

Memory Makes the Difference: Evaluating How Different Memory Roles Shape Conversational Agents

Yuxin Wang^{1*} Paul Thomas² Zhiwei Yu[†] Yuan Gao² Saeed Hassanpour¹
Soroush Vosoughi¹ Robert Sim² Nick Craswell²

¹Dartmouth College ²Microsoft

Abstract

Memory is a critical component of modern conversational systems. Prior research on memory mechanisms in RAG-based conversational systems has focused on how memory is stored and retrieved. However, less is known about how memories with different functional roles influence response quality, shape agents’ responses across conversational contexts, and lead to distinct response behaviors. Existing evaluations are also largely reference-based, often missing response nuances that address users’ preferences differently. In this work, we study the impact of different memory types on agent responses. We present a fine-grained taxonomy of conversational memory, classify retrieved memories by functional role, and design a user-centric evaluation framework that simulates user perspectives. Through comparative experiments on long-term datasets and frontier LLMs, our analysis reveals differentiated effects of memory: *e.g.*, clarifying memory is associated with improved factual accuracy and constraint awareness, making responses more correct and personalized, while irrelevant memory reduces topic relevance and degrades constraint awareness. These findings show that, despite the power of frontier LLMs, different memory types can be leveraged to produce responses better aligned with a user’s stated constraints, and motivate further research in this direction. Our code is available at: <https://github.com/audreycs/MemoryEffect>.

1 Introduction

Conversational memory provides needed context for grounded responses and accurate query understanding (Radlinski and Craswell, 2017), and is now a core feature of contemporary chat systems such as ChatGPT (OpenAI, 2025b) and Google

^{*}This work is done during the author’s internship at Microsoft. Contact address: yuxin.wang.gr@dartmouth.edu or nickcr@microsoft.com.

[†]Independent researcher.

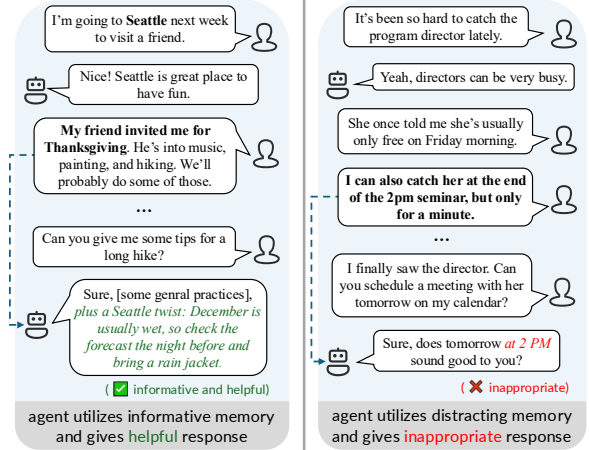


Figure 1: (Left panel) Example of retrieved memory making response more helpful; (Right panel) Example of retrieved memory making response inappropriate.

Gemini (AI and DeepMind, 2025). At a high level, conversation history is logged into an external memory bank to preserve past information. During response generation, chat agents retrieve relevant pieces to surface useful memory. This retrieval-augmented generation (RAG) pipeline has become a standard approach for long-range conversation, helping Large Language Models (LLMs) overcome context-length limits and maintain response coherence and accuracy (Fan et al., 2024).

Existing work on conversational RAG has largely centered on storage and retrieval strategy, *i.e.*, memory structure, retrieval size, and granularity (Xu et al., 2022; Ye et al., 2024; Maharana et al., 2024; Zhong et al., 2024; Wu et al., 2025; Pan et al., 2025; Tan et al., 2025a). These systems typically retrieve a set of stored pieces, concatenate or compress them into the conversational context, then evaluate the aggregated performance of generation system. However, this common practice overlooks the distinct role properties and impact of individual retrieved pieces. Human intuitively recall useful memory to formulate responses, but when it comes

to RAG systems, it is uncertain if retrieved memory which are semantically close are actually useful. The effect of retrieval part often behaves like a black box and hard to interpret or control (Chen et al., 2024). As Fig. 1 shows, some supplementary memory can enrich the response and make it more helpful; while other memories can be distracting and make the response inappropriate and undesired. In this work, our central hypothesis is that **the type of memory matters**: different memory types elicit different agent responses, which in turn shape users’ perceptions of quality. Therefore, understanding the qualitative effect of each memory type and whether agents can leverage a bundle of retrieved memories worth systematic investigation. Building on this hypothesis, we aim to quantify the impact and find their qualitative property. We introduce a fine-grained taxonomy of conversational memory types, determined by their functional roles when the agent reasons over retrieved memory. Unlike prior taxonomies, which organize memory by structure or provenance, ours conditions on the query and decomposes what “relevant” memory does in a dialogue (§2.2). Besides, existing work is usually limited in reference-based evaluation. Prior work leans on question answering (QA) datasets provided with unambiguous gold answers (Anantha et al., 2021; Xu et al., 2022; Pan et al., 2025; Maharana et al., 2024; Chen et al., 2025). Their reference-based evaluation privileges measuring factual accuracy (*e.g.*, F1, ROUGE) while ignoring other qualities of responses like user intent awareness and informativeness, which are critical for dialogic intelligence but difficult to measure using only references.

In this work we analyze how different memory types shape LLM agents’ responses, paired with a holistic, user-centric evaluation framework. Our taxonomy classifies conversational memory into five types — answer, clarifying, enriching, distracting, and irrelevant, judged with awareness of the query and relevant context instead of only the gold answer as in previous analyses (Cuconasu et al., 2024). We conducted comparative experiments on two long-term, multi-session conversational datasets over three popular LLMs. Specifically, we compare varied memory settings with (i) answer memory only and (ii) parametric memory only, to study how the presence of individual types and their combinations affects responses quality. To better differentiate response quality, we design a user-centric evaluation proxy conditioned on user

intent and prior conversation context, capturing relevance, informativeness, and other properties.

Our results reveal many interesting patterns. Frontier LLMs can selectively leverage contextual information with considerable nuance, but agents are less likely to ground responses as retrieval size increases. As for memory types, clarifying memory plays a key role in improving answer accuracy and contextual grounding, and even in isolation it can induce perfect responses. In contrast, irrelevant memory tends to reduce topic relevance and weaken contextual awareness, unlike the positive impact reported for small-scale LLMs such as Llama2 and Phi-2 in previous work (Cuconasu et al., 2024). These also confirm our hypothesis on the impact of memory types. Our taxonomy and user-centric evaluation are general-purpose and can be applied to other conversational corpora and settings. Our main contributions in this work are the following:

- We hypothesize the impact of memory types on agent’s responses, and conducted extensive comparative experiment on frontier LLMs to investigate.
- We propose a fine-grained taxonomy of conversational memory based on functional roles, together with a holistic framework with diverse metrics for evaluation.
- Our experiments qualitatively reveal the distinct impacts of different memory, with several effects diverging from prior findings on small-scale LLMs.

2 Related Work

2.1 Memory-enabled Conversational System

Conversational systems typically frame memory recall as a store-retrieve-generate pipeline, reflecting flexible memory management and working context (Gao et al., 2023). Recent work on LLM-powered conversational systems have largely focused on memory structure (Packer et al., 2023; Rasmussen et al., 2025; Chhikara et al., 2025), retrieval strategy (Tan et al., 2025b; Li et al., 2025; Hu et al., 2025a), and memory update governance (Tan et al., 2025b; Reza-zadeh et al., 2025; Wang et al., 2025). Retrieval commonly uses sparse or dense retrievers, such as BM25 (Robertson et al., 2009) and Contriever (Izacard et al., 2021). To stress-test these systems in realistic settings, several work target long-term, multi-session dialogue, including MT-

Bench-Plus (Lu et al., 2023), LOCOMO (Maharana et al., 2024), LongMemEval (Wu et al., 2025), and InfiniteBench (Zhang et al., 2024).

2.2 Analysis of Retrieval Pieces Type

Several recent studies examine how irrelevant or distracting retrieved pieces affect QA performance, mainly on smaller-scale LLMs. Cuconasu et al. (2024) provide a pioneering analysis of how retrieved document type and position influence model answers. They classify documents into gold, relevant, distracting, and random based on semantic similarity and gold-answer presence, and show that random documents can sometimes improve accuracy. Jin et al. (2025) investigate distracting documents that are relevant to the query but not the gold answer, finding consistent harm to QA performance across models. Later, Amiraz et al. (2025) propose a measure to quantify the distracting effect. Together, these works characterize the impact of irrelevant and distracting evidence in document QA, but give less attention to conversational settings. Inspired by this line of work, we adapt and refine these types for conversational scenarios, shifting from document QA to memory-enabled dialogue, where responses require more nuanced evaluation and multiple forms of “relevant memory” (*e.g.*, clarifying, enriching) must be considered.

2.3 Evaluation of Conversational Agents

Conventional evaluation of conversational systems scores responses against references using n-gram or embedding metrics such as F1, ROUGE (Lin, 2004), and BERTScore (Zhang* et al., 2020). Newer work on long-term, memory-enabled conversations also tracks retrieval accuracy by measuring if particular pieces are retrieved (*i.e.*, needle in the haystack) via metrics like Recall@ k (Maharana et al., 2024; Wu et al., 2025). With the rise of LLM-as-a-judge, LLM-based judges have been increasingly used for holistic chat-quality evaluation (Bai et al., 2024; Zhao et al., 2025; Chhikara et al., 2025). Well-designed rubrics allow them to assess diverse dimensions even without references. For example, Mem0 (Chhikara et al., 2025) prompts an LLM judge to compare responses with references, MT-Bench-101 (Bai et al., 2024) uses pairwise comparison to rank candidate responses, and PersonaLens (Zhao et al., 2025) equips judge with user profiles and rich interaction history to assess whether responses address user preferences.

Memory Type (Abbr.)	Definition
Parametric memory (P)	Exists in LLMs’ parameters through training
Answer memory (A)	Contains the information of the desired answer for user’s query
Clarifying memory (C)	Resolves the identity of the asked-for entity or event in the query
Enriching memory (E)	Makes the final answer more informative or helpful
Distracting memory (D)	Seems to be relevant but is actually misleading
Irrelevant memory (I)	Has no meaningful relevance to the user’s query (off-topic)

Table 1: Our taxonomy of conversational memory types.

3 Taxonomy of Conversational Memory

Retrieved memories form a semantic neighborhood around the query by RAG mechanism, and the agent’s reasoning can be viewed as selectively traversing retrieved memory pieces to assemble the information needed for a response (Appendix D for illustration). Building on this intuition and document RAG classification (Cuconasu et al., 2024; Amiraz et al., 2025), we classify conversational memory by its role in reasoning into five types: *answer*, *clarifying*, *enriching*, *distracting*, and *irrelevant*. Tab. 1 shows their definitions. As shown in Fig. 12, to correctly answer the query, the agent needs clarifying memory to identify the referenced name, then infer answer from the answer memory. Ideally, the agent needs enriching memory for more helpful details, and avoids distracting memory to mitigate false answer. The remaining off-topic memory is irrelevant, serving as neutral noise.

Each memory receives one type per query, although a memory may contain pieces serving multiple roles. For controlled analysis and clear attribution, we adopt single-label assignment as a deliberate methodological simplification, and we apply a labeling precedence: *answer* → *clarifying* → *enriching* → *distracting* → *irrelevant* for disambiguation. This precedence encourages sufficient representation of relatively sparse types in realistic retrieval pools. Importantly, it does not encode a hierarchy of importance or quality of them on response generation. Two consequences follow. First, since clarifying outranks all types but answer, only answer can absorb a clarifying piece, making clarifying a conservative subset; the scarcity of clarifying memory in §7 is thus a lower bound reflecting retrieval behavior, not an artifact of the rule. Second, we do not measure how often pieces are genuinely multi-role, nor how sensitive our conclusions are to a different ordering; multi-label annotation would answer both and expose combinations single labeling hides (*e.g.*, answer-containing yet

distracting). Besides conversational memory, we include *parametric* memory, the LLM’s inherent knowledge learned during training.

To identify types, prior work (Cuconasu et al., 2024; Amiraz et al., 2025) flags distracting or irrelevant retrieved content by string-matching with the gold answer. This works for short, unambiguous targets (e.g., a name or location), but not for open-ended dialogue responses, which are more expansive and lexically diverse. We therefore use a context-aware LLM-as-a-judge to simulate the user and classify retrieved memory pieces. For each candidate memory, the judge receives the query, the candidate piece, and the top 50 retrieved memory pieces as relevant interaction history, helping it infer the query and user’s intent. Fig. 7 presents the details, using a prompt schema adapted from Thomas et al. (2024), which has been shown to be highly effective. We use GPT-5-thinking model (OpenAI, 2025a) as the judge.

Because memory classification underpins the following analyses, we also conduct a human verification study. On a sample of 30 pieces, the judge’s agreement with human annotators is comparable to the agreement among the annotators themselves (Appendix B). We treat this as a pilot sanity check rather than a full validation: the sample is small, absolute agreement is moderate, and we report no per-class results for the sparse clarifying category, which carries our strongest claim.

4 User-centric Response Evaluation

To better capture the sensitivity of responses to different context, we design a holistic, user-centric, and context-aware evaluation framework (Tab. 2). It spans three major aspects (Accuracy, Relevance, and Informativeness), each with finer metrics, by integrating established evaluation frameworks (Liang et al., 2023; Maurya et al., 2025) and adapting them to conversation. At a high level, Accuracy captures factual correctness of responses; Relevance checks alignment with user intent and preferences; Informativeness assesses usefulness beyond the minimal answer. Note that conventional reference-based accuracy metric such as ROUGE is subsumed by Answer Accuracy, which tests whether the response contains the desired answer. We nonetheless report ROUGE-1 Recall separately, alongside Memory Coverage, as a lexical-overlap check. Tab. 2 marks which metrics require prior conversation history (“context-aware” column); for instance, World

Knowledge Accuracy and Internal Coherence can be judged without context.

For robustness, most metrics except Completeness use a binary scale rather than a fine-grained rating scale. Prior work shows binary judgments can improve LLM-judge accuracy and reliability (Dong et al., 2024; Alhawasi and Youssef, 2024), while finer scales introduce subjective design choices (Bansal et al., 2024; Wang et al., 2024). Thus, we believe binary scale is sufficient for our analyses. We use GPT-o3 as the LLM judge, judging metrics under the same aspect together.

5 RAG Pipeline Implementation

Our implementation follows the standard conversational RAG pipeline. In our setting, a *conversation* is the complete interaction history between one user and the chat agent. Each user has a separate conversation, which may span multiple sessions. A *session* is a continuous interaction episode between a user and the agent, typically centered on one topic. An *utterance* is a single text input from either side, and a *turn* consists of one user utterance followed by one agent utterance. For each user query, *context* denotes the retrieved prior memories reachable to the agent.

5.1 Two Long-Term Conversational Datasets

We use two long-term, multi-session conversation datasets for our analyses: LongMemEval-m (Wu et al., 2025) and Long-MT-Bench+ (Zheng et al., 2023), with statistics in Appendix E, Tab. 5. LongMemEval-m contains synthetic conversations paired with human-designed queries, while Long-MT-Bench+ contains human-written conversations and GPT-generated long-range queries. Both datasets feature extensive conversational memory and evaluate memory-related user queries. LongMemEval-m additionally provides metadata such as session timestamps and evidence memories for each query.

5.2 Memory Storage and Retrievers

We use one memory bank for all sessions within the same conversation. When the agent responds to a query, all past sessions are reachable, but different conversations cannot access each other’s memory. Following LongMemEval, we implement the memory bank as key-value pairs {<key>: <value>}, where the key is used for retrieval indexing and the value returns the original content. We **store each**

Aspect	Metric	Description	Range	Context-aware
Accuracy	World Knowledge	Accuracy of the response with respect to general facts and world knowledge	{0, 1}	✗
	Answer Accuracy	Agreement of the response with the gold answer	{0, 1}	✗
	Internal Coherence	Logical consistency of the response within itself	{0, 1}	✗
Relevance	Topic Relevance	Whether the response directly addresses the query	{0, 1}	✓
	Constraint Awareness	Whether the response satisfies constraints stated or implied in prior interactions	{0, 1}	✓
Informativeness	Completeness	Extent to which the response covers relevant information for the query	{0, 1, 2}	✓
	Actionability	Whether the response provides concrete details or clear guidance	{0, 1}	✗
	Memory Coverage	Number of retrieved memory pieces whose information is reflected in the response	$\mathbb{Z}^+$	✓
	ROUGE-1 Recall	Proportion of unigrams in the gold answer that appear in the response	[0, 1]	✗

Table 2: User-centric evaluation framework.

turn as one pair; both key and value are the original turn, although the key can be customized (*e.g.*, as a summary or keyword list). When timestamps are available, we record them in the values. For retrieval, we experiment with BM25 (Robertson et al., 2009), Contriever (Izacard et al., 2021), and OpenAI’s text-embedding-3-small (Neelakantan et al., 2022). We retrieve the top 100 memory pieces by similarity. On average, a memory piece contains 317 tokens in LongMemEval-m and 216 tokens in Long-MT-Bench+, so 100 pieces already form a large context.

5.3 Agent Models and Response Generation

We evaluate three large and popular LLMs as agents: GPT-4o-2024-08-06 (Hurst et al., 2024) (called GPT-4o later), Gemini-2.0-Flash (DeepMind, 2024), and Llama-3.3-70B (Dubey et al., 2024). To vary context size, we select the top k pieces ($k \in \{0, 5, 10, 20, 50, 100\}$) from the 100 retrieved pieces. When timestamps are available, the k pieces are ordered chronologically. We concatenate them and truncate as needed to fit the model’s context window. The full response-generation prompt is shown in Fig. 6 in Appendix. Note that when $k = 0$, the agent relies on parametric memory. All agents use the minimum temperature for near-deterministic output.

5.4 Response Evaluation

We use GPT-o3 with high reasoning effort as the judge, applying aspect-specific rubrics. We carefully designed the rubrics, sanity-checked them on sample queries, and iteratively refined them to align the judge’s scores with human evaluations (Appendix C). Unlike the memory-type judge, however, this judge was aligned to human-checked cases during rubric design but never scored against an independent set of human judgments (see Limitations). Particularly, for the Topic Relevance, refusal responses such as “*I am unable to answer ...*” receive

a score of 0.

6 General Performance On Large LLMs

We first analyze metric-level performance across agents, context sizes, and retrievers, then examine agents’ memory selectivity by type.

6.1 Performance Over Datasets, Agents, Context Size, and Retrievers

Fig. 2 shows average scores under different context sizes (top k) and agents, while Fig. 3 compares retrievers. We find several patterns:

Context Size-Performance Trade-off Metrics such as Answer Accuracy and Constraint Awareness generally rise as context size k increases, then decline. This suggests excessive context can distract agents, echoing prior findings on information overload (Maharana et al., 2024; Pan et al., 2025; Wu et al., 2025). Among the three agents, GPT-4o is most robust to long contexts.

World Knowledge Accuracy and Constraint Awareness Decline In Fig. 2, agents using only parametric memories ($k = 0$) receive higher World Knowledge Accuracy than with retrieval. Inspection suggests added conversational context leads agents to generate more claims and contextualized statements, which judges penalize when unsupported. Thus, external evidence can ground responses but also amplify hallucination risks, consistent with recent studies (Jin et al., 2024; Hu et al., 2025b). Constraint Awareness also degrades once context exceeds a certain size, even when Answer Accuracy stays relatively stable (*e.g.*, on Long-MT-Bench+), highlighting the need to triage historical context, especially in preference-heavy tasks such as recommendation.

Memory Coverage Cap Memory Coverage and Completeness, which measure how much retrieved memory is reflected in the response, show a clear upper bound. Coverage peaks when $k \in [5, 50]$ and then declines. Top-ranked pieces tend to contain

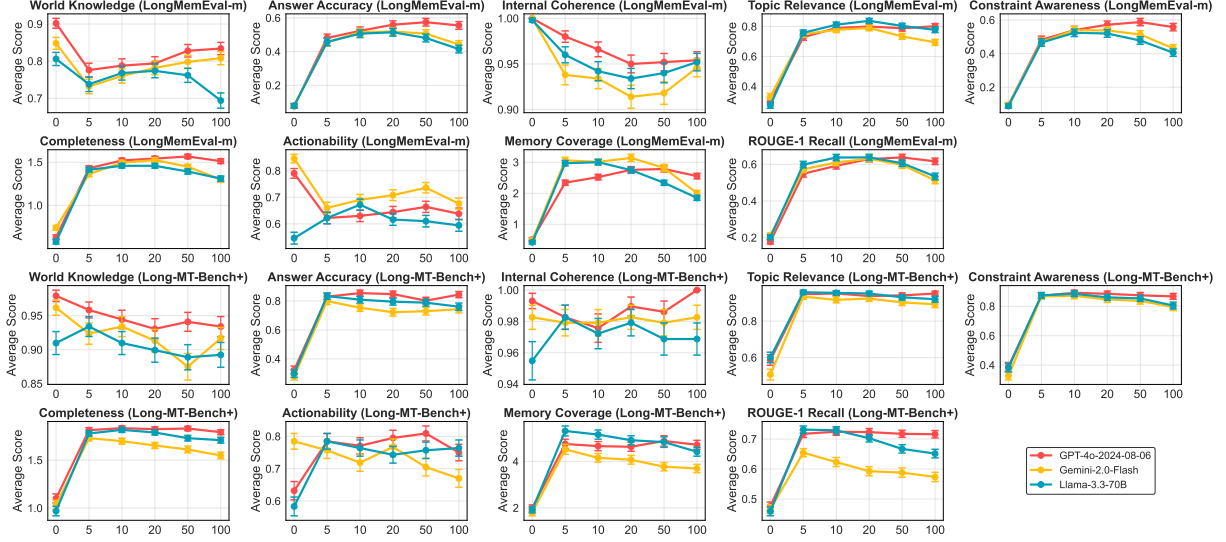


Figure 2: Average metric scores with error bars over queries on LongMemEval-m and Long-MT-Bench+ datasets. Results are based on text-embedding-3-small retriever. X-axis is the context size used in the prompt (top k memory pieces).

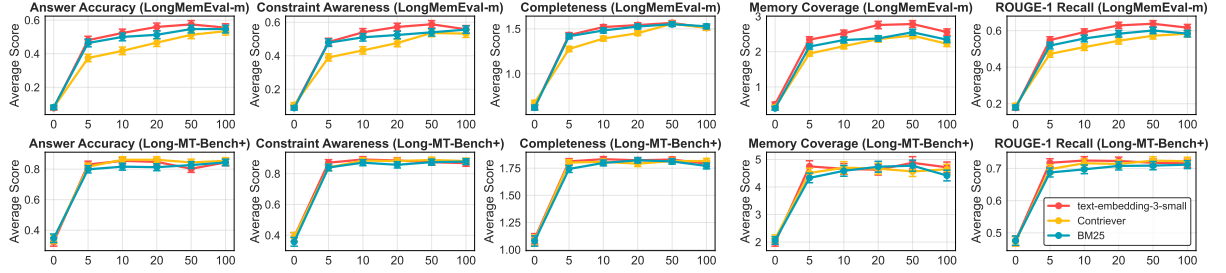


Figure 3: Average metric scores with error bars over context- or reference-aware metrics of three retrieval methods on LongMemEval-m and Long-MT-Bench+ datasets. Results are based on GPT-4o-2024-08-06 agent. X-axis is the context size used in the prompt (top k memory pieces).

the most useful information, while later pieces are less relevant; as context grows, irrelevant pieces distract the agent from using useful ones.

Retriever Comparison Fig. 3 compares retrievers on key retrieval- or reference-dependent metrics. Overall, OpenAI’s text-embedding-3-small outperforms the other two, especially on LongMemEval-m and smaller k , indicating better ranking of useful memory. BM25 remains competitive despite its simplicity, showing lexical match still finds strong anchors. Performance converges when enough memory is retrieved.

6.2 Utilization of Different Memory Types

We also measure how often each memory type appears in responses as context size k varies. Under our taxonomy, irrelevant memory should be ignored and answer memory should be included, providing a sanity check for the labels. For a given type, let a be its number of pieces among

the top k retrieved (the *seen* number) and b the number echoed in the response (the *utilized* number); the coverage rate is b/a , the percentage of seen pieces used. Fig. 4 shows results on GPT-4o. Answer memory is covered most often, while irrelevant memory is covered least, suggesting the agent can distinguish useful from unuseful memory. Clarifying memory is also highly utilized, supporting our taxonomy. As k increases, coverage declines for every type, steeply for distracting memory ($0.37 \rightarrow 0.18$) and more gently for the rest.

7 Impact of Different Memory Types

We next examine how distinct memory types shape response quality. We conduct controlled experiments to characterize systematic associations between memory-type presence and response behavior under realistic retrieval. We run two comparison sets: (i) variant settings that add another memory type to an answer-only baseline (vs. A);

Comparative Experiment (vs. A)	World Knowledge		Answer Accuracy		Internal Coherence		Topic Relevance		Constraint Awareness		Completeness		Actionability		Memory Coverage		ROUGE-1 Recall	
A+C	0.50	1.00	0.60	1.00	–	–	1.00	1.00	0.75	1.00	0.89	1.00	0.67	1.00	0.67	0.75	0.57	0.67
A+E	0.48	0.28	0.53	0.65	0.30	0.33	0.33	0.80	0.50	0.71	0.63	0.68	0.42	0.55	0.73	0.59	0.52	0.55
A+D	0.33	0.20	0.42	0.50	0.33	0.00	0.33	0.50	0.39	0.33	0.49	0.75	0.43	0.00	0.65	0.73	0.39	0.63
A+I	0.46	0.46	0.46	0.53	0.63	0.67	0.11	0.57	0.38	0.36	0.52	0.53	0.40	0.55	0.64	0.59	0.52	0.62
C+E+D+I	0.34	0.23	0.05	0.10	0.67	0.50	0.01	0.08	0.03	0.04	0.09	0.20	0.40	0.30	0.37	0.38	0.16	0.22

Comparative Experiment (vs. P)	World Knowledge		Answer Accuracy		Internal Coherence		Topic Relevance		Constraint Awareness		Completeness		Actionability		Memory Coverage		ROUGE-1 Recall	
C	0.00	0.00	–	0.96	0.00	0.50	0.56	1.00	–	1.00	0.77	0.89	0.43	0.53	1.00	0.87	1.00	0.85
A+C	0.00	–	0.89	1.00	0.50	–	0.90	1.00	1.00	1.00	0.93	0.83	0.80	1.00	1.00	1.00	0.92	1.00
E	0.26	0.04	0.80	0.91	0.00	0.14	0.82	0.96	0.76	0.88	0.74	0.75	0.32	0.69	0.85	0.83	0.82	0.74
D	0.20	0.08	0.86	0.75	–	0.00	0.80	0.87	0.80	0.77	0.70	0.67	0.27	0.62	0.92	0.87	0.79	0.60
I	0.26	0.09	0.50	0.51	0.00	0.29	0.68	0.64	0.43	0.64	0.49	0.46	0.29	0.34	0.86	0.72	0.75	0.57
A+C+E+D+I	0.29	0.18	0.96	0.97	0.00	0.40	0.95	0.96	0.97	0.97	0.95	0.92	0.28	0.75	0.97	0.88	0.93	0.91

Table 3: Improvement rate IR of comparison experiments on LongMemEval-m (left number in each cell) and Long-MT-Bench+ (right number in each cell) on GPT-4o. **Top table:** experimental setting vs. controlled setting (A); **Bottom table:** experimental setting vs. controlled setting (P). IR value stands for the chance of experimental setting wins over controlled setting, and IR = 0.5 means a tie. “—” means a lack of data point for the result. We highlight $IR \in [0.8, 1.0]$ with heavy green, $[0.6, 0.8)$ with light green, $(0.2, 0.4]$ with light red, and $[0, 0.2]$ with heavy red.

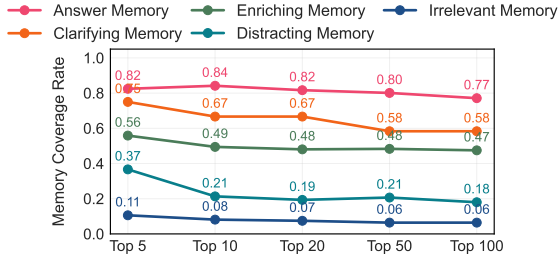


Figure 4: Memory coverage rate of different memory types under different retrieval size k . Results are on GPT-4o.

and (ii) variant setting compared to a parametric-only baseline (vs. P). The answer-only baseline reflects the RAG assumption that isolated answer-containing pieces should make the query easy to answer. Because real retrieval also includes non-answer memory, we test how C/E/D/I memory changes responses. The parametric-only baseline tests the converse assumption: without retrieval from conversation history, the agent cannot answer user queries well. We study the top 15 retrieved pieces, determine their types, vary memory settings, and use text-embedding-3-small as the retriever. Other setups remain unchanged.

7.1 Comparison with Answer Baseline

We compare the vs. A baseline with five variants: A+C, A+E, A+D, A+I, and the non-answer bundle C+E+D+I, yielding six settings and five pairwise comparisons. Because memory types are judged with query and context, the composition and

counts within the top 15 pieces vary; a retrieval may contain one clarifying (C) piece but more than ten irrelevant (I) pieces. We therefore retain original retrieved counts to reflect realistic retrieval. Queries without answer memory (A) are excluded because they lack a comparable answer baseline. This buys ecological validity at a cost: piece count — and hence token budget and evidence position — correlates with role composition, since C memory typically adds one or two pieces while I memory can add ten. Our comparisons therefore identify **associations between role composition and response behavior under realistic retrieval**, not isolated causal effects; an effect attributed to a role may partly reflect context length, attention dilution, or position. Separating them requires count-, token-, and rank-matched variants holding the prompt budget fixed, so we state findings at the associational level throughout.

The idea to quantify each variant setting’s impact is checking how its metrics scores change compared to baselines. Formally, let $\mathcal{Q}$ denote the query set. For metric m and memory setting i , let $s_q^m(i)$ be the score of the agent’s response to query $q \in \mathcal{Q}$. The per-query score difference between settings i and j is $\text{diff}_q^m(i, j)$:

$$\text{diff}_q^m(i, j) = s_q^m(i) - s_q^m(j).$$

For instance, $\text{diff}_q^m(A + C, A) > 0$ means adding clarifying memory improves metric m for query q over the answer-only baseline. We measure how often setting i outperforms setting j using

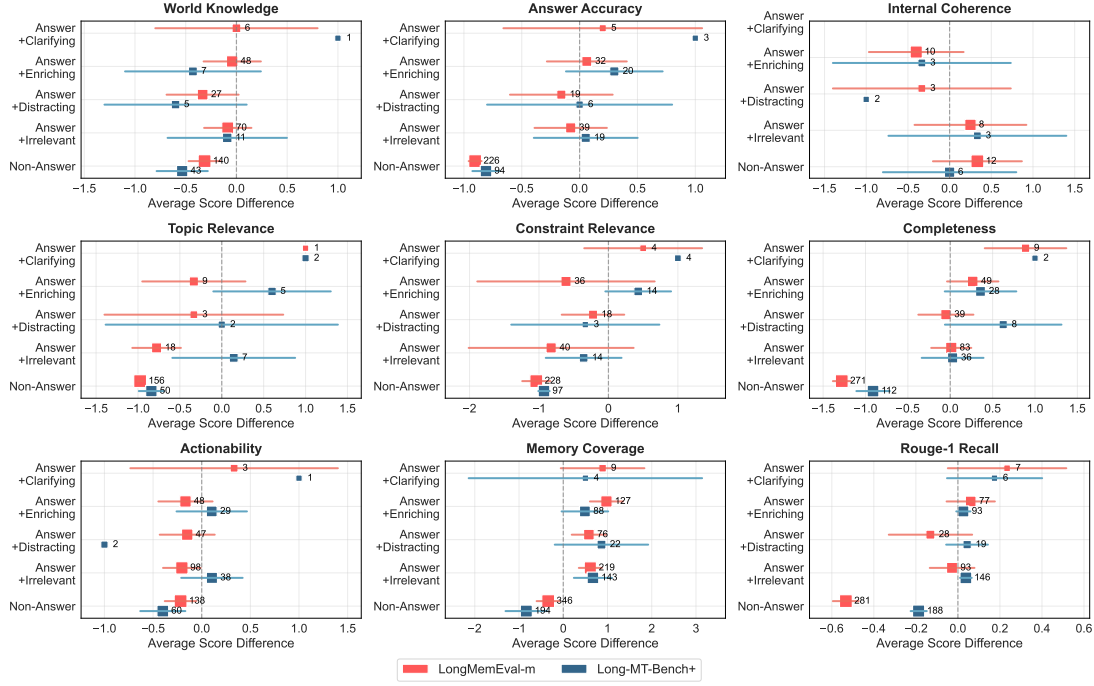


Figure 5: Mean per-query score differences with 95% confidence intervals when adding a memory setting relative to the answer-only baseline on GPT-4o. Squares mark the mean, bars the interval. Square size and the adjacent number give the count of queries with a non-zero change; no square means no qualifying queries.

Improvement Rate (IR):

$$\text{IR}^m(i, j) = \frac{\#\{\text{diff}_q^m(i, j) > 0\}}{\#\{\text{diff}_q^m(i, j) \neq 0\}} \in [0, 1].$$

IR reflects the direction of score changes under matched retrieval settings. An $\text{IR}^m(i, j) = 0.7$ means that, switching from setting j to setting i , 70% of the **score changes** improve. Ties are excluded by construction: most metrics are binary, so unaffected queries would otherwise dominate the denominator and compress every comparison toward 0.5. IR thus reports direction among affected queries, not effect magnitude nor the share of the query set that benefits — improving one query and leaving ninety-nine unchanged also gives $\text{IR} = 1.00$ — and should not be read as an effect size. We report IR for each metric. The top table in Tab. 3 shows GPT-4o results; other agents are in the Appendix.

Because memory-type distributions vary in the top 15 pieces, some types (*e.g.*, clarifying) are scarce while others (*e.g.*, irrelevant) are common, leading to uneven sample sizes. To assess statistical significance, we collect $L^m(i, j) = \{\text{diff}_q^m(i, j) \mid q \in \mathcal{Q}, \text{diff}_q^m(i, j) \neq 0\}$ and calculate its mean with 95% confidence interval (ci_lo , ci_hi). The interval is Student- t on the sample mean, $\bar{x} \pm t_{0.975, n-1} s / \sqrt{n}$, with $n = |L^m(i, j)|$. Like IR

it conditions on non-zero differences, so it gives the size of a change *given that* the score changed, not the average paired effect over all queries, which would be attenuated toward zero. Fig. 5 visualizes these results with blobbograms. Key findings are:

Significance of Clarifying Memory The top table in Tab. 3 shows that, among queries whose scores change, clarifying memory (C) improves them almost uniformly on both datasets, especially for Relevance and Informativeness, and Fig. 5 shows consistently positive means in the first row. We rest the claim on this directional stability across agents and datasets, not on magnitude: clarifying pieces are rare in the top 15, so most cells here rest on single-digit counts and the values of 1.00 summarize a handful of queries. At that strength, the pattern suggests lightweight disambiguation signals are high-leverage when retrieved at all — a hypothesis worth targeting, not an established effect size.

Enriching Memory Improves Informativeness Enriching memory (E) generally makes responses more informative, reflected in higher Completeness and Memory Coverage. Tab. 3 also shows a small, statistically non-significant drop in Internal Coherence (Fig. 5). Manual inspection suggests the judge penalizes verbosity and unfocused discussion: responses become richer but less organized.

Distracting Memory is Harmful The top table in

Tab. 3 shows that distracting memory (D) degrades performance across most metrics, most clearly on World Knowledge, Internal Coherence, and Constraint Awareness; Answer Accuracy and Topic Relevance are essentially unchanged on Long-MT-Bench+. This indicates that superficially similar but incorrect content can mislead the model.

Necessity of Answer Memory Relative to the non-answer bundle, answer memory remains essential. Without it, agents are more likely to hedge or decline, lowering Informativeness. This suggests GPT-4o is reluctant to fabricate answers even when some relevant context (C/E/D) is present.

We also observe that **irrelevant memory is correlated with reduced constraint awareness** while slightly increasing Memory Coverage and ROUGE-1 Recall. Overall, irrelevant memory does not help. This extends and challenges claims that random/irrelevant documents can improve answer-containing Accuracy on small LLMs (Cuconasu et al., 2024): large LLMs may behave differently, and irrelevant memory may divert attention from contextual constraints.

7.2 Comparison with Parametric Baseline

For the parametric-only baseline, we evaluate six settings: C, A+C, E, D, I, and the all-types bundle. The IR results for GPT-4o are in the bottom table of Tab. 3, with statistical analysis in Fig. 15. Key findings are:

Relevant Memory Helps Accuracy and Relevance The bottom table in Tab. 3 and Fig. 15 show that adding semantically relevant memory (answer, clarifying, enriching, or distracting) generally improves Answer Accuracy, Topic Relevance, and Constraint Relevance over parametric-only baseline. This suggests the necessity of past history when agents generating responses. Otherwise, the responses are seriously off-topic and inaccurate.

Clarifying Memory Remains Critical Among the settings, the A+C and all type bundle settings yield the largest average gains, highlighting the central role of answer-containing evidence and disambiguation information; for GPT-4o, these settings also correspond to higher Actionability. Adding clarifying memory (C) alone also brings significant improvement.

Irrelevant Memory Offers Little Improvement Irrelevant memory provides no clear improvement over parametric-only, yet is still frequently echoed in responses under Memory Coverage, suggesting the agent repeats irrelevant retrieved text without

improving the answer.

RAG Can Reduce Fidelity and Coherence

Across models, adding retrieved memory often lowers World Knowledge Accuracy and Internal Coherence relative to parametric memory alone. Manual analysis suggests additional context encourages more specific claims, which the LLM judge treats as higher factual-error risk, whereas parametric-only responses are shorter, less specific, and safer.

Non-significant Results Some effects are non-significant across agents and datasets, such as whether distracting memory reduces Internal Coherence relative to answer-only, or whether enriching memory affects Topic Relevance. These are small and variable, so stronger evidence is needed to claim reliability.

8 Conclusion and Future Directions

In this work, we examine how memory types affect conversational LLMs using a novel taxonomy, user-centric evaluation, and controlled comparisons. Experiments across agent models, context sizes, and retrievers confirm known patterns and reveal new findings in long-term conversational RAG. Our comparisons show that performance depends not only on retrieving relevant context, but on retrieving the right functional memory types.

These findings point to memory-type-aware retrieval and selection as a promising direction for IR. First, retrieval systems can benefit from memory-role awareness, enabling category-aware curation, ranking, and diversification rather than treating all retrieved context uniformly. Second, our results motivate collections that evaluate different grounding roles in RAG for memory and other corpora. Finally, it is important to develop and validate taxonomies as analytical tools: the most useful taxonomy is not necessarily the most granular, but the one that best explains retrieval behavior and optimize end-to-end outcomes.

Limitations

While our work provides many new findings, it also has several limitations that should be acknowledged.

LLM-as-a-Judge Evaluation Our experiments heavily rely on LLM judges for response scoring and memory type identification. Though we chose advanced models GPT-5-thinking and GPT-o3 as the judges and carefully designed the rubric, LLM judge is known to occasionally produce inconsis-

tent or imperfect assessments. This limitation may introduce uncertainty into the reported results and our analysis.

Computational Cost Our experiments involve large-scale experiments with proprietary LLMs such as GPT models and Gemini. Reproducing these results requires substantial API usage, which may be costly and limit full replication by others.

Dataset Size and Noise The two datasets we used LongMemEval-m and Long-MT-Bench+ contain a few hundred test queries (see their statistics in Table 5). While labeling errors are expected to be rare, any noise in such a small test size can still introduce minor changes in the reported results. Moreover, some of our results are based on very few data points, which limits their statistical significance and should be interpreted with caution.

References

- Google AI and Google DeepMind. 2025. Google gemini. <https://gemini.google.com>.
- Hadel Alhawasi and Abdou Youssef. 2024. Using llms for evaluating qa systems: Exploration and assessment. In *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, pages 462–469. IEEE.
- Chen Amiraz, Florin Cuconasu, Simone Filice, and Zohar Karnin. 2025. [The distracting effect: Understanding irrelevant passages in RAG](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18228–18258. Association for Computational Linguistics.
- Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2021. Open-domain question answering goes conversational via question rewriting. *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.
- Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, and 1 others. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7421–7454.
- Hritik Bansal, John Dang, and Aditya Grover. 2024. Peering through preferences: Unraveling feedback acquisition for aligning large language models. In *The Twelfth International Conference on Learning Representations*.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. [Benchmarking large language models in retrieval-augmented generation](#). In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI Press.
- Nuo Chen, Hongguang Li, Jianhui Chang, Juhua Huang, Baoyuan Wang, and Jia Li. 2025. Compress to impress: Unleashing the potential of compressive memory in real-world long-term conversations. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 755–773. Association for Computational Linguistics.
- 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*.
- Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonello, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.
- Google DeepMind. 2024. Introducing gemini 2.0: our new ai model for the agentic era. Blog post. <https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/>.
- Yijiang River Dong, Tiancheng Hu, and Nigel Collier. 2024. [Can LLM be a personalized judge?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10126–10141. Association for Computational Linguistics.
- 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.
- Wenqi Fan, Yajuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.
- Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2023. *Neural approaches to conversational information retrieval*, volume 44. Springer.
- Mengkang Hu, Tianxing Chen, Qiguang Chen, Yao Mu, Wenqi Shao, and Ping Luo. 2025a. [HiAgent: Hierarchical working memory management for solving long-horizon agent tasks with large language model](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 32779–32798. Association for Computational Linguistics.

- Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025b. [Removal of hallucination on hallucination: Debate-augmented RAG](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15839–15853. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning. *arXiv preprint arXiv:2112.09118*.
- Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2024. Long-context llms meet rag: Overcoming challenges for long inputs in rag. *arXiv preprint arXiv:2410.05983*.
- Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan O Arik. 2025. Long-context LLMs meet RAG: Overcoming challenges for long inputs in RAG. In *The Thirteenth International Conference on Learning Representations*.
- Hao Li, Chenghao Yang, An Zhang, Yang Deng, Xiang Wang, and Tat-Seng Chua. 2025. [Hello again! LLM-powered personalized agent for long-term dialogue](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5259–5276. Association for Computational Linguistics.
- Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew A. Hudson, and 31 others. 2023. [Holistic evaluation of language models](#). *Transactions on Machine Learning Research*.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Junru Lu, Siyu An, Mingbao Lin, Gabriele Pergola, Yulan He, Di Yin, Xing Sun, and Yunsheng Wu. 2023. [Memochat: Tuning llms to use memos for consistent long-range open-domain conversation](#). *Preprint*, arXiv:2308.08239.
- 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. Association for Computational Linguistics.
- Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. [Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico.
- Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Halsey, and 1 others. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.
- OpenAI. 2025a. Chatgpt (gpt-5). <https://chat.openai.com/>. Large language model.
- OpenAI. 2025b. Gpt-5 system card. <https://cdn.openai.com/gpt-5-system-card.pdf>.
- Charles Packer, Sarah Wooders, Kevin Lin, Vivian Fang, Shishir G Patil, Ion Stoica, and Joseph E Gonzalez. 2023. Memgpt: Towards llms as operating systems. *arXiv preprint arXiv:2310.08560*.
- 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. Secom: On memory construction and retrieval for personalized conversational agents. In *The Thirteenth International Conference on Learning Representations*.
- Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 conference on conference human information interaction and retrieval*, pages 117–126.
- Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. 2025. Zep: a temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*.
- Alireza Rezazadeh, Zichao Li, Ange Lou, Yuying Zhao, Wei Wei, and Yujia Bao. 2025. [Collaborative memory: Multi-user memory sharing in llm agents with dynamic access control](#). *Preprint*, arXiv:2505.18279.
- Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Rajan Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025a. [In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents](#). In *Proceedings of the 63rd Annual Meeting of*

- the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8416–8439. Association for Computational Linguistics.
- Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, Long T. Le, Yiwen Song, Yanfei Chen, Hamid Palangi, George Lee, Anand Iyer, Tianlong Chen, Huan Liu, Chen-Yu Lee, and Tomas Pfister. 2025b. [In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents](#). *Preprint*, arXiv:2503.08026.
- Paul Thomas, Seth Spielman, Nick Craswell, and Bhaskar Mitra. 2024. [Large language models can accurately predict searcher preferences](#). In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 1930–1940, New York, NY, USA.
- Bo Wang, Weiyi He, Shenglai Zeng, Zhen Xiang, Yue Xing, Jiliang Tang, and Pengfei He. 2025. [Unveiling privacy risks in LLM agent memory](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25241–25260. Association for Computational Linguistics.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. [Large language models are not fair evaluators](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450. Association for Computational Linguistics.
- Di Wu, Hongwei Wang, Wenhao Yu, Yuwei Zhang, Kai-Wei Chang, and Dong Yu. 2025. Longmemeval: Benchmarking chat assistants on long-term interactive memory. In *The Thirteenth International Conference on Learning Representations*.
- Jing Xu, Arthur Szlam, and Jason Weston. 2022. [Beyond goldfish memory: Long-term open-domain conversation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5180–5197. Association for Computational Linguistics.
- Linhao Ye, Zhikai Lei, Jianghao Yin, Qin Chen, Jie Zhou, and Liang He. 2024. Boosting conversational question answering with fine-grained retrieval-augmentation and self-check. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2301–2305.
- Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and 1 others. 2024. ∞ Bench: Extending long context evaluation beyond 100K tokens. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15262–15277.
- Zheng Zhao, Clara Vania, Subhradeep Kayal, Naila Khan, Shay B Cohen, and Emine Yilmaz. 2025. [PersonaLens: A benchmark for personalization evaluation in conversational AI assistants](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 18023–18055. Association for Computational Linguistics.
- 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.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [Memorybank: Enhancing large language models with long-term memory](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.

Appendix

A Prompt for Response Generation

The full prompts we use for response generation are shown in Figure 6.

system: You are a chat assistant. You previously had a conversation with a user. Now the user is sending you a new message and you need to respond to it. I will provide you with some historical chat pieces with dates between you and the user which you may find helpful to give response. user: Historical Chats: {<i>retrieved pieces</i>} User’s message: {<i>query</i>} Your response:
system: You are a chat assistant. You previously had a conversation with a user. Now the user is sending you a new message and you need to respond to it. I will provide you with some historical chat pieces between you and the user which you may find helpful to give response. user: Historical Chats: {<i>retrieved pieces</i>} User’s message: {<i>query</i>} Your response:
system: You are a chat assistant. Now the user is sending you a new message and you need to respond to it. user: User’s message: {<i>query</i>} Your response:

Figure 6: Prompt for prompting conversational agent to generate response. **(top)** prompt for memory with timestamp and memory piece size $k > 0$; **(middle)** prompt for memory without timestamp and $k > 0$; **(bottom)** prompt for $k = 0$.

B Human Verification on Memory Type

We randomly sampled five queries from LongMemEval-m and five from Long-MT-Bench+ datasets. For each query, we randomly sampled three retrieved memory pieces among the top-15 pieces, whose types are predicted by GPT-5-thinking in our pipeline. In total, we have 30 memory pieces for human evaluation.

We designed a user study questionnaire, mirroring the same instruction and rubric used for the LLM judge, providing example tasks and detailed guidance. Each task is a multiple-choice question to classify the memory into one of the five types. We recruited five human participants from Prolific¹, a third-party data annotation platform, for the study. Our recruitment follows the below criteria:

¹<https://www.prolific.com/>

	GPT-5	H1	H2	H3	H4	H5
GPT-5	–	0.417	0.618	0.512	0.578	0.532
H1	0.417	–	0.426	0.245	0.441	0.573
H2	0.618	0.426	–	0.514	0.615	0.524
H3	0.512	0.245	0.514	–	0.467	0.384
H4	0.578	0.441	0.615	0.467	–	0.538
H5	0.532	0.573	0.524	0.384	0.538	–
Average	0.531	0.420	0.540	0.425	0.528	0.510

Table 4: Pairwise kappa scores among GPT-5-thinking and human participants (H1-H5). GPT-5’s scores suggest it achieves human-level reliability for memory type classification.

- Native English speakers residing in English-speaking countries;
- Holding bachelor degree or above;
- High historical approval rate ($\geq 90\%$);
- Age between 18 to 80.

We rejected participants whose submissions are with unrealistically short completion times and re-recruit. The compensation is at Prolific’s recommended rate.

Using the 30 annotated memory pieces, we calculate the pairwise Cohen’s Kappa k as their inter-annotator agreement, where k is standard unweighted and $k = 0$ corresponds to chance-level agreement. The results are shown in Tab 4. All pairwise k scores between GPT-5-thinking and human annotators are well above chance, and GPT-5-thinking is comparable to human-human agreement.

C Evaluation with LLM-as-a-Judge

We carefully designed and iteratively refined the prompts used for LLM judges to ensure reliable and consistent evaluation. The design process followed an iterative “prompt-test-refine” loop. Starting from a minimal baseline, we tested responses to identify issues such as instruction drift or unstable reasoning. We then applied human-in-the-loop refinement to adjust phrasing and structure, validating each revision on representative cases until the prompts produce desired outputs on a set of human-validated test samples. We provide judgment cases for memory type classification and response evaluation in Table 10, 11, and 12.

We use GPT-5-thinking as the judge for determining memory types, setting its reasoning effort to medium, and max completion tokens to 8,000. Figure 7 is the full prompt for it. For the response evaluation, we use GPT-o3, setting its reasoning effort to high, and max completion tokens to 10,000.

	LongMemEval-m	Long-MT-Bench+
total # conversation	500	11
total # query	500	288
avg. # query/conversation	1	26.2
avg. # session/conversation	501.9	4.9
avg. # utterance/conversation	9.8	26.7
avg. # token/conversation	776, 472.4	14, 000.9

Table 5: Statistics of two long-term conversation datasets for experiment. Token counts are measured as space-separated chunks.

Figure 8 is the prompt for evaluating Accuracy; Figure 9 is for evaluating Relevance; Figure 10 is for evaluating Memory Coverage; Figure 11 is for evaluating Informativeness.

D Memory Relation Illustration

Figure 12 illustrates the semantic-neighborhood view that motivates our taxonomy of conversational memory. It shows how retrieved memory pieces can connect to the user query and to one another, and how different memory roles support or interfere with the agent’s reasoning path.

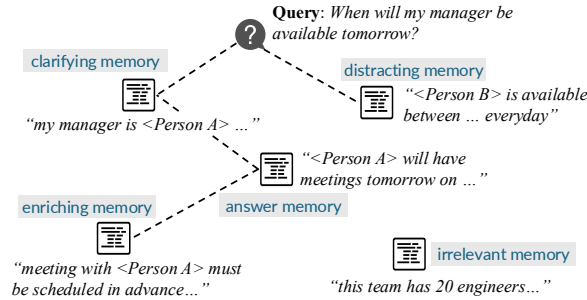


Figure 12: Illustration of relations between memory pieces and user query in agent’s reasoning space. Links stand for semantic connection between two.

E Dataset Statistics

Table 5 summarizes the scale and structure of the two long-term conversational datasets used in our experiments. These statistics contextualize the retrieval and response-generation settings in the main experiments, including the number of conversations, queries, sessions, utterances, and average token counts.

F Memory Type Classification Details

We basically rely on prompt in Figure 7 to determine the memory type. For the experiment on Long-MT-Bench+, since it contains oracle labels for gold pieces, we directly classify the gold pieces as answer memory.

G General Performances on Contriever and BM25

Figure 13 and Figure 14 show the general performance over metrics of three LLMs using Contriever and BM25 as retrieval method.

H Comparative Experiments on Gemini and Llama

Table 6, 7, 8 and 9 show the improvement rate IR results of Gemini-2.0-Flash and Llama-3.3-70B on different comparative experiments. Figure 15 show the statistical results of $L^m(i, j)$ of GPT-4o-2024-08-06 on experiment comparing with parametric memory only. Figure 16, 17, 18, and 19 are statistical results of specific model and comparison setting.

user: You are a quality rater evaluating the relevance of a memory piece to a user's query. The memory piece is from the user's conversation with a chat assistant. Your task is to assign an integer score on a scale from 0 to 4 based on the following criteria:

Scoring rubric (check in order 4→3→2→1→0):

4 - Answer: the memory piece clearly answers the query and itself contains the answer or one of the answers for the question.

3 - Clarifying: the memory piece is necessary to resolve the identity of the asked-for entity/event (e.g., coreference, alias, where/which/company disambiguation) so that an answer stated elsewhere can be correctly applied. Without this piece, you cannot derive a correct or confident answer from the remaining information.

2 - Enriching: the memory piece refers to the same specific entity/event (or uniquely narrows to it) and adds useful ancillary detail (e.g., who/where/what/when) that would make the final answer more informative or helpful, even though it does not provide the asked-for attribute. Domain context that is clearly connected to the query also counts as Helpful.

1 - Distracting: the memory piece seems to reveal the answer but is actually misleading.

0 - Irrelevant: the memory piece has no meaningful semantic connection to the user's query domain.

Now I will provide you with a user's query, one memory piece, and other pieces that can be helpful for you to judge. Your task is to assign a relevance score for this memory piece, following the rubric above.

Note: The score should reflect the memory piece itself. The other pieces are provided only to help you understand the memory piece better.

User's Query:

{*user query*}

Memory Piece:

{*memory piece*}

Other Pieces:

{*other memory pieces*}

Your Score:

Figure 7: Prompt used by the LLM judge to classify retrieved memory pieces.

user: You are an objective and fair evaluator. You are provided with a user's query to a chatbot, the chatbot's response, and the gold response. Your task is to judge the accuracy of the chatbot's response based on three aspects:

1. World knowledge accuracy - How accurate are the factual statements in the response with respect to general facts and world knowledge. A response that does not mention world knowledge is not considered inaccurate.

2. Answer accuracy - How accurate is the response with respect to the answer provided in the gold response.

3. Internal consistency - How logically consistent is the response internally.

You should give **binary scores** for each aspect where 1 means accurate or consistent and 0 means not, and the explanation for the score.

###User's query###

{*question*}

###Chatbot's response###

{*agent response*}

###Gold answer###

{*gold answer*}

You should output only three lines. The first line is the score and explanation for World knowledge accuracy, the second line is the score and explanation for Contextual accuracy, and the third line is the score and explanation for Internal consistency. Each line's format should be 'Score: <score>; Explanation: <explanation>'. No index is needed. Now tell me your scores:

Figure 8: Prompt for LLM judge to determine the Accuracy of responses.

user: You are an objective and fair evaluator. You are provided with a user's query to a chatbot, the user's intent behind the query, the chatbot's response to the query, and the prior conversation history. Your task is to judge the relevance of the chatbot's response based on two aspects:

1. Topical relevance - How directly does the response address the query?
2. Constraint relevance - How well does the response meet all implicit or explicit constraints and conditions stated or implied by the user's intent and prior conversation?

You should give **binary scores** for each aspect where 1 means relevant and 0 means not relevant, and the explanation for the score.

###User's query###
{question}

###User's intent### (optional)
{intent}

###Chatbot's response###
{agent response}

###Prior conversation history###
{context}

You should output only two lines. The first line is the score and explanation for Topical relevance, the second line is the score and explanation for Constraint relevance. Each line's format should be 'Score: <score>; Explanation: <explanation>'. No index is needed. Now tell me your scores:

Figure 9: Prompt for LLM judge to determine the Relevance of responses.

user: You are a helpful and fair judge. I will provide you with:

1. A user's query to a chatbot;
2. The user's intent behind the query;
3. The chatbot's response;
4. The prior conversation history between the user and the chatbot.

Your task is to identify all pieces of information in the chatbot's response that are recalled from the prior conversation history but were **not** restated in the current user query.

- Count each such recalled item.
- Consider paraphrases or indirect references as recalls.
- Ignore generic pleasantries or facts that never appeared in the history.

###Prior conversation history###
{prior conversation history}

###User's query###
{query}

###User's intent###
{intent}

###Chatbot's response###
{agent response}

You should output two lines. The first line is the count, the second line is explanation listing the recalled items and in which piece they are recalled from. Now tell me your score and explanation:

Figure 10: Prompt for LLM judge to determine the Memory Coverage of responses.

user: You are an objective and fair evaluator. You are provided with a user’s query to a chatbot, the user’s intent behind the query, the chatbot’s response to the query, and the prior conversation history. Your task is to judge the informativeness of the chatbot’s response based on two aspects:

1. Completeness - How well does the response cover all sub-questions or necessary information relevant to the query and intent?
2. Actionability - How well does the response provide concrete details or clear guidance?

Particularly, for aspect Completeness, this is a more detailed scale:

Score 0 (Absent): No sign the response noticed the user’s motivations, constraints, or intents. Response is completely generic or contradicts the context.

Score 1 (Partial): The response realizes some of the user’s intent and motivations and mentions at least one contextual element. But it does not fully address the user’s intent.

Score 2 (Comprehensive): The response fully addresses the user’s intent and motivations, providing a thorough and contextually relevant answer.

For aspect Actionability, this is a more detailed scale:

Score 0 (Not actionable): The response is vague, generic, or purely descriptive. Lacks concrete items.

Score 1 (Actionable): The response provides a clear sequence of actions or names specific items.

You should give an integer score from 0 to 2 for aspect Completeness and binary score for aspect Actionability. You should also provide explanation for the score.

###User’s query###

{question}

###User’s intent###

{intent}

###Chatbot’s response###

{agent response}

###Prior conversation history###

{context}

You should output only two lines. The first line is the score and explanation for Completeness, the second line is the score and explanation for Actionability. Each line’s format should be ‘Score: <score>; Explanation: <explanation>’. No index is needed. Now tell me your scores:

Figure 11: Prompt for LLM judge to determine the Informativeness of response.

Comparative Experiment (vs. A)	World Knowledge	Answer Accuracy	Internal Coherence	Topic Relevance	Constraint Awareness	Completeness	Actionability	Memory Coverage	ROUGE-1 Recall
A+C	0.60	–	0.75	0.00	0.67	–	1.00	1.00	0.83
A+E	0.42	0.30	0.53	0.44	0.45	0.00	0.25	0.50	0.59
A+D	0.36	0.60	0.47	0.40	0.71	1.00	0.60	1.00	0.42
A+I	0.43	0.64	0.40	0.41	0.60	0.57	0.30	0.31	0.30
C+E+D+I	0.27	0.14	0.05	0.04	0.55	0.55	0.05	0.01	0.06

Table 6: Gemini-2.0-Flash’s results of IR of experimental setting (index column) vs. controlled setting (answer memory only) on LongMemEval-m (left number in each cell) and Long-MT-Bench+ (right number in each cell).

Comparative Experiment (vs. A)	World Knowledge	Answer Accuracy	Internal Coherence	Topic Relevance	Constraint Awareness	Completeness	Actionability	Memory Coverage	ROUGE-1 Recall
A+C	0.25	0.00	0.60	1.00	1.00	0.00	–	1.00	0.50
A+E	0.53	0.29	0.47	0.73	0.38	0.00	0.83	0.67	0.48
A+D	0.36	0.50	0.39	0.50	0.64	0.50	0.00	0.50	0.52
A+I	0.46	0.50	0.36	0.61	0.56	0.56	0.19	0.00	0.30
C+E+D+I	0.24	0.08	0.04	0.06	0.44	0.57	0.02	0.00	0.02

Table 7: Llama-3.3-70B’s results of IR of experimental setting (index column) vs. controlled setting (answer memory only) on LongMemEval-m (left number in each cell) and Long-MT-Bench+ (right number in each cell).

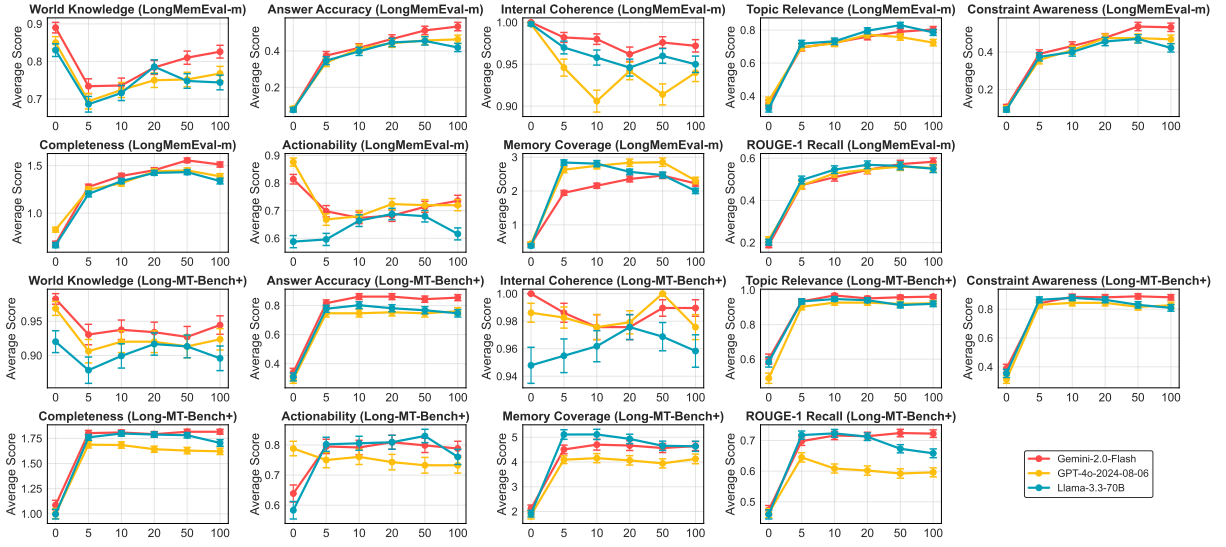


Figure 13: Average scores over different metrics of three agent models on LongMemEval-m and Long-MT-Bench+ datasets. Results are based on Contriever. Average scores are calculated over queries.

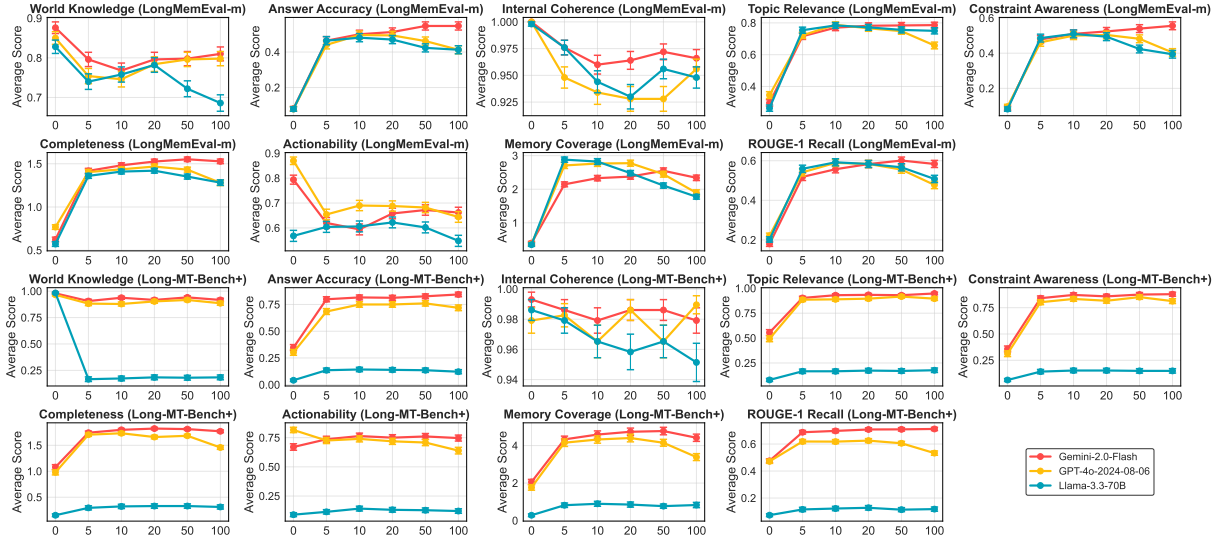


Figure 14: Average scores over different metrics of three agent models on LongMemEval-m and Long-MT-Bench+ datasets. Results are based on BM25 retriever. Average scores are calculated over queries.

Comparative Experiment (vs. P)	World Knowledge		Answer Accuracy		Internal Coherence		Topic Relevance		Constraint Awareness		Completeness		Actionability		Memory Coverage		ROUGE-1 Recall	
C	0.17	0.12	0.00	0.91	0.00	0.00	0.88	0.92	—	0.88	0.42	0.81	0.25	0.22	0.87	0.87	0.58	0.76
A+C	0.57	0.00	0.89	0.75	0.00	—	1.00	1.00	1.00	0.80	0.93	0.83	0.40	0.00	0.88	0.86	0.82	1.00
E	0.20	0.08	0.76	0.82	0.00	0.40	0.84	0.87	0.72	0.88	0.62	0.75	0.20	0.37	0.87	0.84	0.63	0.69
D	0.23	0.00	0.69	0.62	0.00	0.17	0.76	0.73	0.68	0.62	0.48	0.63	0.16	0.35	0.89	0.81	0.59	0.45
I	0.23	0.06	0.50	0.39	0.00	0.50	0.68	0.35	0.44	0.43	0.41	0.22	0.12	0.18	0.88	0.58	0.60	0.29
A+C+E+D+I	0.38	0.28	0.95	0.92	0.00	0.50	0.93	0.96	0.95	0.96	0.92	0.78	0.29	0.47	0.95	0.81	0.89	0.67

Table 8: Gemini-2.0-Flash’s results of IR of experimental setting (index column) vs. controlled setting (parametric memory only) on LongMemEval-m (left number in each cell) and Long-MT-Bench+ (right number in each cell).

Comparative Experiment (vs. P)	World Knowledge		Answer Accuracy		Internal Coherence		Topic Relevance		Constraint Awareness		Completeness			Actionability	Memory Coverage		ROUGE-1 Recall	
C	0.25	0.00	–	0.95	0.00	0.00	0.89	0.95	1.00	0.96	0.91	0.93	0.60	0.50	0.95	0.93	1.00	0.90
A+C	0.50	0.00	0.88	1.00	–	0.00	0.92	1.00	1.00	1.00	1.00	1.00	0.89	1.00	0.94	1.00	1.00	1.00
E	0.22	0.25	0.82	0.83	0.00	0.70	0.90	0.88	0.79	0.81	0.79	0.76	0.44	0.61	0.96	0.87	0.79	0.83
D	0.08	0.08	0.53	0.67	0.00	0.43	0.84	0.75	0.57	0.57	0.62	0.77	0.26	0.50	0.96	0.76	0.70	0.73
I	0.25	0.13	0.33	0.44	0.06	0.65	0.74	0.47	0.38	0.37	0.50	0.39	0.22	0.29	0.93	0.75	0.71	0.56
A+C+E+D+I	0.45	0.47	0.96	0.98	0.03	0.71	0.96	0.98	0.97	0.95	0.94	0.91	0.59	0.74	0.95	0.89	0.94	0.88

Table 9: Llama-3. 3-70B’s results of IR of experimental setting (index column) vs. controlled setting (parametric memory only) on LongMemEval-m (left number in each cell) and Long-MT-Bench+ (right number in each cell).

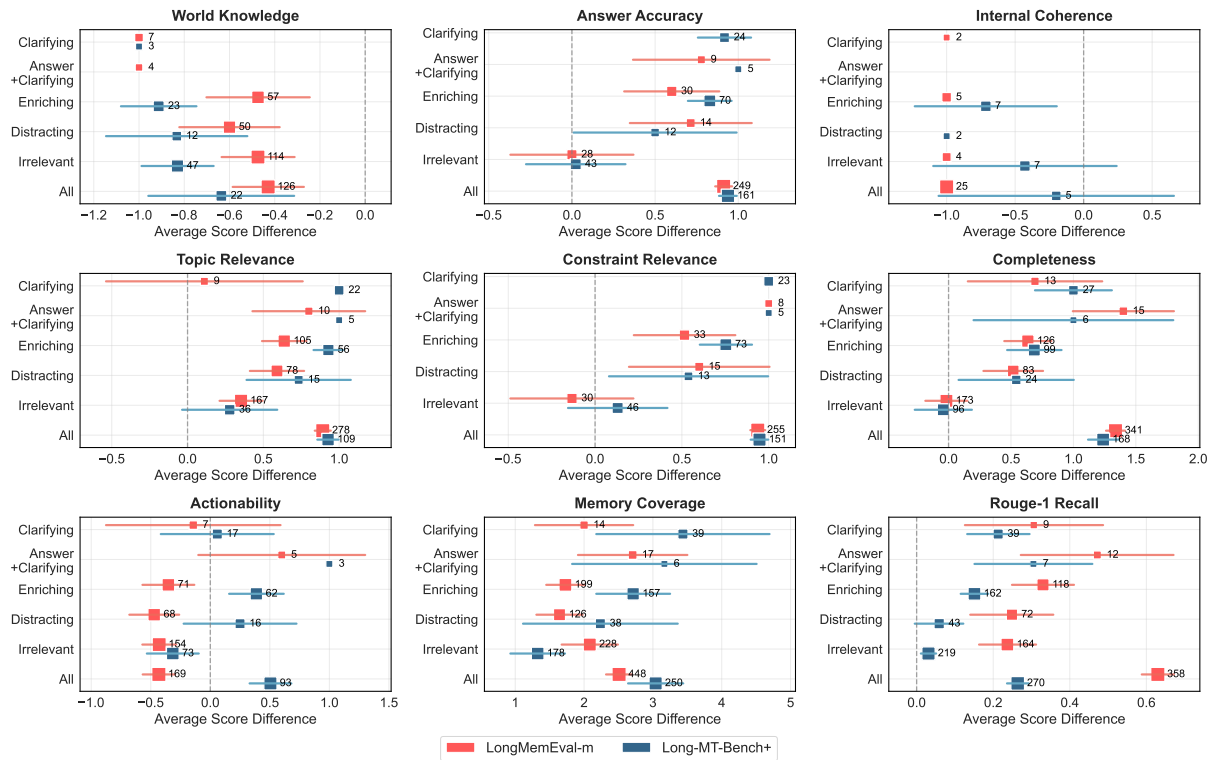


Figure 15: Results of mean per-query score differences with confident intervals when adding a memory setting relative to the **parametric-only baseline** on GPT-4o. Squares mark the mean difference; horizontal bars show the confidence interval. Square size is proportional to the number of queries with a non-zero change, and the number shown beside each square equals that count.

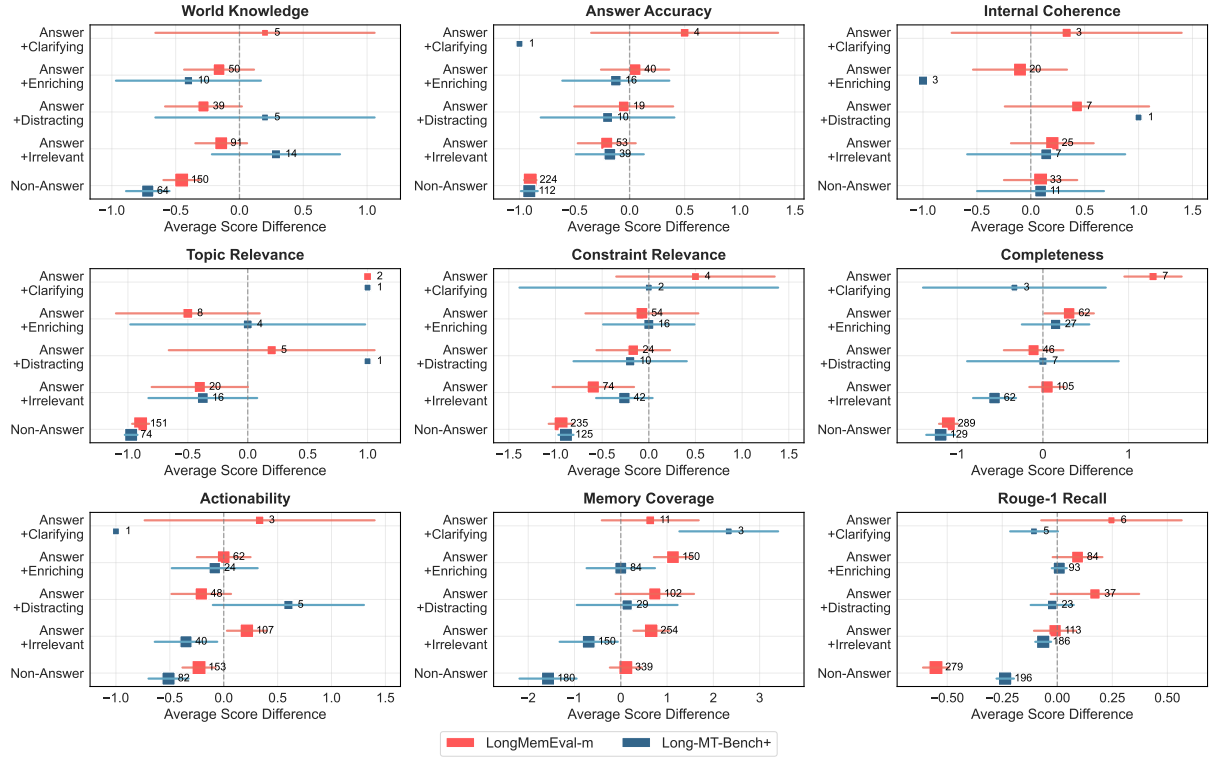


Figure 16: Results of mean per-query score differences with confident intervals when adding a memory setting relative to the **answer-only baseline** on Gemini-2.0-Flash. Squares mark the mean difference; horizontal bars show the confidence interval. Square size is proportional to the number of queries with a non-zero change, and the number shown beside each square equals that count.

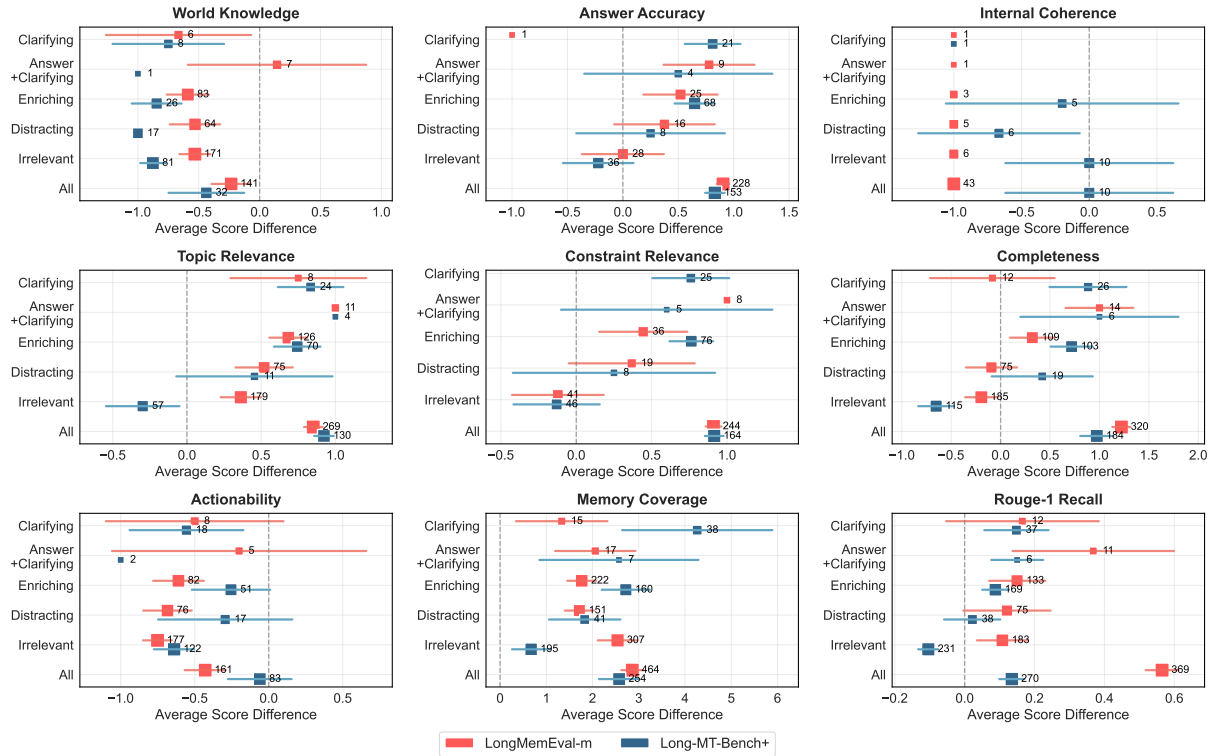


Figure 17: Results of mean per-query score differences with confident intervals when adding a memory setting relative to the **parametric-only baseline** on Gemini-2.0-Flash. Squares mark the mean difference; horizontal bars show the confidence interval. Square size is proportional to the number of queries with a non-zero change, and the number shown beside each square equals that count.

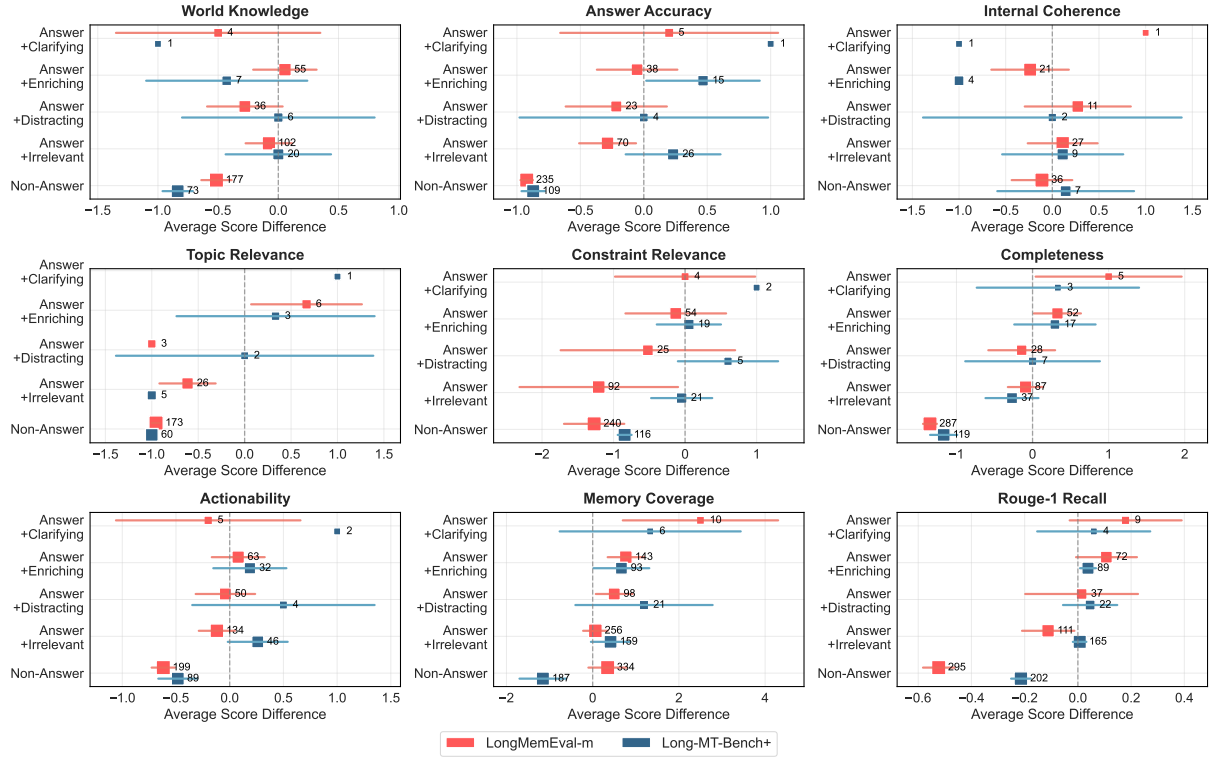


Figure 18: Results of mean per-query score differences with confident intervals when adding a memory setting relative to the **answer-only baseline** on Llama-3. 3-70B. Squares mark the mean difference; horizontal bars show the confidence interval. Square size is proportional to the number of queries with a non-zero change, and the number shown beside each square equals that count.

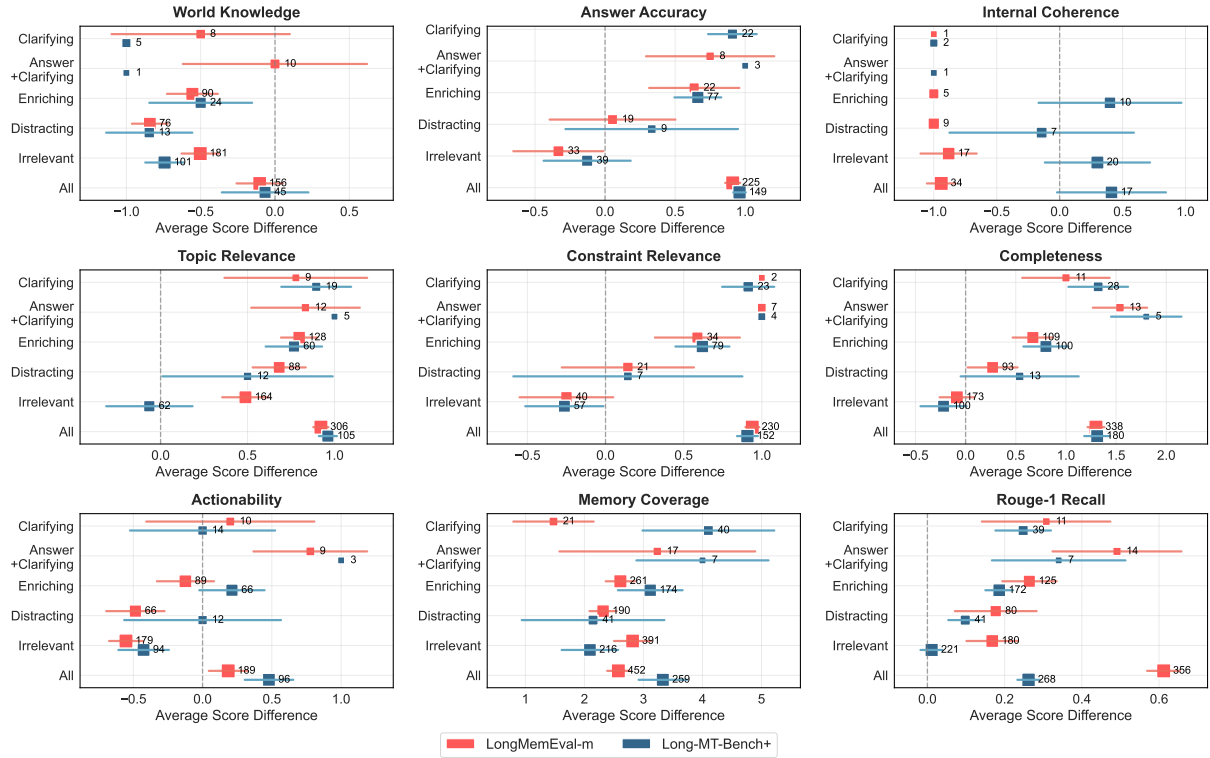


Figure 19: Results of mean per-query score differences with confident intervals when adding a memory setting relative to the **parametric-only baseline** on Llama-3. 3-70B. Squares mark the mean difference; horizontal bars show the confidence interval. Square size is proportional to the number of queries with a non-zero change, and the number shown beside each square equals that count.

Instruction: ... User query: I'm planning to visit Bandung again and I was wondering if you could remind me of the name of that restaurant in Cihampelas Walk that serves a great Nasi Goreng? Gold answer: Miss Bee Providore	
Memory piece: [User]: I think I'll start with Cihampelas Walk to check out the denim street. Can you recommend any good restaurants or cafes in the area? [Assistant]: Sure! There are many good restaurants and cafes in the Cihampelas Walk area. Here are a few options: 1. Miss Bee Providore: This restaurant serves a mix of western and Indonesian cuisine and has a cozy and stylish interior. 2. Bakery & Co: A great place to relax and enjoy a cup of coffee and a pastry. 3. Raa Cha Suki & BBQ: If you're looking for something spicy and delicious, this Thai restaurant is a great option. 4. Sumoboo: This Japanese shaved ice dessert place features a wide variety of flavors, toppings, and combinations. 5. Holycow! Steakhouse by Chef Afrit: This steakhouse has a warm atmosphere and serves high-quality steaks at reasonable prices. These are just a few of the many options you can explore in the area. Make sure to do some research and read reviews before making a final decision. Enjoy your visit!	Enriching
Memory piece: [User]: I'm definitely going to try out Miss Bee Providore for some delicious food. Do they have any signature dishes I should try? [Assistant]: Miss Bee Providore offers a variety of delicious dishes, both Western and Indonesian cuisine. Here are a few signature dishes you should definitely try: 1. Miss Bee's Nasi Goreng: Their take on the classic Indonesian fried rice dish is a must-try! It's served with a side of prawn crackers and topped with a fried egg. 2. Yoghurt Granola Bowl: This is a great option if you're looking for something healthy and refreshing. It's made with homemade granola, fresh fruit, and yogurt. 3. Fried Chicken Burger: This signature burger features a juicy chicken patty, crispy bacon, lettuce, tomato, and a special aioli sauce. It's served with a side of fries. 4. Beef Rendang: If you're looking for a traditional Indonesian dish, try their version of beef rendang. It's a slow-cooked beef dish in a flavorful, spicy coconut milk sauce. 5. Chocolate Brownie: Finish off your meal with a sweet treat - Miss Bee's chocolate brownie is a local favorite and is served warm with vanilla ice cream and chocolate sauce. These are just a few of the many delicious options, so feel free to explore their menu and try something new!	Answer
Memory piece: [User]: Yum, all of those dishes sound amazing! I'm definitely going to try the Miss Bee's Nasi Goreng and finish it off with the chocolate brownie. [Assistant]: Enjoy your meal at Miss Bee Providore and have a great time shopping at Cihampelas Walk!	Answer
Memory piece: [User]: I'm so excited to explore more of Bandung! Do you have any recommendations for other fun activities to do in the city? [Assistant]: Sure, Bandung has plenty of fun activities to offer! Here are a few recommendations: 1. Visit a volcano: Bandung is located near several active and dormant volcanoes, including Tangkuban Perahu and Kawah Putih. You can take a day trip to visit these fascinating natural wonders. 2. Explore the tea plantations: Bandung is also known for its tea plantations, such as the Malabar Tea Plantation and the Puncak Tea Garden. You can take a tour of these plantations, enjoy a cup of tea, and enjoy the beautiful scenery. 3. Visit the Trans Studio Bandung: This is one of Indonesia's largest indoor theme parks, with rides, attractions, and entertainment for all ages. 4. Go shopping at the Paris Van Java mall: This mall is known for its luxurious and trendy boutiques, restaurants, and cafes, as well as its modern architecture. 5. Take a cooking class: Bandung is famous for its cuisine, and you can take a cooking class to learn how to make traditional Indonesian dishes such as nasi goreng, sate, and rendang. 6. Visit the Saung Angklung Udjo: This is a cultural center that features traditional music and dance performances with the angklung, a traditional Indonesian instrument made from bamboo. These are just a few of the many activities that you can do in Bandung. Make sure to do some research and plan your trip accordingly, depending on your interests and preferences. Have fun exploring!	Irrelevant
Memory piece: [User]: What are some unique shopping experiences to be had in Bandung? [Assistant]: As an AI language model, I don't have personal experiences, but based on research, here are some unique shopping experiences to be had in Bandung: 1. Factory outlets: Bandung is well known for being a factory outlet (FO) paradise. You can find many factory outlets selling local and international brands at discounted prices. 2. Pasar Baru: This is Bandung's oldest and largest traditional market, selling everything from clothing and accessories to food and home goods. It's a great place to experience the hustle and bustle of a traditional market. 3. Braga Street: This historic street is lined with old colonial buildings that have been turned into trendy cafes, restaurants, and boutique shops. It's a great place to find unique souvenirs and gifts. 4. Cihampelas Walk: Also known as Ciwalk, this shopping center is famous for its denim street, where you can find a wide range of denim products, from jeans to jackets. 5. Dago Street: This street is home to many boutiques, art galleries, and cafes, making it a popular spot for young and trendy shoppers. 6. Rumah Mode: This popular factory outlet is famous for its Balinese-inspired architecture and stylish clothing and accessories. 7. Kartika Sari: This bakery is famous for its chocolate brownies and other Indonesian snacks, making it a great place to pick up some tasty treats to take home.	Irrelevant

Table 10: A judgment case for GPT-5-reasoning to determine the memory type. Left column is each memory piece and right column is the determined type. See the full prompt in Figure 7.

Instruction: ... User query: I'm planning to revisit Orlando. I was wondering if you could remind me of that unique dessert shop with the giant milkshakes we talked about last time? Gold answer: The Sugar Factory at Icon Park.	
Memory piece: [User]: Wow, these dining options sound amazing! Do you have any recommendations for a fun dessert spot that my family can check out after dinner? [Assistant]: Absolutely! Here are some fun dessert spots that your family might enjoy after dinner: 1. The Sugar Factory - A sweet shop located at Icon Park that offers an enormous menu of sweet treats, including specialty drinks and giant milkshakes. 2. Wondermade - A gourmet marshmallow shop located in Sanford, just north of Orlando, that offers unique flavored marshmallows and s'mores kits. 3. Gideon's Bakehouse - A bakery located at Disney Springs that offers delicious cookies, cakes, and other baked treats. 4. Kelly's Homemade Ice Cream - A small-batch ice cream shop located in Orlando that offers unique flavors like Banana Cream Pie and Strawberry Jalapeno. 5. Voodoo Doughnut - A doughnut shop located at Universal CityWalk that offers imaginative and creative doughnuts with unique toppings and flavors. 6. The Glass Knife - A dessert and coffee shop located in Winter Park that serves up beautiful cakes, pastries, and other sweet treats in a stylish setting. 7. Lickety Splitz - A dessert shop located in Orlando that serves up handmade frozen yogurt, sorbet, and ice cream. 8. Better Than Sex - A dessert shop located in the Ivanhoe Village neighborhood of Orlando that specializes in decadent desserts like chocolate fondue and cheesecake. 9. The Greenery Creamery - An ice cream shop located in the Thornton Park neighborhood of Orlando that offers all-natural ice cream flavors like lavender, honeycomb, and matcha. 10. The Pastry Studio - A bakery located in Daytona Beach, just east of Orlando, that offers beautiful and delicious custom cakes and other desserts.	Answer
Memory piece: [User]: Can you recommend any places to eat in Orlando that are family-friendly? [Assistant]: Certainly! Orlando has many family-friendly dining options. Here are some recommendations: 1. Toothsome Chocolate Emporium & Savory Feast Kitchen - A restaurant located at Universal CityWalk that offers a unique dining experience with imaginative dishes and delicious desserts. 2. The Cheesecake Factory - A casual dining restaurant located in The Mall at Millenia that offers a diverse menu that includes everything from burgers to salads and, of course, an extensive selection of cheesecakes. 3. Shake Shack - A fast-food restaurant located in Winter Park that offers burgers, hot dogs, and shakes. 4. Hash House A Go Go - A restaurant located on International Drive that serves up twisted farm food that is both delicious and visually stunning. 5. Bahama Breeze - A Caribbean-themed restaurant located on International Drive that has a family-friendly atmosphere and offers a menu of tropical dishes and drinks. 6. Blaze Pizza - A fast-casual restaurant located in Disney Springs that offers build-your-own Neapolitan pizzas and salads. 7. T-Rex Cafe - A dinosaur-themed restaurant located in Disney Springs that offers a unique prehistoric dining experience for families. 8. Uncle Julio's - A Tex-Mex restaurant located in Orlando and Kissimmee that offers a menu of authentic Mexican dishes and delicious drinks. 9. Pio Pio - A restaurant located in the Dr. Phillips neighborhood that serves up delicious Peruvian dishes. 10. Sweet Tomatoes - A buffet restaurant chain located throughout Orlando that offers a variety of healthy and delicious salads, soups, and baked goods.	Distracting
Memory piece: [User]: Wow, those tips are great! I'm excited to take on this dessert adventure with my family. Do you have any more recommendations for other fun and unique things to do in Orlando? [Assistant]: Yes, there are plenty of other fun and unique things to do in Orlando! Here are some more recommendations: 1. Escape Rooms - Orlando has several escape rooms that offer immersive and challenging puzzle-solving experiences. 2. Cirque du Soleil - Watch a mesmerizing acrobatic and theatrical performance by the world-famous Cirque du Soleil, which often performs in Orlando. 3. Airboat Tours - Take an airboat tour through the Florida Everglades and see alligators, birds, and other wildlife up close. 4. The Orlando Eye - Take a ride on the 400-foot observation wheel and enjoy panoramic views of Orlando. 5. Orlando Science Center - Visit the Orlando Science Center, which has interactive exhibits and educational programs that entertain visitors of all ages. 6. The Orlando Museum of Art - See beautiful artwork from various periods and artists at the Orlando Museum of Art. 7. Amusement Parks - Orlando has many amusement parks besides Disney and Universal, such as Fun Spot America, ICON Park, and SeaWorld. 8. The Charles Hosmer Morse Museum of American Art - See beautiful works of art by Louis Comfort Tiffany, including iconic stained glass lamps. 9. Dr. Phillips Center for the Performing Arts - Enjoy a variety of dance, music and theater performances at this beautiful venue. 10. The Orlando Shakespeare Theater - Experience the theater with classic and contemporary works in a unique setting.	Irrelevant
Memory piece: [User]: I don't really trust buffet restaurants. Can you suggest any other family-friendly dining options in Orlando? [Assistant]: Sure, here are some additional family-friendly dining options in Orlando: 1. The Cowfish Sushi Burger Bar - A unique restaurant located at Universal CityWalk that specializes in creative sushi and gourmet burgers. 2. BJ's Restaurant and Brewhouse - A casual dining restaurant located in Orlando that offers a diverse menu of American style dishes and beers brewed on site. 3. Mediterranean Blue - A Greek restaurant located in Orlando that serves up classic Greek dishes like souvlaki and gyros. 4. Black Bean Deli - A casual Cuban restaurant located in Orlando that serves up delicious Cuban sandwiches, plantains, and more. 5. Toothsome Chocolate Emporium & Savory Feast Kitchen - As mentioned before, this is a great family-friendly option for lunch or dinner. 6. Dandelion Communita Cafe - A vegetarian and vegan restaurant located in Orlando that serves up fresh and healthy salads, sandwiches, and smoothies. 7. The Whiskey - A restaurant located in Orlando that has a family-friendly outdoor seating area and offers a menu of American dishes with a Southern twist. 8. Blaze Pizza - Another option mentioned before, this fast-casual pizzeria is great for a quick meal. 9. Kobe Japanese Steakhouse - A Japanese hibachi grill restaurant located in Orlando that offers an entertaining dining experience and delicious seafood, meat, and vegetable dishes. 10. Soco Restaurant - A restaurant located in the Thornton Park neighborhood of Orlando that offers Southern-style dishes with a modern twist. They also have a great weekend brunch.	Distracting

Table 11: A judgment case for GPT-5-reasoning to determine the memory type. Left column is each memory piece and right column is the determined type. See the full prompt in Figure 7.

Instruction: ... Prior conversation history: ... User query: I wanted to check back on our previous conversation about Netflix. I mentioned that I wanted to be able to access all seasons of old shows? Do you remember what show I used as an example, the one that only had the last season available? Gold answer: Doc Martin.						
Answer ($k = 0$): I'm sorry, I don't have the ability to recall past conversations. However, if you let me know the name of the show again, I can try to help you with any information or updates regarding its availability on Netflix.						
World Knowledge	Answer Accuracy	Internal Coherence	Topic Relevance	Constraint Awareness	Completeness	Actionability
1	0	1	1	0	0	0
Answer ($k = 5$): Yes, I remember! You mentioned the show "Doc Martin" as an example where only the last season was available on Netflix. If you have any more questions or need further assistance, feel free to ask!						
1	1	1	1	1	2	0
Answer ($k = 10$): Yes, I remember! You mentioned that you wanted to have access to all seasons of old shows on Netflix, and you used "Doc Martin" as an example, noting that only the last season was available.						
1	1	1	1	1	2	0
Answer ($k = 20$): Yes, I remember! You mentioned that you wanted to have access to all seasons of old shows and used "Doc Martin" as an example, noting that only the last season was available.						
1	1	1	1	1	2	0
Answer ($k = 50$): Yes, you mentioned that you wanted to be able to access all seasons of old shows on Netflix, and you used "Doc Martin" as an example, noting that it only had the last season available. If you have any more questions or need further assistance, feel free to ask!						
1	1	1	1	1	2	0
Answer ($k = 100$): Yes, I remember our conversation about Netflix. You mentioned that you wanted to be able to access all seasons of old shows and used "Doc Martin" as an example, noting that it went down after a while, and now only the last season is available. If you have any more questions or need further assistance, feel free to ask!						
1	1	1	1	1	2	0

Table 12: A judgment case for GPT-3o to score the response on different metrics. We provide responses under different retrieval size (k) and their scores.