs are not prematurely truncated.

D Additional Results

D.1 Evaluation on trained specialized RMs

As shown in Table 5, Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2026) and Selene-1-Mini-Llama-3.1-8B (Alexandru et al., 2025) attain overall average scores of 64.79 and 46.83, respectively. Skywork-Reward-V2 is particularly strong on Long-form Generation, reaching an average score of 77.38, while both models achieve consistently high scores on the DROP settings in Long-context Reasoning. These results suggest that well-trained open reward models can serve as competitive evaluators for memory-management quality. Their strong performance is likely attributable to their reward-model training objective, which explicitly teaches them to distinguish preferred from rejected responses using preference data. This objective is closely aligned with the pairwise evaluation protocol in MemoryRewardBench, enabling the models to identify salient quality differences caused by memory errors, especially omitted key information and downstream response degradation. Nevertheless, their comparatively weaker performance on the NOISE settings and the MM setting indicates that subtle or redundant memory perturbations remain more challenging, leaving room for further improvement in trajectory-level memory evaluation.

D.2 Human Verification of Gold Labels

To assess whether the process-based gold labels align with human preferences, we conduct a blinded human evaluation on a randomly sampled subset of MemoryRewardBench. We sample 30 preference pairs from each task—Long-context Reasoning, Multi-turn Dialogue Understanding, and Long-form Generation—resulting in 90 pairs in total. Each pair is independently evaluated by three annotators. The task identities are anonymized during annotation, the order of the two trajectories is randomized, and the original process-based gold label is hidden from annotators.

For each pair, annotators select the trajectory they prefer overall, considering both final-answer correctness and the quality of intermediate memory management, including faithfulness, conciseness, and logical coherence. We define the agreement

rate as the proportion of human judgments that match the corresponding process-based gold label. To further examine the validity of the synthetic perturbations, annotators are also asked to identify whether the rejected trajectory contains an injected memory error.

As shown in Figure 18, the agreement rates between individual annotators and the process-based gold labels are consistently high across all three tasks. These results support the reliability of the process-based gold labels and the fidelity of the injected memory perturbations.

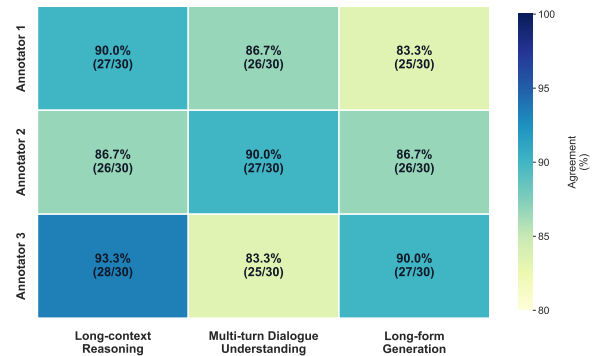


Figure 18: Agreement between human annotators and process-based gold labels.

D.3 Parsing Robustness and Format Compliance

Format compliance is an integral part of model evaluation in MemoryRewardBench: a model must produce an output that can be parsed into the required preference decision. To distinguish format-related failures from failures on parseable outputs, we separately report per-model parsing success and parsing failure rates in Table 6. A parsing failure is defined as an output from which the evaluation pipeline cannot extract a valid prediction.

Overall, 8 out of 12 evaluated models achieve parsing success rates above 90%, indicating that format-related failures are limited for most models. Claude-Opus-4.5 achieves perfect format compliance, while Qwen3-32B, Qwen3-8B, and GLM4.5-106A12B also achieve success rates above 97%. The remaining models exhibit non-negligible parsing failure rates, ranging from 14.08% to 29.46%. Reporting these rates separately makes the contribution of format violations transparent when interpreting the overall MemoryRewardBench scores and cross-model comparisons. We retain parsing failures in the primary evaluation, since reliably

Models	Long-context Reasoning					Multi-turn Dialogue Understanding					Long-form Generation			Avg.
	S-NOISE	S-DROP	M-NOISE	M-DROP	Avg.	MO	MM	AO	AM	Avg.	S	P	Avg.	
Skywork-Reward-V2-Llama-3.1-8B	53.33	85.82	46.74	81.06	65.13	25.75	16.98	75.11	70.61	51.88	71.55	85.13	77.38	64.79
Selene-1-Mini-Llama-3.1-8B	46.28	54.48	41.30	69.16	53.00	37.73	38.99	36.25	37.96	37.63	51.86	47.23	48.88	46.83

Table 5: Supplementary results on MemoryRewardBench.

Model	Parsing Success (%) $\uparrow$
Claude-Opus-4.5	100.00
Gemini3-Pro	92.54
Qwen3-Max	93.13
Qwen3-235A22B	91.50
GLM4.5-106A12B	97.13
Qwen2.5-72B	91.67
Llama3.3-70B	84.50
Qwen3-32B	98.41
Qwen3-8B	97.79
Llama3.1-8B-Ins	81.08
Qwen2.5-7B-Ins	70.54
Qwen3-4B	85.92

Table 6: Per-model parsing success rates in the main experiments.

Models	Sequential			Parallel		
	LR	LG	Avg.	LR	LG	Avg.
GLM4.5-106A12B	54.9	79.7	70.8	49.5	77.6	67.7
Llama-3.3-70B-Instruct	40.8	60.6	53.5	33.7	64.1	53.4
Llama-3.1-8B-Instruct	37.6	46.2	43.2	42.9	43.1	43.0
Qwen3-4B	53.3	56.2	55.2	46.2	57.1	53.3

Table 7: Performance comparison at seed 42 between Sequential and Parallel pattern. "LR", "LG" refer to **Long-context Reasoning** and **Long-form Generation** respectively.

following the required output format is itself part of the evaluation capability.

E Details of Ablation Study

E.1 Memory Management Patterns

We report a more thorough version of Figure 3 in Table 7, which shows that the Long-context Reasoning (LR) task accuracy of Llama-3.1-8B-Instruct outperforms the Long-form Generation (LG) task, while the LG task accuracy of Llama-3.3-70B-Instruct outperforms the LR task. Given the similarity in patterns between these two tasks, we compute a weighted average of their results to remove task-specific bias, and ultimately conclude that the sequential pattern is easier than the parallel pattern.

E.2 Failure Case Analysis of Large-scale LLMs

As shown in Fig 19, the superior performance of Qwen3-14B over both Qwen2.5-72B-Instruct and

Models	8k	16k	32k	64k	128k	Avg.
🔒 Closed-source Models						
Claude-Opus-4.5	70.9	79.7	78.3	73.4	68.8	74.8
Gemini-3-Pro	71.6	69.0	75.4	75.3	63.3	71.6
Qwen3-Max	68.5	72.2	68.9	64.7	63.9	67.8
🔓 Open-source Models						
GLM4.5-106A12B	67.8	63.6	76.7	66.8	63.9	68.3
Qwen3-235A22B	64.9	71.8	<u>70.1</u>	<u>64.8</u>	58.5	<u>66.8</u>
Qwen3-32B	64.0	62.0	66.3	62.0	<u>58.7</u>	62.8
Qwen3-14B	63.0	60.4	60.8	60.4	55.9	60.6
Qwen3-8B	59.0	58.4	59.7	55.1	53.6	57.2
Llama3.3-70B	<u>65.4</u>	58.1	58.0	44.6	35.8	57.8
Qwen2.5-72B	55.1	58.1	53.2	46.2	45.6	51.9
Llama3.1-8B	45.1	48.1	49.8	44.2	26.4	43.9
Qwen2.5-7B	28.9	28.8	44.2	42.8	45.6	38.0

Table 8: Results on MemoryRewardBench-L (Length Perspective).

Llama-3.3-70B-Instruct can be attributed to its significant improvement in reasoning capabilities following post-training. Qwen3-14B excels not only in accurately identifying constraint violations but also in maintaining strict adherence to specified requirements. As demonstrated in the comparison between Agent A and Agent B, Qwen3-14B strikes a careful balance between descriptive quality and instruction adherence, ensuring that each output is both detailed and fully compliant with the established constraints. In contrast, while Llama-3.3-70B-Instruct prioritizes descriptive richness, it struggles with consistently following instructions, leading to inaccurate or poorly structured outputs, as seen in its failure to correctly allocate coffee shops on the required floors. Furthermore, Qwen2.5-72B-Instruct lacks the content precision and accuracy of Qwen3-14B, as evidenced by its less detailed descriptions and the omission of crucial design features.

E.3 The impact of trajectory length on RM

The impact of trajectory length manifests in two primary aspects: the accuracy of reward evaluation and the consistency thereof.

Regarding accuracy, we present in Table 8 the statistical evaluation results categorized by trajectory length, indicating that model performance ex-

hibits a downward trend as trajectory length increases. Regarding consistency, as indicated in Table 11, we initially reversed the contextual order of the chosen and rejected responses during evaluation and tested several models, revealing inconsistencies in their outputs. To investigate this phenomenon, we further conducted consistency experiments across all tasks. As demonstrated in Section 5.3, we provide a more thorough version of length-induced bias shown in Figure 20. We also report the consistency score of some open-sourced models. As shown in Table 9, models with enhanced reward evaluation capability also tend to exhibit stronger consistency.

E.4 Global Constraint of Long-form Generation

As shown in Figure 21, we present overall performance trends of LongEval, LongGenBench, and LongProc on long-form generation tasks as the constraint density varies. A unified constraint density control strategy is adopted across all three datasets to ensure comparability across task settings. For the **LongProc** dataset, we vary the availability of route constraints by progressively partitioning the complete set of known routes. A constraint rate of 0% corresponds to the setting where no known routes are provided to the model, in which case the RMs infer the path solely from the given start and end points. In contrast, a constraint rate of 100% supplies the model with the full set of known routes, enabling the reward model to be fully informed of all possible path transitions.

For **LongEval** and **LongGenBench**, constraint density is controlled by selecting subsets of the complete constraint set while preserving the semantic integrity of each individual constraint. As shown in Figure 22, under the 0% constraint setting, the model is provided only with a minimal prompt to generate long-form text, without any explicit content or structural requirements. At the 100% constraint level, all predefined constraints are supplied. Intermediate constraint ratios are obtained by proportionally sampling from the full constraint set, enabling a systematic analysis of RMs behavior under varying degrees of constraint supervision.

E.5 Multi-turn Understanding with Auxiliary Signals

We investigate the role of auxiliary signals, such as tags, in the A-Mem memory system (Xu et al., 2025b). As illustrated in Figure 23, we present two

scenarios: one where auxiliary signals are used, and another where they are absent. In the first scenario, A-Mem utilizes semantic tags to summarize dialogue content; in the second, these tags are removed, causing them to lose part of their semantic function. This comparison highlights the impact of auxiliary signals on the system’s performance.

The experimental results show that when the data contains meaningful semantic tags, the model is able to more effectively distinguish between different memory processes and dialogue segments, leading to significant improvements in performance on the memsys task. As shown in Figure 23, tags in the A-Mem system not only serve as metadata for structured memory but also play a critical role in semantic indexing and organization. They provide additional context for the model, helping it locate key information and understand the structure of long conversations, thereby optimizing memory management.

In contrast, removing the tags significantly impairs the model’s ability to distinguish memory processes, resulting in a decline in performance. This validates the importance of designing effective memory organization mechanisms (such as tagging) when building memory systems to enhance a model’s ability to comprehend long conversations.

E.6 Sources of Task Difficulty

To investigate the relationship between the difficulty levels of MemoryRewardBench and those of the original tasks, we evaluate Llama-3.1-8B-Instruct and Qwen3-8B on the source benchmarks corresponding to the three MemoryRewardBench subtasks. Since the source benchmarks lack the fine-grained sub-classifications present in MemoryRewardBench, we aggregate the experimental results solely based on its three primary categories. Table 12 demonstrates a descending difficulty profile for LR, MD, and LG on their original benchmarks. Interestingly, this trend does not persist within MemoryRewardBench, implying that our benchmark introduces difficulty factors distinct from the source tasks. For instance, LR performs surprisingly well despite its greater inherent difficulty, whereas LG remains challenging despite being the easiest among the source tasks. This discrepancy arises because MemoryRewardBench is designed as a discriminative task rather than a generative one; models simply need to distinguish the better trajectory from a pair of candidates rather than solve the task. Meanwhile, the variation in

Models	Long-context Reasoning					Multi-turn Dialogue Understanding					Long-form Generation			Avg.
	S-Noise	S-Drop	M-Noise	M-Drop	avg.	MO	MM	AO	AM	avg.	S	P	avg.	
GLM4.5-106A12B	34.12	82.84	26.09	86.34	55.25	69.46	71.7	55.1	65.94	64.5	71.99	65.89	69.38	63.04
Qwen2.5-72B	21.57	<u>74.63</u>	22.28	<u>74.45</u>	45.63	46.71	50.31	25.71	<u>37.99</u>	38.50	44.20	39.94	42.38	42.17
Llama-3.3-70B	24.71	50.00	9.78	60.79	35.75	53.29	<u>59.12</u>	26.94	31.88	<u>40.25</u>	47.05	34.40	41.63	39.21
Qwen3-8B	40.78	47.76	41.85	61.23	48.00	<u>55.09</u>	52.83	27.35	27.51	38.25	<u>59.74</u>	51.02	56.00	47.42
Llama-3.1-8B	27.06	44.03	19.02	50.22	34.63	38.92	33.33	24.08	22.27	28.50	27.13	20.70	24.38	29.17
Qwen2.5-7B	10.20	15.67	20.11	22.91	17.00	31.74	34.59	13.88	14.41	21.88	32.82	31.20	32.13	23.67
Qwen3-4B	35.69	58.96	28.80	62.56	45.63	39.52	44.03	<u>37.55</u>	36.25	38.88	44.86	41.98	43.63	42.71

Table 9: Results on MemoryRewardBench-Consistency.

Models	subtasks		avg
	With	WithOut	
GLM4.5-106A12B	0.759	0.620	0.690
Qwen3-14B	0.690	0.540	0.603
Qwen2.5-14B-Instruct	0.621	0.500	0.567
Qwen3-4B	0.655	0.510	0.585

Table 10: Impact of structured tags on LLM performance in memsys subtasks (WithTags: memsys_v2 with structured tags; WithOutTags: memsys without tags)

Models	Chosen-First			Rejected-First		
	LR	MD	Avg.	LR	MD	Avg.
GLM4.5-106A12B	84.1	52.4	68.3	53.9	63.4	58.6
Llama-3.3-70B-Ins	65.6	58.6	62.1	39.6	27.5	33.6
Llama-3.1-8B-Ins	56.8	38.5	47.6	40.1	31.4	35.8
Qwen3-4B	61.5	37.2	49.4	57.9	52.1	55.0

Table 11: Performance comparison at seed 42 on MultiHop-QA task. “LR” and “MD” refer to Long-context Reasoning and Multi-turn Dialogue Understanding, respectively.

memory construction methods further contribute to the observed difficulty levels.

F Discussion

F.1 The Necessity of Reward Models

The extent to which a Reward Model (RM) facilitates the development of advanced memory modules is primarily governed by two factors: the underlying RLHF algorithm employed and the accuracy and density of the reward signals generated by the RM. Our work specifically targets the latter. By providing a standardized evaluation framework that captures representative memory-management patterns under controlled settings, MemoryRewardBench offers a proxy evaluation environment. Consequently, an RM’s performance on this benchmark may indicate its potential effectiveness in actual deployment; however, this relationship remains to be validated through downstream experiments.

Models	Source			MemoryRewardBench		
	LR	MD	LG	LR	MD	LG
Llama-3.1-8B-Ins	23.7	42.5	50.1	34.6	28.5	24.3
Qwen3-8B	25.3	45.4	68.5	48.0	38.2	56.0

Table 12: Performance comparison between source benchmarks and MemoryRewardBench.

Furthermore, our analysis reveals a critical disconnect between a model’s general capabilities and its discriminative prowess in memory-intensive contexts. As demonstrated by the varying performance across source tasks—and their altered difficulty profiles upon memory-based construction—an RM’s ability to solve a task does not seamlessly translate to evaluating complex memory trajectories. MemoryRewardBench exposes these vulnerabilities, highlighting the necessity for specialized training paradigms. RMs are indispensable because they can be explicitly trained to enhance sensitivity to memory retention, factual consistency, and reasoning over extended contexts, which general outcome-based models often fail to capture.

F.2 Limitations of Rule-Based Evaluation

While rule-based heuristics are computationally efficient, they are fundamentally inadequate for evaluating complex memory trajectories. Memory operations in multi-turn dialogues involve abstract summarization and semantic evolution, which rigid lexical matching or predefined templates cannot effectively capture. For example, rule-based systems struggle to identify nuanced contextual errors, such as mistakenly discarding a historically vital detail. In contrast, RMs possess the necessary semantic depth for accurate, process-level evaluations.

We also acknowledge certain limitations in the current iteration of our benchmark. First, our memory construction relies on a bounded set of templates, making it challenging to fully capture the unpredictable, entangled, and noisy nature of real-

world user interactions within an offline, static dataset. Second, an offline evaluation framework cannot completely account for dynamic challenges encountered during online reinforcement learning, such as reward hacking or out-of-distribution shifts.

Looking ahead, a natural next step is integrating MemoryRewardBench into an active RLHF pipeline to quantify the correlation between offline discriminative metrics and online policy improvements. Additionally, expanding the benchmark to include diverse memory architectures—such as retrieval-augmented generation (RAG) or graph-based systems—will further stress-test contemporary RMs. Ultimately, we hope MemoryRewardBench serves as a stepping stone toward more reliable, memory-capable AI agents.

1. Input & Ground Truth

Question: Which bands has Dave enjoyed listening to?

Reference Context :

[Conversation 3] Dave says: "Wow, there were so many great bands! If I had to pick a favorite, it would definitely be Aerosmith. Their performance was incredible. I'll show you a pic I took when they were playing."

[Conversation 12] Dave says: "The Fireworks headlined the festival."

Answer (Ground Truth): Dave has enjoyed listening to Aerosmith and The Fireworks.

2. Memory Process

Chosen

[Conversation 3]

Context: The speaker enthusiastically praises the band Aerosmith, specifically calling their performance "incredible" and mentioning a photo taken during the set.

Keywords: ['Aerosmith', 'band', 'performance', 'incredible', 'pic']

Tags: ['music', 'concerts', 'personal favorites', 'rock']

[Conversation 12]

Context: The speaker mentions that a specific band named "The Fireworks" was the headliner for the festival.

Keywords: ['The Fireworks', 'headlined', 'festival', 'band']

Tags: ['music', 'live events', 'festival lineup']

Rejected

[Conversation 3]

Context: The speaker discusses attending a music event with many bands and shares a picture they took of the stage performance. (Generic interpretation)

Keywords: ['bands', 'pic', 'playing', 'great']

Tags: ['Entertainment', 'Photography', 'Events']

[Conversation 12]

Context: The speaker describes a festival event where a fireworks display was the main attraction to entertain the crowd. (Entity linking error: Pyrotechnics vs Band)

Keywords: ['fireworks', 'festival', 'entertainment']

Tags: ['Public Events', 'Celebrations', 'Pyrotechnics']

3. Output

Chosen

From the context provided, it's clear that Dave has particularly enjoyed Aerosmith, describing their performance as incredible. Additionally, Dave mentioned The Fireworks as the headlining band of a festival, suggesting that he also has an appreciation for their music.

Rejected

From the context provided, it is clear that Dave has enjoyed listening to Aerosmith, as he mentioned their performance was incredible. Additionally, Dave noted that The Fireworks headlined a festival.

Figure 8: A case showing critical differences in their memory quality. Although both chosen and rejected samples yield correct final answers, closer inspection reveals critical differences in their memory quality. (1)Chosen keywords: ['Aerosmith', 'band', 'performance', 'incredible', 'pic'] The chosen trajectory successfully retains "Aerosmith", whereas the rejected sample omits this key information. (2)Rejected trajectory: The model erroneously interprets the band name as a festival, generating: "The speaker describes a festival event where a fireworks display was the main attraction to entertain the crowd." This represents a critical factual distortion in memory. Despite producing a superficially correct answer, the underlying memory state contains misinformation that undermines reliability.

A_Mem and Mem0 Case

[A_Mem system]

Question:

What is James' favorite game?

A_Mem_Chosen:

[Excerpt from a larger conversation memory pool]

[conversation_72]

Context: The main topic is about playing the game Apex Legends with a team, emphasizing the intensity and sharing a screenshot of the gameplay.

Keywords: ['Apex Legends', 'game', 'team', 'screenshot', 'intense']

Tags: ['gaming', 'social media', 'entertainment', 'team gameplay']

A_Mem_Rejected:

[Excerpt from a larger conversation memory pool]

[conversation_72]

Context: The main topic is about a player sharing their experience and excitement playing the multiplayer video game Apex Legends with their team, and inviting others to view a screenshot of their gameplay.

Keywords: ['game', 'team', 'screenshot', 'playing']

Tags: ['Video Games', 'Gaming', 'Team Play']

Answer:

James's favorite game is Apex Legends.

[Mem0 system]

Question:

What is James interested in?

Mem0_Chosen:

[Representative memory excerpt]

'id': '72eae632-3310-4ec0-93c5-97d535408d2f',

'memory': 'Jame enjoys reading books.',

'hash': 'a65072fab28c8d0ef7aff048a3af4ae1',

'metadata': None,

'created_at': '2025-12-01T20:37:46.562661-08:00'

Mem0_Rejected:

[Representative memory excerpt]

'id': '367e4e8b-8cd9-4266-a2e4-e33dc8c09b18',

'memory': "[5:35 pm on 22 March, 2023] Speaker Jame says: 'I've recently developed a strong interest in reading. I enjoy immersing myself in books, especially how reading allows me to explore different ideas, stories, and perspectives in a focused and reflective way.'",

'hash': '8544b92aab390170f4f19d309e6d192a',

'metadata': None,

'created_at': '2025-12-04T07:04:52.138025-08:00'

Answer:

Reading books.

Figure 9: Illustration of chosen and rejected memory excerpts in a_mem and mem0 systems under different question types.

LongProc Case

[sequential scheme]

Question:

[TASK]

In a completely hypothetical world, there are a number of cities. Each city has a one-way connection to only one other city via a specific transit method (bus, train, plane, or ferry). Your task is to provide a route from a city to another city. You should follow the specific instruction provided later and output the route following the format provided in the instruction. . .

[PROBLEM]

In a hypothetical world, there are a number of cities. Each city has a one-way connection to only one other city via a specific transit method. The details of the cities are as follows: . . . Garland is a lively city. You can travel from Garland to Royston by ferry. Ivanovo is a lively city. You can travel from Ivanovo to Frome by plane. Frome is a lively city. You can travel from Frome to Swift Current by bus. Ostrava is a lively city. You can travel from Ostrava to Frome by train. Rennes is a lively city. You can travel from Rennes to Royston by bus. **Royston is a lively city. You can travel from Royston to Redhill by plane. . .**

Now find the route from Warsaw to Burnley based on the information above. Some reminders: - All connections are one-way. You can solve the problem by iteratively finding the next city to travel to until you reach the destination city. - Follow the specific format for the route output. Mark the route with <Route> and </Route> tags. - Describe in detail how you determine the route step by step (at least 550 sentences)."

Chosen:

From Warsaw, take a plane to Lichfield. From Lichfield, take a ferry to Bern. From Bern, take a bus to Porto. From Porto, take a train to Rennes. From Rennes, take a bus to Royston. **From Royston, take a plane to Redhill.** From Redhill, take a plane to Hitchin. . . From Argenteuil, take a train to Kingston. From Kingston, take a plane to Vienna. From Vienna, take a train to Burnley.

Rejected:

From Warsaw, take a plane to Lichfield. From Lichfield, take a ferry to Bern. From Bern, take a bus to Porto. From Porto, take a train to Rennes. From Rennes, take a bus to Royston. **From Royston, take a plane to Frome.** From Frome, take a bus to Swift Current. . . From Argenteuil, take a train to Kingston. From Kingston, take a plane to Vienna. From Vienna, take a train to Burnley.

Figure 10: LongProc Case.

LongEval Case

[Sequential scheme]

Question:

Title: Building Conversational AI Agents By Integrating Reasoning, Speaking & Acting With LLMs

please generate a blog based on the given bullet point: . . . **Employing AI Agents for engaging conversations with customers. . .**

Chosen:

Based on the title Building Conversational AI Agents By Integrating Reasoning, Speaking & Acting With LLMs, **the previously generated text**, and the given bullet points . . . **Employing AI Agents for engaging conversations with customers. . .**, **I have composed the following section. . .** These advanced AI agents are designed not just to perform routine operations but to **engage in meaningful and dynamic interactions with customers. . .**

Rejected:

Based on the title Building Conversational AI Agents By Integrating Reasoning, Speaking & Acting With LLMs, **the previously generated text**, and the given bullet points . . . **Employing AI Agents for automated decision-making without human intervention. . .**, I have composed the following section. . . these agents can understand context, make informed decisions, and **execute actions with minimal human oversight. . .**

[Parallel scheme]

Question:

Title: Microsoft's Magnetic-One: The Dawn of Truly Autonomous AI

please generate a blog based on the given bullet point: . . . **The system is capable of performing various tasks such as booking flights and writing code. . .**

Chosen:

Based on the title Microsoft's Magnetic-One: The Dawn of Truly Autonomous AI and the given bullet points . . . **The system is capable of performing various tasks such as booking flights and writing code. . .**, I have composed the following section . . . from **booking flights and writing code** to potentially even making a mean cup of coffee . . .

Rejected:

Based on the title Microsoft's Magnetic-One: The Dawn of Truly Autonomous AI and the given bullet points . . . **The system is capable of performing various tasks, but it primarily focuses on entertainment and gaming. . .**, I have composed the following section . . . While its primary focus is currently on **entertainment and gaming**, the potential applications are vast . . .

Figure 11: LongEval Case.

LongGenBench Case

[sequential scheme]

Question:

As Chef Roy Hee, a world-renowned chef at a globally renowned restaurant, you are tasked with designing an entire year's menu for 2018. Your menu should be varied and innovative, adhering to the following guidelines: ... 2) **Feature 'Autumn Harvest Fest' from Week 15 to Week 16 with dishes like: Roasted Squash and Root Vegetables.** ...

Chosen:

Based on the events: Autumn Harvest Fest featuring Roasted Squash and Root Vegetables, **the previously generated text**, I have composed the following section. Menu Week 15 (April 9th - April 15th) Welcome to the ****Autumn Harvest Fest****! ... 2. Main Courses: 2.1) ****Roasted Squash and Root Vegetables****: A medley of butternut squash, parsnips, and carrots, roasted to perfection with thyme and rosemary. ...

Rejected:

Based on the events: Autumn Harvest Fest featuring Roasted Squash and Root Vegetables, **the previously generated text**, I have composed the following section. Menu Week 15 (April 9th - April 15th) This week's menu is a tribute to the bold and hearty flavors of the Middle East, where spices, grains, and legumes come together in harmony. ... 2. Main Courses: 2.1) Lamb Kofta with Sumac and Mint: Spiced lamb meatballs with a bright, herbaceous garnish. 2.2) Grilled Salmon with Za'atar and Sumac: A zesty, aromatic fish dish with a tangy, smoky flavor.

[Parallel scheme]

Question:

Noah is a historian with a passion for chronicling his vibrant life through weekly diary entries. Captures: ... 3) **Participate in a language exchange meetup every 5 weeks on weekends, starting from week 13.** ...

Chosen:

Based on the events: Participate in a language exchange meetup, I have composed the following section. Week 13 (March 25th - March 31st): ... **This week also marked the start of my language exchange meetups, which take place every five weeks on weekends.** I attended the first session on the 24th, and it was both exciting and slightly intimidating to engage in conversation with people from all over the world. ...

Rejected:

Based on the events: Participate in a language exchange meetup, I have composed the following section. Week 13 (March 25th - March 31st): The air is finally beginning to feel like spring, and the days are lengthening in a most welcome way. Temperatures have warmed to a pleasant 12°C by the end of the week, coaxing the first hints of green to peek through the remaining patches of snow in the park.

Figure 12: LongGenBench Case.

S-Drop Case

Question:

Here is a long context but only a few clues are useful, while others are interferences, you should answer the question following the context after carefully reading it.

[Context] Bassel El-Gharbawy ... Abdelouahed Idrissi Chorfi Abdelouahed Idrissi Chorfi (born 1 January 1969) is a Moroccan judoka. He competed at the 1996 Olympics as an extra-lightweight, coming seventeenth. Achievements References Category:1969 births Category:Living people Category:Moroccan male judoka Category:Judoka at the 1996 Summer Olympics ... Majemite Omagbaluwaje (born 10 June 1969) is a Nigerian judoka. He competed at three Olympic Games. [Question]Which athlete has a more diverse set of achievements, Bassel El-Gharbawy or Majemite Omagbaluwaje, considering their participation in various games?

Chosen:

Step1:

<problem> Which athlete has a more diverse set of achievements, Bassel El-Gharbawy or Majemite Omagbaluwaje, considering their participation in various games? </problem>

<memory> No previous memory </memory>

<section> Bassel El-Gharbawy Bassel El-Gharbawy (born 10 April 1977) is an Egyptian judoka. ... </section>

<wrong-updated-memories> No wrong updated memory </wrong-updated-memories>

Updated memory: <updated-memory> Bassel El-Gharbawy (born 10 April 1977) is an Egyptian judoka. ... </updated-memory>

Explanation: I have updated the memory by retaining the correct information from the wrong updated memory and the section, ...

Step2: ... Step3: ... Step4: ... Step5: ...

Updated memory: <problem> Which athlete has a more diverse set of achievements, Bassel El-Gharbawy or Majemite Omagbaluwaje, considering their participation in various games? </problem>

<memory> ... I will continue to gather more information to make a more informed decision.

Comparing Maj

Rejected:

Step1:

<problem> Which athlete has a more diverse set of achievements, Bassel El-Gharbawy or Majemite Omagbaluwaje, considering their participation in various games? </problem>

<memory> No previous memory </memory>

<section> Abdelouahed Idrissi Chorfi

Abdelouahed Idrissi Chorfi (born 1 January 1969) is a Moroccan judoka. He competed at the 1996 Olympics as an extra-lightweight, coming seventeenth. ...

</section>

<wrong-updated-memories> No wrong updated memory </wrong-updated-memories>

Updated memory: <updated-memory> ... No information is available about Bassel El-Gharbawy or Majemite Omagbaluwaje, so it's not possible to compare their achievements.

Step2:

<problem> Which athlete has a more diverse set of achievements, Bassel El-Gharbawy or Majemite Omagbaluwaje, considering their participation in various games? </problem>

<memory> <updated-memory> ... </memory> <section> None </section>

<wrong-updated-memories> No wrong updated memory </wrong-updated-memories>

Updated memory: Since we have no information about Bassel El-Gharbawy and Majemite Omagbaluwaje, and the previous wrong answer was based on an incorrect inference, we cannot make a comparison between their achievements.

It is not possible to compare their achievements as there is no information available about Bassel El-Gharbawy and Majemite Omagbaluwaje.

Figure 13: S-Drop Case.

P-Noise Case

Question:

Here is a long context but only a few clues are useful, while others are interferences, you should answer the question following the context after carefully reading it. [Context] Sandra travelled to the bedroom. ... Sandra grabbed the apple there. ... Sandra gave the apple to John. John journeyed to the bathroom. ... John discarded the milk. ... Sandra went back to the bathroom. John put down the apple there. ... Daniel went to the bathroom... [Question]How many objects is John carrying?

Chosen:

The entire agent workflow consists of parallel processing components and a summarization processing component. Each parallel processing component is as follows: [Part 1] Step 1: ... Step 2: ...

<wrong-updated-memories> No wrong updated memory </wrong-updated-memories>

Updated memory: <updated-memory> There is no information about John in the provided section. ... From the given section, we can see that the narrative is focused on the little girl, Mary, Daniel, and old Hans. ... Upon re-examining the section, I noticed that there is no mention of John whatsoever. ... The scene is indeed taking place in a rural setting, with the sound of a creek and old Hans' axe in the background. The background sounds do not provide any information about John's presence or actions. Therefore, my updated memory is that there is still no information about John, ...

[Part 2] Step 1: ... Step 2: ... Updated memory: <updated-memory> There is a house with a courtyard and stables situated on the fourth side of the quadrangle. The stables are not being used much, as the horses are grazing in an adjoining paddock. John is seen grabbing something in the stables, specifically the milk. ... </updated-memory>

[Part 3] Step 1: ... Step 2: ... Updated memory: <updated-memory> John was carrying an apple initially. He then placed it down in the bathroom. There is no information about John carrying any other objects. </updated-memory>

After processing in parallel, the summarization processing component is as follows: There are 3 processed components (Memories): [Part 1] <updated-memory> There is no information about John in the provided section. However, it does not necessarily mean he is not carrying any objects. We would need more context or information about John to determine the number of objects he is carrying. ... [Part 2] <updated-memory> There is a house with a courtyard and stables situated on the fourth side of the quadrangle. The stables are not being used much, as the horses are grazing in an adjoining paddock. ... </updated-memory> [Part 3] <updated-memory> John was carrying an apple initially. He then placed it down in the bathroom. There is no information about John carrying any other objects. </updated-memory> The summarization of them is as follows: There is no information about John in the provided section. ... The text also mentions that Sandra grabbed an apple and gave it to John. ...

Based on the memory, the process of obtaining the answer is as follows: <problem> How many objects is John carrying? </problem> <memory> <updated-memory> John was carrying an apple initially. He then placed it down in the bathroom. There is no information about John carrying any other objects. </updated-memory> </memory>

Updated memory: Based on the given memory, John was initially carrying an apple. He then placed it down in the bathroom, which means he is no longer carrying it. Therefore, John is carrying 0 objects.

Rejected:

The entire agent workflow consists of parallel processing components and a summarization processing component. Each parallel processing component is as follows: [Part 1] Step 1: ... Step 2: ...

<wrong-updated-memories> **There is no information about John in the provided section. ... From the given section, we can see that the narrative is focused on the little girl, Mary, Daniel, and old Hans. ... Upon re-examining the section, I noticed that there is no mention of John whatsoever. ... The scene is indeed taking place in a rural setting, with the sound of a creek and old Hans' axe in the background. The background sounds do not provide any information about John's presence or actions. Therefore, my updated memory is that there is still no information about John, ...** </wrong-updated-memories>

Updated memory: <updated-memory> There is no information about John in the provided section. ... From the given section, we can see that the narrative is focused on the little girl, Mary, Daniel, and old Hans. ... Upon re-examining the section, I noticed that there is no mention of John whatsoever. ... The scene is indeed taking place in a rural setting, with the sound of a creek and old Hans' axe in the background. The background sounds do not provide any information about John's presence or actions. Therefore, my updated memory is that there is still no information about John, ...

[Part 2] Step 1: ... Step 2: ... Updated memory: <updated-memory> There is a house with a courtyard and stables situated on the fourth side of the quadrangle. The stables are not being used much, as the horses are grazing in an adjoining paddock. John is seen grabbing something in the stables, specifically the milk. ... </updated-memory>

[Part 3] Step 1: ... Step 2: ... Updated memory: <updated-memory> John was carrying an apple initially. He then placed it down in the bathroom. There is no information about John carrying any other objects. </updated-memory>

After processing in parallel, the summarization processing component is as follows: There are 3 processed components (Memories): [Part 1] <updated-memory> There is no information about John in the provided section. However, it does not necessarily mean he is not carrying any objects. We would need more context or information about John to determine the number of objects he is carrying. ... [Part 2] <updated-memory> There is a house with a courtyard and stables situated on the fourth side of the quadrangle. The stables are not being used much, as the horses are grazing in an adjoining paddock. ... </updated-memory> [Part 3] <updated-memory> John was carrying an apple initially. He then placed it down in the bathroom. There is no information about John carrying any other objects. </updated-memory> The summarization of them is as follows: There is no information about John in the provided section. ...

Based on the memory, the process of obtaining the answer is as follows: <problem> How many objects is John carrying? </problem> <memory> <updated-memory> John was carrying an apple initially. He then placed it down in the bathroom. There is no information about John carrying any other objects. </updated-memory> </memory>

Updated memory: Based on the given memory, John was initially carrying an apple. He then placed it down in the bathroom, which means he is no longer carrying it. Therefore, John is carrying 0 objects.

Figure 14: M-Noise Case. The bold parts are redundant noise.

M-Drop Case

Question:

Here is a long context but only a few clues are useful, while others are interferences, you should answer the question following the context after carefully reading it. [Context] A man was seriously wounded in a Long Beach shooting, authorities said Monday.

The shooting was reported at 10:40 p.m. Sunday at Atlantic Avenue and Burnett Street, Long Beach police Lt. James Richardson said.

Arriving officers found a man shot in the upper body, Richardson said.

Paramedics took the man to a hospital where he was in serious condition, the lieutenant said. ...

[Question]What was the condition of the man found by the arriving officers, where was he taken, and what are the details of the incident?

Chosen:

The entire agent workflow consists of parallel processing components and a summarization processing component. Each parallel processing component is as follows: [Part 1] Step 1: ... Step 2: ...

[Part 2] Step 1: ... Step 2: ...

After processing in parallel, the summarization processing component is as follows: There are 2 processed components (Memories): [Part 1] A shooting incident occurred at 10:45 p.m. Tuesday night in the 2700 block of East Seventh Street in Long Beach. Multiple suspects were at large after the shooting. The incident was described as a "mass-casualty incident" and was initially reported as a house party shooting. However, it's unclear whether the shooting occurred during the party or before it. Three men were kil... [Part 2] The incident involved gunfire at a Long Beach home in Southern California, killing at least three people and wounding several others, possibly at a Halloween party that was being held at the house. ... Based on the memory, the process of obtaining the answer is as follows:

<problem> What was the condition of the man found by the arriving officers, where was he taken, and what are the details of the incident? </problem>

<memory> Updated memory:

The incident involved gunfire at a Long Beach home in Southern California, killing at least three people and wounding several others, possibly at a Halloween party that was being held at the house. The exact number of people shot is not immediately available, with reports ranging from 7 to 10 people. ... </memory>

Updated memory: Based on the provided memory and problem, I will answer the question.

The condition of the man found by the arriving officers is not specified. They were taken to a nearby yard outside the Long Beach home. The incident involved a shooting at a house party in Long Beach, Southern California, where three people were killed and nine others were hospitalized.

Rejected:

The entire agent workflow consists of parallel processing components and a summarization processing component. Each parallel processing component is as follows: [Part 1] Step 1: ...

[Part 2] Step 1: ... Step 2: ...

After processing in parallel, the summarization processing component is as follows: There are 2 processed components (Memories): [Part 1] The updated memory will be based on the provided section, as there is no previous memory to retain. Three men were killed in a shooting incident at a Halloween party in the 2700 block of East Seventh Street in Long Beach on Tuesday night. The incident occurred around 10:45 p.m. Multiple suspects were at large following the shooting. ...

[Part 2] The man found by the arriving officers was among the nine people injured in the shooting.

The condition of the man is not immediately available, however, it is mentioned that some of the wounded, including the man, suffered life-threatening gunshot wounds. ...

<problem> What was the condition of the man found by the arriving officers, where was he taken, and what are the details of the incident? </problem>

<memory> The man found by the arriving officers was among the nine people injured in the shooting.

The condition of the man is not immediately available, however, it is mentioned that some of the wounded, including the man, suffered life-threatening gunshot wounds. ... </memory>

Updated memory: Based on the provided memory and problem, I will answer the question.

The condition of the man found by the arriving officers is life-threatening as he suffered a gunshot wound. The location where he was taken is not specified in the article. The details of the incident include a shooting at a Long Beach home in Southern California, resulting in at least three deaths and nine injuries, with the shooter firing indiscriminately into a crowd of 25 to 30 people from an alley behind the residence.

Figure 15: M-Drop Case.

S-Noise Case

Question:

Here is a long context but only a few clues are useful, while others are interferences, you should answer the question following the context after carefully reading it. [Context] Early math knowledge is as important as early literacy for children's subsequent achievement. ... Now a new longitudinal study has found that young children whose mothers supported them during play, specifically in their labeling of object quantities, had better math achievement at ages 4-½ and 5 years. ...

[Question]What did a new longitudinal study find about the impact of mothers supporting their children's labeling of object quantities on math achievement?

Chosen:

Step1:... Step2: <problem> What did a new longitudinal study find about the impact of mothers supporting their children's labeling of object quantities on math achievement? </problem> <memory> <updated-memory> Early math knowledge is crucial for children's subsequent achievement, predicting later school success better than early reading skills and even income in adulthood. ... Mothers' support during play, specifically in labeling object quantities, was found to have a positive impact on math achievement at ages 4-½ and 5 years. ... To assess mothers' support of their children's math skills, researchers examined how moms supported and guided their 3-year-olds' learning as they played with a toy cash register and blocks. ... </memory> <section> While many studies have been conducted on infants' and preschoolers' math competencies, few have evaluated how toddlers' ... Geary and his team including Alex Moore, a postdoctoral fellow, and Kristy vanMarle, an assistant professor in the ... "We measured participants' math skills at the beginning of preschool and again at the end of preschool," Geary said. ... Geary and his team will continue to follow their participants through first grade in this four-year longitudinal study where </section> <wrong-updated-memories> No wrong updated memory </wrong-updated-memories>

Updated memory: <updated-memory> Early math knowledge is crucial for children's subsequent achievement, predicting later school success better than early reading ... Mothers' support during play, specifically in labeling object quantities, was found to have a positive impact on math achievement at ages 4-½ and 5 years. ... To assess mothers' support of their children's math skills, researchers examined how moms supported and guided their 3-year-olds' learning as they played ... Step3:...

Updated memory: <problem> What did a new longitudinal study find about the impact of mothers supporting their children's labeling of object quantities on math achievement? </problem> <memory> ... (same as provided) </memory> <wrong-answer> No wrong answer </wrong-answer>

Since the problem is about the impact of mothers supporting their children's labeling of object quantities on math achievement, and the memory provides information about a study conducted by researchers at Boston College, which found that mothers' support during play, specifically in labeling object quantities, had a positive impact on math achievement at ages 4-½ and 5 years. Therefore, the answer is: Mothers' support during play, specifically in labeling object quantities, has a positive impact on math achievement.

Rejected:

Step1: ... Step2: <problem> What did a new longitudinal study find about the impact of mothers supporting their children's labeling of object quantities on math achievement? </problem> <memory> <updated-memory> Early math knowledge is crucial for children's subsequent achievement, predicting later school success better than early reading skills and even income in adulthood. ... Mothers' support during play, specifically in labeling object quantities, was found to have a positive impact on math achievement at ages 4-½ and 5 years. ... To assess mothers' support of their children's math skills, researchers examined how moms supported and guided their 3-year-olds' learning as they played with a toy cash register and blocks. ... </memory> <section> While many studies have been conducted on infants' and preschoolers' math competencies, few have evaluated how toddlers' ... Geary and his team including Alex Moore, a postdoctoral fellow, and Kristy vanMarle, an assistant professor in the ... "We measured participants' math skills at the beginning of preschool and again at the end of preschool," Geary said. ... Geary and his team will continue to follow their participants through first grade in this four-year longitudinal study where </section> <wrong-updated-memories> **<updated-memory> Early math knowledge is crucial for children's subsequent achievement, predicting later school success better than early reading skills and even income in adulthood. ... Mothers' support during play, specifically in labeling object quantities, was found to have a positive impact on math achievement at ages 4-½ and 5 years. ... To assess mothers' support of their children's math skills, researchers examined how moms supported and guided their 3-year-olds' learning as they played with a toy cash register and blocks. ...** </wrong-updated-memories>

Updated memory: <updated-memory> Early math knowledge is crucial for children's subsequent achievement, predicting later school success better than early reading skills and even income in adulthood. ... Mothers' support during play, specifically in labeling object quantities, was found to have a positive impact on math achievement at ages 4-½ and 5 years. ... To assess mothers' support of their children's math skills, researchers examined how moms supported and guided their 3-year-olds' learning as they played with a toy cash register and blocks. ... Step3: ...

Updated memory: <problem> What did a new longitudinal study find about the impact of mothers supporting their children's labeling of object quantities on math achievement? </problem> <memory> ... (same as provided) </memory> <wrong-answer> No wrong answer </wrong-answer>

Since the problem is about the impact of mothers supporting their children's labeling of object quantities on math achievement, and the memory provides information about a study conducted by researchers at Boston College, which found that mothers' support during play, specifically in labeling object quantities, had a positive impact on math achievement at ages 4-½ and 5 years. Therefore, the answer is: Mothers' support during play, specifically in labeling object quantities, has a positive impact on math achievement.

Figure 16: S-Noise Case. The bold parts are redundant noise.

Evaluation Prompt

[System Prompt of Understanding Tasks]

You are given two memory update trajectories from two different memory-augmented agents (Agent A and Agent B). Each trajectory represents the step-by-step evolution of the agent's memory as it processes a shared input context and iteratively refines its memory to answer a specific question. Both agents start with an initial memory state and update it over several steps, ultimately producing a final answer.

Your task is to compare and evaluate the two memory update paths and select the better agent based on a detailed analysis of their workflows and outputs. Focus on the following aspects:

Evaluation Criteria:

1. Correctness of Each Step in the Work Process:

- Strict Constraint Adherence: Did the agent follow every specific instruction, numerical constraint, and formatting rule mentioned in the question and context? Failure to meet a hard constraint is a critical failure.
- Check for proper preservation of critical information: Did the agent retain all essential details from the context and question at each step? Identify any step where crucial information was incorrectly added, modified, or discarded.
- Ensure faithfulness: Did any step involve hallucination or fabrication of facts not present in the original context? Each update must strictly derive from the provided input.

2. Correctness of Each Step vs. Constraints:

- Constraint Audit: For every step, explicitly check if the agent respected the specific limits.
- Information Fidelity: Identify if any crucial detail was dropped or altered.

Analysis Instructions:

- **Overall Judgment:** Decide which agent has a higher-quality memory update path. Justify your choice by highlighting strengths and weaknesses in their workflows.

- Explanation:

Analyze both trajectories sequentially. Identify specific instances where one agent demonstrates superior reasoning, information fidelity, or efficiency. For example, note steps where Agent A correctly integrates key details while Agent B omits them, or where Agent B avoids redundancy that Agent A introduces.

- Remember, the goal is to determine which agent is better overall, not necessarily which one is flawless. Even small advantages in step correctness or answer quality can sway the verdict.

First, output your verdict by strictly following this format: "[[A]]" if Agent A is better, "[[B]]" if Agent B is better. Then provide your explanation.

[System Prompt of Generation Tasks]

You are given two memory update trajectories from two different memory-augmented agents (Agent A and Agent B). Each trajectory represents the step-by-step evolution of the agent's memory as it processes a shared input context and iteratively refines its memory to answer a specific question. Both agents start with an initial memory state and update it over several steps, ultimately producing a final answer.

Your task is to compare and evaluate the two memory update paths and select the better agent based on a detailed analysis of their workflows and outputs. Focus on the following aspects:

Evaluation Criteria:

1. Correctness of Each Step in the Work Process:

- Strict Constraint Adherence: Did the agent follow every specific instruction, numerical constraint, and formatting rule mentioned in the question and context? Failure to meet a hard constraint is a critical failure.
- Check for proper preservation of critical information: Did the agent retain all essential details from the context and question at each step? Identify any step where crucial information was incorrectly added, modified, or discarded.
- Ensure faithfulness: Did any step involve hallucination or fabrication of facts not present in the original context? Each update must strictly derive from the provided input.

2. Correctness of Each Step vs. Constraints:

- Constraint Audit: For every step, explicitly check if the agent respected the specific limits.
- Information Fidelity: Identify if any crucial detail was dropped or altered.

Analysis Instructions:

- **Overall Judgment:** Decide which agent has a higher-quality memory update path. Justify your choice by highlighting strengths and weaknesses in their workflows.

- **Constraint-Based Comparison:** In your step-by-step analysis, use a "Pass/Fail" mindset for the constraints. For example: "Agent A followed the 5-week interval constraint for VR sessions, whereas Agent B missed the interval in Week 30."

- **Explanation:** Analyze both trajectories sequentially. Identify specific instances where one agent demonstrates superior reasoning, information fidelity, or efficiency. For example, note steps where Agent A correctly integrates key details while Agent B omits them, or where Agent B avoids redundancy that Agent A introduces.

- Remember, the goal is to determine which agent is better overall, not necessarily which one is flawless. Even small advantages in step correctness or answer quality can sway the verdict.

First, output your verdict by strictly following this format: "[[A]]" if Agent A is better, "[[B]]" if Agent B is better. Then provide your explanation.

[User Template]

[Question]

{question (along with a long context)}

[The Start of Agent A's Trajectory]{Response A} [The End of Agent A's Trajectory]

[The Start of Agent B's Trajectory]{Response B} [The End of Agent B's Trajectory]

Figure 17: Evaluation Prompt.

Qwen3-14B Performance Case: Enhanced Reasoning Process

Question:

Construct a skyscraper with 100 floors. Please follow the detailed floor assignments below:

- 1) Specific floors requirements: - Designate Floor 41 for spa and wellness center use. - Designate Floor 38 for martial arts dojo use.- Designate Floor 75 for photography studio use. - Designate Floor 21 for daycare center use. - Designate Floor 35 for private investigator office use.
 - 2) Range floors requirements: Allocate Floors 7 to 11 for a advanced technology and research institute.
 - 3) Periodic floors requirements: Include a coffee shop every 15 floors, starting from Floor 35
 - 4) Document each floor independently with detailed descriptions of the intended facilities, architectural features, and unique design elements.
 - 5) Use '###' to separate the documentation for each floor (e.g. ### Floor 1: [...])
- Ensure that the document consists of 100 entries, each containing at least 150 words. Ensure clarity and continuity without interruptions or omissions in the narrative throughout the document. When the design of all 100 floors is complete, use '*** finished' to indicate the end of the document. *** started ##### Floor 1:

Respond_A: [...]

Respond_B: [...]

Qwen3-14B:

Step-by-Step Analysis and Comparison:

1. Specific Floor Requirements: [...]
2. Range Requirements (Floors 7-11): [...]
3. Periodic Requirements (Coffee Shop Every 15 Floors): [...]
4. Document Clarity and Continuity: [...]
5. Information Fidelity and Faithfulness: [...]
6. Overall Document Quality: [...]

Final Verdict: Agent B produced superior output by strictly following all constraints, ensuring fidelity, and maintaining complete coherent narrative without hallucinations or redundancy. Agent A violated multiple requirements and produced incomplete output. Thus, the better agent is: [[B]].

Qwen2.5-72B-Instruct:

Step-by-Step Analysis:

Floors 1-6 (Initial Floors): [...]

Floors 7-11(Advanced Tech Institute): [...]

Periodic Floors (35-100): [...]

Overall Judgment Final Verdict: Agent A demonstrates stronger ability to create vivid, immersive narratives despite some redundancy. Better agent: [[A]]

Llama-3.3-70B-Instruct:

The skyscraper design presented by Agent A demonstrates a more comprehensive and detailed approach to meeting the requirements outlined in the prompt. Each floor is meticulously described, with a focus on architectural features, amenities, and the specific needs of various users, from corporate executives and entrepreneurs to wellness enthusiasts and residents. The inclusion of specific details such as materials used for flooring and walls, the type of lighting, and the integration of technology for smart building management showcases a thorough understanding of what makes a modern, functional, and luxurious skyscraper. Agent A's design also prioritizes sustainability, wellness, and community, with the inclusion of green spaces, fitness centers, meditation rooms, and social lounges. This holistic approach to building design not only enhances the quality of life for occupants but also contributes to a healthier and more sustainable urban environment. Furthermore, Agent A's trajectory adheres closely to the specified requirements, including the allocation of specific floors for particular uses (e.g., the spa and wellness center on Floor 41, the photography studio on Floor 75), and incorporates periodic floors for coffee shops starting from Floor 35, as per the instructions. This adherence to the brief demonstrates a strong ability to follow instructions and meet client needs. In contrast, while Agent B's design also showcases luxury and functionality, it lacks the depth and specificity of Agent A's descriptions. Agent B's floors, though well-appointed and modern, sometimes lack the detailed consideration of user experience and the integration of sustainable and wellness-focused elements that are hallmarks of Agent A's design. Therefore, based on the criteria of adherence to instructions, detail and specificity of design, and consideration of user experience and sustainability, the better agent is: [[A]]

Key Insight: Qwen3-14B demonstrates enhanced post-training reasoning by identifying *constraint violations* that larger models missed.

Figure 19: Case study demonstrating Qwen3-14B's enhanced reasoning process after post-training. The model correctly identifies constraint violations and maintains strict fidelity to requirements, while baseline models prioritize descriptive quality over instruction adherence.

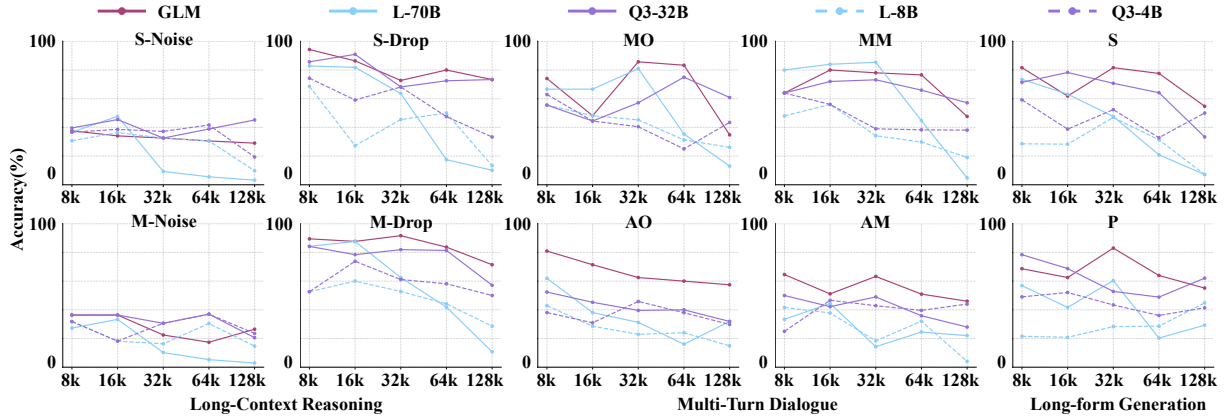


Figure 20: Trends in RM consistency performance of each subtask with respect to memory management trajectory length.

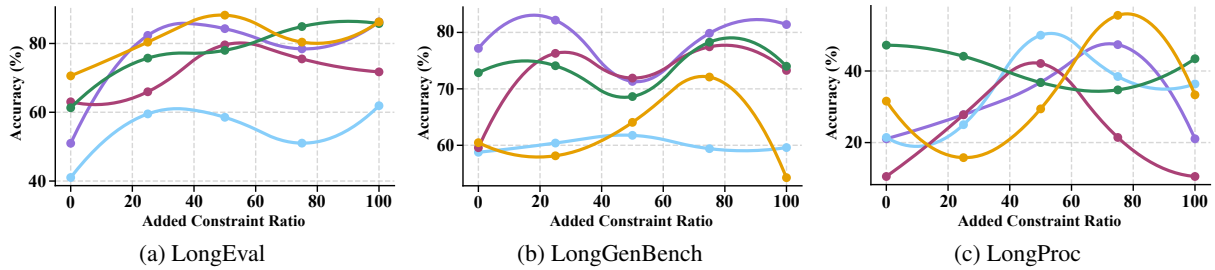


Figure 21: Overall performance trends on LongEval, LongGenBench, and LongProc as the constraint density varies in long-form generation tasks.

LongGen Ablation

[0%]

Generate a complete weekly diary for Benjamin for the entire year of 2018

[25%]

Benjamin is a architect with a passion for chronicling his vibrant life through weekly diary entries. Captures:1) Family member birthday: husband (yourself) (birthday on December 07), my wife (birthday on November 21)

[50%]

Benjamin is a architect with a passion for chronicling his vibrant life through weekly diary entries. Captures:1) Family member birthday: husband (yourself) (birthday on December 07), my wife (birthday on November 21), my child_1 (birthday on September 21), my child_2 (birthday on January 19), my child_3 (birthday on November 20)

[75%]

Benjamin is a architect with a passion for chronicling his vibrant life through weekly diary entries. Captures:1) Family member birthday: husband (yourself) (birthday on December 07), my wife (birthday on November 21), my child_1 (birthday on September 21), my child_2 (birthday on January 19), my child_3 (birthday on November 20), 2) Attending a traditional dance festival in India in week 30-31.3) Attend a DIY pottery and ceramics class every 5 weeks on weekends, starting from week 11.

[100%]

Benjamin is a architect with a passion for chronicling his vibrant life through weekly diary entries. Captures:1) Family member birthday: husband (yourself) (birthday on December 07), my wife (birthday on November 21), my child_1 (birthday on September 21), my child_2 (birthday on January 19), my child_3 (birthday on November 20), 2) Attending a traditional dance festival in India in week 30-31.3) Attend a DIY pottery and ceramics class every 5 weeks on weekends, starting from week 11. 4) Weekly updates on weather changes, work developments, family life, and other interesting topics.5) Use '##' to separate each weekly entry (e.g. ## Week 1 (January 1st - January 7th):

Figure 22: LongGen Ablation.

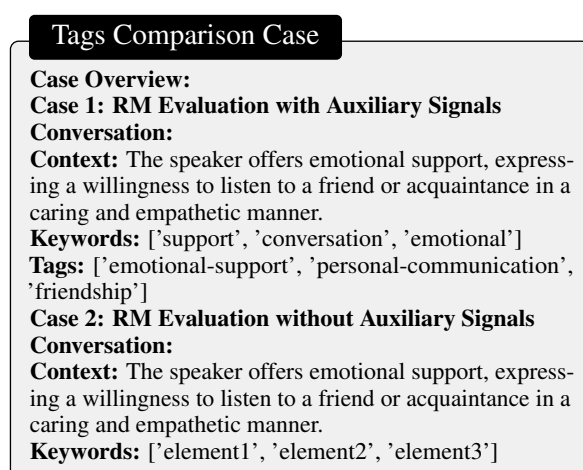


Figure 23: Comparison of RM evaluation for multi-turn dialogue understanding in a memory system, with and without auxiliary signals.