UTILMEM: Benchmarking Evidence Utilization in Long-Term Conversational Memory

Peijun Qing[1] **Fobo Shi**[2] **Soroush Vosoughi**[1]
[1]Dartmouth College [2]Wuhan University
peijun.qing.gr@dartmouth.edu

Abstract

Long-term memory is central to conversational agents, yet existing benchmarks largely evaluate systems through pointwise factual recall against atomic gold answers, assessing whether systems can recover specific facts or event-level details from past conversations and answer localized questions. However, real-world memory use often requires integrating distributed, implicit, and noisy evidence across long interaction histories into coherent task-oriented outputs. We introduce UTILMEM, a 1,717-instance, five-domain diagnostic benchmark designed to evaluate this requirement, which we term memory utilization. UTILMEM targets four challenges that remain underexplored in existing memory benchmarks: dense multi-session reasoning, retrieval of implicitly relevant memory, evidence composition into summaries, analyses, and plans, and robustness to semantically similar distractors. Across a range of retrieval-based and memory-augmented systems, we find that strong performance on existing factual memory benchmarks does not reliably transfer to strong utilization performance. Even when relevant evidence is successfully retrieved, systems often fail to integrate information across sessions and separate relevant evidence from distractors. Our results reveal a substantial gap between memory retrieval and effective memory utilization in current long-term memory systems, highlighting the need for architectures that support robust evidence integration under retrieval interference. Our code is available at https://github.com/peijunallin/UtilMem.

1 Introduction

Large language model (LLM) assistants are increasingly deployed in long-horizon, multi-session settings such as personal companions, tutors, financial advisors, and health coaches, where users expect the system to remember past interactions and act on them (Zhong et al., 2024; Packer et al., 2023; Chhikara et al., 2025; Qing et al., 2026b). This regime has motivated a wave of memory-augmented architectures that compress, index, and retrieve from chat histories (Packer et al., 2023; Xu et al., 2025; Gutiérrez et al., 2024; Wang and Chen, 2025; Diao et al., 2025; Qing et al., 2026a), alongside benchmarks that aim to measure how well these systems support long-term interaction (Maharana et al., 2024; Wu et al., 2025; Hu et al., 2026b; Tan et al., 2025; Bian et al., 2026). Yet what existing benchmarks measure is substantially narrower than what deployed memory systems are asked to do.

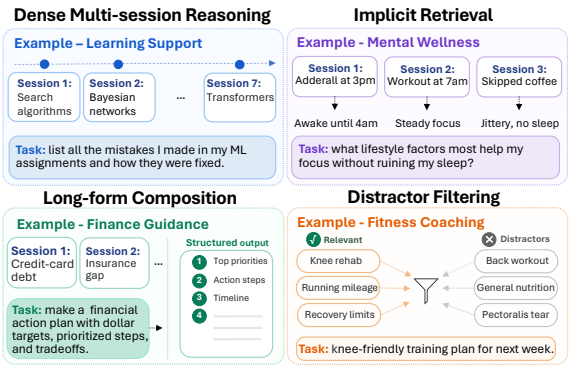

Figure 1: Memory utilization tasks in UTILMEM.

Existing long-term memory benchmarks are primarily evaluated through pointwise factual recall against atomic gold answers. Benchmarks such as LoCoMo (Maharana et al., 2024) and LongMemEval (Wu et al., 2025) primarily evaluate whether a system can retrieve a relevant span from conversation history and answer a narrowly scoped question. However, real memory usage is rarely a retrieval-only problem. Current evaluations leave several critical aspects underconstrained: **(1) localized evidence**, where supporting information is concentrated within a small number of sessions and long-horizon reasoning collapses into local retrieval; **(2) limited distractor pressure**, where distractors are weak or topically distant, allow-

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 37249–37272
October 24-29, 2026 ©2026 Association for Computational Linguistics

ing systems to improve performance simply by increasing top-k retrieval; **(3) short-form outputs**, where tasks require only atomic factual responses rather than coherent long-form synthesis; and **(4) shortcut-friendly evaluation**, where strong benchmark performance can be achieved through event-level retrieval, atomic fact extraction, or aggressive memory compression that converts raw interaction histories into compact summaries or structured slots. As a result, these limitations undermeasure an essential goal of long-term memory that we operationalize as *memory utilization*: the ability to synthesize distributed, implicit, and noisy evidence from long interaction histories into coherent task-oriented responses.

To systematically evaluate this requirement, we introduce UTILMEM, which targets four challenges (see Figure 1) largely absent from recall-oriented evaluations: **Dense Multi-session Reasoning**, where evidence is distributed across temporally separated sessions and topics; **Implicit Retrieval**, where systems recover latent user preferences, behavioral patterns, and evolving states rather than explicit facts; **Long-form Composition**, where fragmented and partially redundant evidence must be organized into coherent outputs; and **Distractor Filtering**, where semantically similar but functionally irrelevant sessions create retrieval and generation interference.

UTILMEM is built from realistic multi-session interaction trajectories spanning learning support, finance guidance, mental wellness, fitness coaching, and document analysis. We generate evidence-dependent tasks requiring cross-session integration and apply counterfactual filtering to retain only instances whose answer quality degrades along the sampled sequential evidence-removal path. We further inject adversarial distractors that overlap lexically and semantically with the evidence while differing in intent, forcing systems to separate grounded memory from plausible interference.

We evaluate retrieval-augmented and memory-augmented systems on UTILMEM. Results reveal a consistent gap between memory access and memory utilization. Systems performing strongly on factual memory benchmarks degrade substantially when tasks require cross-session reasoning under retrieval noise. Even when relevant evidence is retrieved, models often fail to compose coherent grounded outputs or instead rely on semantically plausible distractors that produce confident but unsupported generations. For the evaluated systems and configurations, these findings diagnose failure modes that recall-oriented evaluation does not expose; they do not establish a universal mechanism for all long-term memory systems. Our contributions are summarized as follows:

- A benchmark for evaluating memory utilization in long-term conversational agents.
- A scalable synthesis pipeline that produces evidence-sensitive evaluation instances with adversarial retrieval interference.
- An empirical evaluation showing that strong performance on existing factual memory benchmarks does not guarantee robust memory utilization.

2 Related Work

2.1 Benchmarks for Long-Term Conversational Memory

Long-term memory benchmarks have largely evolved within a single evaluation paradigm: factual question answering over multi-session dialogue histories. Early work established the format with multi-day chat probes (Du et al., 2024; Kim et al., 2024; Zhong et al., 2024) and human–human conversations spanning single-hop, multi-hop, temporal, and adversarial questions (Maharana et al., 2024). LongMemEval (Wu et al., 2025) consolidated this protocol around five recall-oriented abilities for user–assistant interactions, namely information extraction, multi-session reasoning, temporal reasoning, knowledge updates, and abstention. Recent extensions diversify the protocol along several axes: broader question taxonomies that include reflective memory and observer scenarios (Tan et al., 2025), operation-level decomposition that localizes failures within the memory pipeline (Chen et al., 2025), finer cognitive competencies under incremental multi-turn formats (Hu et al., 2026b), project-oriented interactions with interwoven queries (Bian et al., 2026), and large-scale conversational corpora with implicit cues (Pakhomov et al., 2025).

Despite this breadth, the underlying evaluation signal remains pointwise factual recall against atomic gold answers, with evidence concentrated in one or two sessions and limited adversarial interference. As summarized in Table 1, none of the benchmarks compared jointly covers the four utilization

Benchmark	#Q	Context Depth	Eval. Metric	Memory Utilization Dimensions			
				D1: Dense Multi-session	D2: Implicit Retrieval	D3: Long-form Composition	D4: Distractor Filtering
LoCoMo (Maharana et al., 2024)	7,512	∼9k	F1 / BLEU / ROUGE	✓	✗	•	✗
LongMemEval (Wu et al., 2025)	500	115k–1.5M	LLM-judge accuracy	•	✗	✗	✗
MemoryBank (Zhong et al., 2024)	194	∼5k	Manual + LLM judge	✗	✗	✗	✗
MemBench (Tan et al., 2025)	1,039	10k–100k	MCQ Acc. / Recall@10	✓	✓	✗	•
MemoryAgentBench (Hu et al., 2026b)	2,071	103k–1.44M	LLM-judge accuracy	✓	•	✗	✗
RealMem (Bian et al., 2026)	1,415	∼269k	QA Score / Recall@K / NDCG	✓	✓	✗	✗
ConvoMem (Pakhomov et al., 2025)	75,336	1k–3M	LLM-judge accuracy	•	✓	✗	✗
UtilMem (Ours)	**1,717**	**80k–220k**	**Rubric LLM-judge**	✓	✓	✓	✓

Table 1: Comparison of long-term conversational memory benchmarks. ✓: fully supported; •: partially supported; ✗: not supported. Among the benchmarks compared here, UtilMem is the only one that jointly covers all four utilization dimensions.

dimensions we identify: dense multi-session evidence, implicit retrieval, long-form composition, and distractor filtering. UtilMem targets this gap by evaluating memory utilization: long-form composition over distributed, implicit evidence under semantically similar distractors. Noise sensitivity is also studied in retrieval-augmented generation. Prior work shows that irrelevant context can distract language models (Shi et al., 2023), that useful evidence can be underused depending on its position in a long context (Liu et al., 2024), and that RAG systems vary in noise robustness, negative rejection, information integration, and counterfactual robustness (Chen et al., 2024). UtilMem extends this diagnostic setting to long-term conversational memory: its distractors are complete conversation sessions that deliberately share lexical and semantic cues with the evidence while remaining functionally irrelevant to the user's request, and the output requires synthesis across multiple sessions rather than passage-level extraction.

2.2 Memory Systems for Long-Term Interaction

A parallel line of work develops explicit memory systems that compress, organize, and retrieve from interaction histories (Wang et al., 2024; Zhang et al., 2024; Du et al., 2025; Liu et al., 2025; Mei et al., 2025). Existing designs cluster into three directions. The first stores experiences as linear or hierarchical streams with policies for retention and forgetting (Packer et al., 2023; Zhong et al., 2024; Liang et al., 2023; Park et al., 2023). The second imposes structure through graphs, trees, or temporal indices to support compositional retrieval and update (Xu et al., 2025; Chhikara et al., 2025; Rasmussen et al., 2025; Rezazadeh et al., 2025). The third integrates heterogeneous memory types, including parametric, activation, procedural, and multi-agent memory, under unified lifecycle control (Li et al., 2025; Kang et al., 2025). Retrieval-augmented generation and its graph-augmented variants remain widely used as memory backends (Lewis et al., 2020; Edge et al., 2024; Gutiérrez et al., 2024; Guo et al., 2024).

These systems target distinct facets of memory, including storage efficiency, update consistency, relational structure, and retrieval precision. Their reported gains, however, are often evaluated with recall-oriented QA signals, leaving open whether architectural progress translates to tasks that require evidence integration rather than fact lookup. UtilMem provides such a testbed.

3 Benchmark Construction

This section starts with the formulation of our task. As illustrated in Figure 2, we first synthesize realistic multi-session interaction histories (§3.1), then generate evidence-dependent questions (§3.2), filter for evidence sensitivity (§3.3), and finally inject adversarial distractors to construct retrieval-interfering haystacks (§3.4).

Problem Formulation A benchmark instance is a tuple $(\mathcal{H}, q, \mathcal{E})$, defined as follows. Let $\mathcal{H} = \{s_1, s_2, \ldots, s_{|\mathcal{H}|}\}$ denote a user's complete interaction history, consisting of $|\mathcal{H}|$ chat sessions. Each session s_i is a multi-turn dialogue between a user and an assistant:

$$s_i = \left[(u_i^{(1)}, a_i^{(1)}), \ldots, (u_i^{(T_i)}, a_i^{(T_i)})\right], \quad (1)$$

where $(u_i^{(t)}, a_i^{(t)})$ is the t-th turn consisting of a user utterance and an assistant response, and T_i is the number of turns in session s_i. The history $\mathcal{H}$ is partitioned into two disjoint subsets:

$$\mathcal{H} = \mathcal{E} \cup \mathcal{D}, \quad \mathcal{E} \cap \mathcal{D} = \emptyset, \quad (2)$$

where $\mathcal{E} = \{e_1, \ldots, e_K\}$ is the set of *evidence sessions* containing information relevant to the evaluation task, and $\mathcal{D} = \{d_1, \ldots, d_L\}$ is the set of

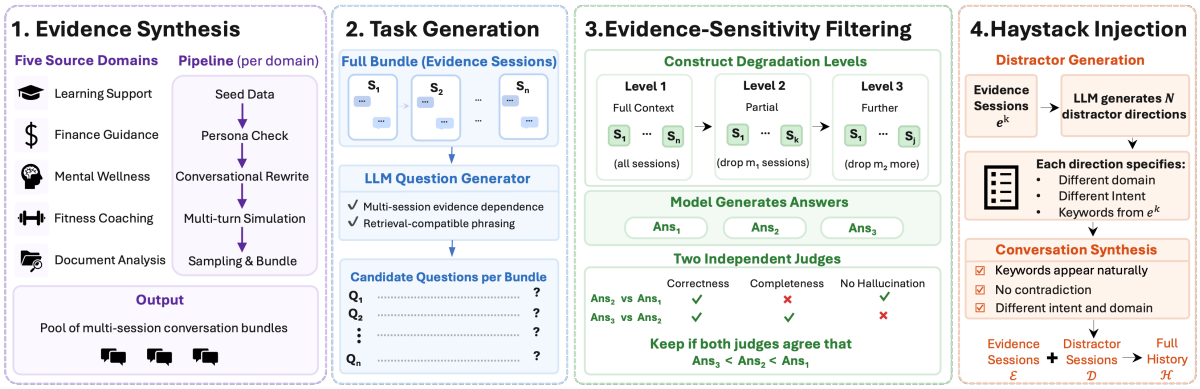

Figure 2: Overview of the UTILMEM construction pipeline, including multi-session evidence synthesis, task generation, evidence-sensitivity filtering via counterfactual degradation, and adversarial distractor injection for haystack construction.

haystack sessions. Associated with each instance is a *question pool*

$$\mathcal{Q} = \{q_1, q_2, \ldots, q_n\}, \tag{3}$$

where each $q_j \in \mathcal{Q}$ is an open-ended question designed to require synthesis across multiple evidence sessions in $\mathcal{E}$. At evaluation time, for each question $q \in \mathcal{Q}$ a system receives the history $\mathcal{H}$ and the question q, and produces a response $\hat{a}$.

3.1 Evidence Synthesis

To ensure realism and diversity, we seed UTILMEM with content drawn from five public datasets, each capturing a distinct mode of user–assistant interaction. Learning Support draws from StudyChat (McNichols et al., 2025), a corpus of student conversations with an AI tutor across seven course topics in an undergraduate AI class. Finance Guidance draws from Personal-Finance-Queries (Theerthala, 2025), a Reddit-sourced question–answer dataset covering budgeting, investing, debt, insurance, and savings. Mental Wellness draws from a Reddit mental health corpus (Solomonk, 2022) containing first-person posts from subreddits such as *depression*, *ADHD*, *OCD*, and *ptsd*. Fitness Coaching draws from a fitness question–answer dataset[1] covering exercise programming, nutrition, and recovery. Document Analysis draws from S&P 500 10-K filings[2], where each filing is decomposed into standardized items (e.g., Business, Risk Factors, Legal Proceedings). Each domain is processed through a synthesis pipeline that shares a common four-step structure. (1) Sampling. Source items are grouped into bundles $B = \{s_1, \ldots, s_K\}$ representing a single user's history, with sessions drawn from different topical subareas within the domain to encourage breadth of evidence. (2) Persona consistency. An LLM judge rejects bundles containing contradictions in stable user attributes such as conflicting ages or mutually exclusive circumstances, ensuring each bundle plausibly originates from one individual. (3) Conversational rewriting. Source items that are not already in dialogue form are rewritten into natural user–assistant exchanges, after which a back-and-forth loop between an agent model and a user model expands each seed into a multi-turn session while preserving all concrete details. (4) Deduplication. Bundle-level deduplication ensures no two samples share the same source-item set. The output is a pool of conversation bundles, each a plausible multi-session interaction history for a single user. Detailed per-domain processing is provided in Appendix C.

3.2 Question Generation

For each bundle B, we generate candidate questions satisfying two properties. *Multi-session dependence* is a generation target: a candidate question q should benefit from aggregating information distributed across the sessions in B. *Retrieval-compatible phrasing* requires that q uses terms appearing in the conversation text (e.g., *my AI course assignments*, *credit card debt*), ensuring that retrieval-based systems can operate on the questions as posed. Beyond these two properties, we additionally require questions to be *deterministic and non-judgmental*, asking for organization, integration, listing, or planning operations whose

[1] https://huggingface.co/datasets/its-myrto/fitness-question-answers

[2] https://huggingface.co/datasets/jlohding/sp500-edgar-10k

outputs are grounded directly in the conversation content, rather than for subjective diagnoses (e.g., *weaknesses*, *most important takeaway*) that would yield unstable answers across model runs. A language model receives the complete text of all sessions in B with internal session labels stripped, and emits ten candidate questions per bundle. Domain-specific instantiations of the prompt guide the question types appropriate to each domain. The full system prompt and the per-domain templates are provided in Appendix G.

3.3 Evidence-Sensitivity Filtering

Not all candidate questions genuinely require all sessions. Some are answerable at comparable quality from a subset due to information redundancy or underspecification. We filter these via a counterfactual degradation test. For each bundle with K evidence sessions, we construct three context levels by progressively removing sessions:

$$\mathcal{E}^{(1)} = \mathcal{E}, \quad \mathcal{E}^{(2)} = \mathcal{E}^{(1)} \setminus R_1, \quad \mathcal{E}^{(3)} = \mathcal{E}^{(2)} \setminus R_2, \tag{4}$$

where $R_1, R_2 \subset \mathcal{E}$ are randomly selected removal sets. The removal sizes $|R_1|$ and $|R_2|$ are adapted proportionally to K, targeting approximately 50% total removal while ensuring $|\mathcal{E}^{(3)}| \geq 1$. A target model generates answers $\hat{a}^{(1)}, \hat{a}^{(2)}, \hat{a}^{(3)}$ at each context level. Two independent judge models then perform pairwise comparisons, each rendering a binary verdict on whether the degraded answer is worse than the fuller-context answer. We verify *monotonic degradation* and retain a question only if both judges agree that

$$\text{quality}(\hat{a}^{(1)}) > \text{quality}(\hat{a}^{(2)}) > \text{quality}(\hat{a}^{(3)}), \tag{5}$$

which requires four concordant binary verdicts from two judges across two adjacent comparisons. This conservative sampled degradation test retains questions whose answer quality decreases along one sequential removal path, R_1 followed by R_2. It establishes evidence sensitivity under the sampled removals. The system prompt used for the pairwise judges is provided in Appendix H. Human verification and case study are provided in Appendix E.1.

3.4 Haystack Construction

Weak distractors allow systems to succeed through shallow retrieval, reducing evaluation to recall rather than utilization. To stress-test robustness under retrieval interference, we construct adversarial distractors that remain semantically close to the evidence sessions $\mathcal{E}$ while being functionally irrelevant to the target question.

Distractor synthesis. For each evidence session $e_k \in \mathcal{E}$, we generate distractors in two stages. First, an LLM reads e_k and produces N distractor directions sampled from curated cross-domain families. Each direction specifies a target domain, a conversational intent, a domain distinguisher, and 5–10 keywords shared with e_k. The shared keywords preserve lexical overlap with the evidence, while the new domain and intent shift the underlying semantics. Second, each direction is instantiated into M multi-turn conversations with turn counts sampled uniformly from $[\tau_{\min}, \tau_{\max}]$. The synthesis prompt enforces three constraints: (1) shared keywords must appear naturally, (2) domain-specific terminology must dominate the conversation so the distractor intent is immediately recognizable, and (3) no evidence content may be copied or contradicted. As a result, distractors remain retrieval-plausible but answer-irrelevant.

Haystack assembly. Each evidence session generates $N \cdot M$ in-domain distractors, and all distractors are aggregated into a single haystack $\mathcal{D}$. To additionally simulate the heterogeneous content found in real user histories, the haystacks of the five domains are then merged so that each sample's history $\mathcal{H} = \mathcal{E} \cup \mathcal{D}$ contains its own in-domain evidence and strong distractors alongside cross-domain sessions that act as weak distractors. Real timestamps from the fitness domain are used as temporal anchors, while sessions from other domains receive synthesized timestamps that preserve their relative ordering. Distractors are distributed throughout the timeline before the full history is sorted chronologically for evaluation.

This construction separates evidence dependence from retrieval robustness. The evidence-sensitivity filter selects questions that degrade along the sampled removal path, while the adversarial haystack tests whether systems can recover the relevant sessions under semantically similar interference. Appendix D presents 15 representative questions drawn from the released data, and Appendix E.2 gives two human-checked development examples in which strong lexical overlap masks a functionally irrelevant context. Benchmark statistics summarizing the resulting evidence counts, haystack composition, and per-sample token bud-

get are reported in Appendix A.

4 Experimental Setup

Baselines. We evaluate representative long-term memory systems spanning retrieval-based and structured memory architectures. Retrieval-only baselines use NaiveRAG with Nomic-embed-text-v1.5 (Nussbaum et al., 2024), Qwen3-Embedding-4B, and Qwen3-Embedding-8B (Zhang et al., 2025). Structured and agentic memory systems include A-MEM (Xu et al., 2025), Mem0 (Chhikara et al., 2025), Mem0+Graph, MemOS (Li et al., 2025), LangMem (The LangChain Team, 2025), and EverMemOS (Hu et al., 2026a). These systems cover diverse memory designs including raw retrieval, graph-augmented retrieval, hierarchical memory organization, profile-based aggregation, and write-time compression.

Evaluation protocol. UTILMEM requires long-form synthesis across multiple sessions, so exact-match is inappropriate. We instead measure how faithfully a memory system preserves answer quality under realistic retrieval, relative to a noise-free upper bound. For each instance $(\mathcal{H}, q, \mathcal{E})$, we obtain a *reference answer* a_{oracle} by conditioning the generator on only the oracle evidence $\mathcal{E}$ and the query q. Each memory system under test ingests the full history $\mathcal{H}$, and the generator is conditioned on the memory *that system retrieves* for q, yielding $\hat{a}$. We pin a_{oracle} to score 10 as a protocol reference shared by all systems; it is not a human optimum, and the resulting comparison is relative rather than an absolute measure of answer quality (see Appendix E for detailed discussion). A judge then scores $\hat{a}$ against this reference, conditioned on $\mathcal{E}$ and q. To reduce judge variance, we use two independent judge models and average their scores. The judge first classifies $\hat{a}$ along five sub-dimensions into ordinal severity levels (*intact*, *mild*, *clear*, *severe*): **factual fidelity** (whether claims are consistent with the oracle context or the reference answer), **completeness preservation** (whether question-relevant content from the context is covered to a comparable extent), **hallucination absence** (whether the answer fabricates facts that are absent from both the context and the reference), **semantic equivalence** (whether a reader would reach the same conclusion or take the same action from $\hat{a}$ as from a_{oracle}), and **ungrounded inference** (whether the answer speculates or fills gaps beyond what the context supports). Conditioned on these severities, the judge then assigns a single integer score $s \in [1, 10]$ via explicit anchors (full prompt in Appendix F). The score is anchor-based rather than an average over sub-dimensions, so that a single severe failure caps s and cannot be offset by strength on unaffected dimensions. We provide analysis of consistency between the two judges in Appendix E.3.

We report three metrics, aggregated over the N evaluated instances:

- **Robustness Score (RS)** $\in [1, 10]$: the mean per-instance score, $\text{RS} = \frac{1}{N}\sum_i s_i$, with the oracle reference pinned at 10.
- **Normalized Robustness (NR)** $\in [0, 100]$: a rescaling of RS to a 0–100 range, $\text{NR} = \frac{\text{RS}-1}{9} \times 100$, read directly as the percentage of the clean-reference score under the evaluation protocol a system retains. We treat NR as our headline metric.
- **Degradation Rate (DR)** $\in [0, 100]$: the fraction of instances with $s_i \leq 6$, the first rubric region containing a clear, user-relevant loss. Score 7 retains the core facts and conclusion, whereas scores at or below 6 contain a bounded clear error or a more serious failure.

A system succeeds only if $\hat{a}$ approaches the clean-reference score: high NR with DR near zero.

Implementation. We use Qwen3-235B-A22B-Instruct-2507 for benchmark construction and QA generation. Temperatures are set to 0.7 during benchmark data synthesis and 0 for QA generation and evaluation to ensure reproducibility. We use GPT-5.2-1211-global and GPT-5.4-0305-global as judge models, with scores averaged across both judges. For memory-specialized systems, we follow the recommended ingestion and retrieval pipeline from MemBase (Fang et al., 2025). We provide a detailed breakdown of models used at each stage in Appendix B.

5 Results

5.1 Compression-Heavy Systems on Detail-Sensitive Utilization Tasks

Table 2 shows that preserving raw conversational evidence is more important than increasing memory structure or embedding capacity. **1) Factual-oriented write-time compression creates irreversible information loss.** Mem0 performs worst overall at 17.3 NR and 98.3 DR, while Mem0+Graph only slightly improves to 18.9 NR and 97.8 DR. The failure persists even with graph

Method	Learning Support		Finance Guidance		Mental Wellness		Fitness Coaching		Document Analysis		Average	
	NR↑	DR↓	NR↑	DR↓	NR↑	DR↓	NR↑	DR↓	NR↑	DR↓	NR↑	DR↓
Retrieval-only baseline (NaiveRAG) with varying embedding backbones												
NaiveRAG + *Nomic-embed-text-v1.5*	48.0	74.4	**69.7**	25.4	54.9	62.7	67.0	32.6	55.6	52.9	58.9	49.6
NaiveRAG + *Qwen3-Embedding-4B*	**49.2**	**73.1**	68.4	26.2	**55.7**	**61.5**	66.1	33.8	**56.8**	**51.7**	59.2	49.3
NaiveRAG + *Qwen3-Embedding-8B*	48.7	73.8	**69.7**	**24.9**	55.2	62.1	**67.6**	**31.9**	55.1	53.4	**59.3**	**49.2**
Agentic / structured memory systems												
A-MEM	45.6	78.2	67.4	28.6	51.8	68.9	64.2	36.8	50.9	59.4	56.0	54.4
Mem0	7.4	99.7	22.1	99.4	30.8	93.7	24.4	98.9	1.7	100.0	17.3	98.3
Mem0 + *Graph*	8.9	99.4	24.7	98.7	32.1	92.8	26.2	98.1	2.8	99.8	18.9	97.8
MemOS	39.8	80.7	57.2	55.4	47.9	72.6	54.1	61.8	38.5	78.9	47.5	69.9
LangMem	33.2	87.4	49.6	69.2	42.8	80.1	47.3	75.8	30.7	90.6	40.7	80.6
EverMemOS	41.7	82.9	58.4	51.6	49.2	70.4	55.9	58.7	36.8	81.3	48.4	69.0

Table 2: Pairwise generation quality against the oracle answer across five domains. For each domain we report **NR** and **DR** against an oracle answer generated from complete, noise-free ground-truth evidence sessions. **Higher is better** for **NR** (↑), while **lower is better** for **DR** (↓). The shaded columns mark the average across all five domains. The best results are shown in bold.

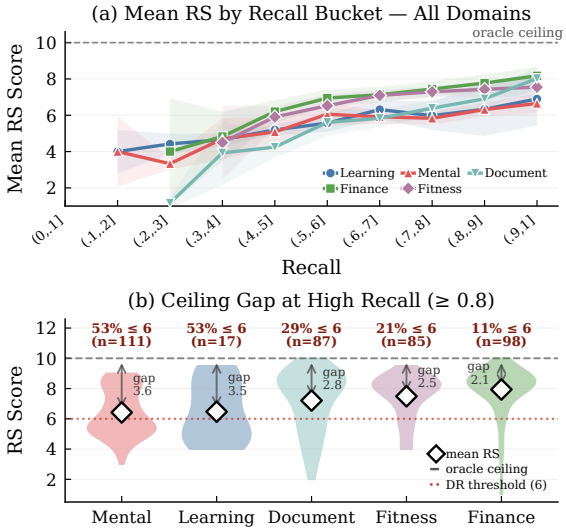

Figure 3: **(a)** Mean RS rises monotonically with unit-level recall in every domain. **(b)** Conditioning on instances where retrieval already succeeds (recall ≥ 0.8), the RS distribution per domain remains far below the clean-reference score of 10. White diamonds mark per-domain means; arrows mark the residual reference gap; red labels report the share of high-recall instances that still degrade (RS ≤ 6).

structure, indicating that relational organization cannot recover contextual and reasoning information discarded during fact extraction. **2) Partial evidence preservation improves robustness but still lags behind raw-turn retrieval.** A-MEM achieves the strongest structured performance at 56.0 NR and 54.4 DR, while MemOS, LangMem, and EverMemOS fall between 40.7–48.4 NR and 69.0–80.6 DR. These systems preserve more conversational context than Mem0, but still rely on summarization, rewriting, or profile abstraction that weakens implicit recall and cross-session synthesis. Overall, the results suggest that utilization-oriented memory depends primarily on preserving original evidence rather than constructing increasingly compact memory representations.

In benchmark construction, the sampled evidence-sensitivity filter preferentially retains questions whose answers lose useful detail when sessions are removed. Such tasks are deliberately demanding for representations that discard contextual details, so strong performance on existing factual memory benchmarks does not reliably transfer to UTILMEM-style evidence-utilization tasks. By contrast, the three NaiveRAG embedding variants remain tightly clustered at 58.9–59.3 NR and 49.2–49.6 DR, suggesting that, within these settings, embedding scale matters less than the evidence representation presented to the generator. Appendix E.4 presents two raw-log-verified Mem0 failures and confirms that retrieval was non-empty in both cases.

5.2 Recall Is Necessary but Not Sufficient for Utilization

To examine this, we compare per-instance RS with unit-level recall, defined as the fraction of ground-truth evidence sessions retrieved in the top-k. We use NAIVERAG for this analysis because its retrieved units are directly observable, unlike memory systems that delete or update memory at ingestion time. In Figure 3 (a), RS increases monotonically with recall across all domains. This confirms that the benchmark is evidence-sensitive: retrieving more relevant evidence consistently improves answer quality. However, high recall alone does not close the gap to the oracle. Figure 3 (b) isolates instances with recall ≥ 0.8. Even when most evi-

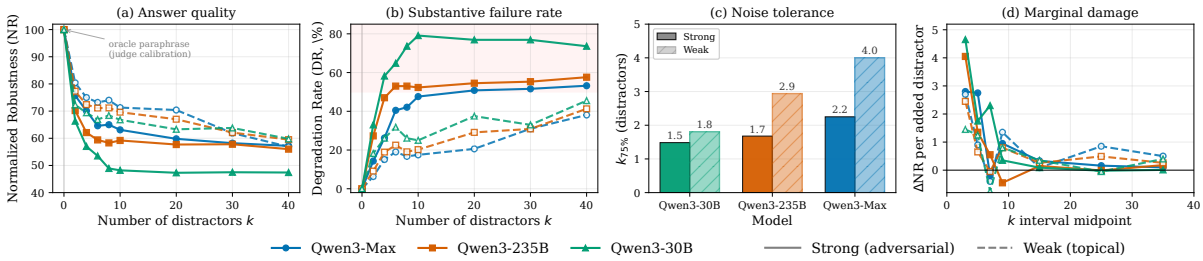

Figure 4: Effect of distractor count k at recall@k=1.0 across three generators and two noise types. **(a)** Normalized Robustness (NR): all curves decay monotonically from the oracle-paraphrase anchor at k=0. **(b)** Degradation Rate (DR, fraction of instances with $s_i \leq 6$): rises sharply, plateauing above 50% for strong noise across all models. **(c)** Noise tolerance, measured as $k_{75\%}$, the number of distractors required to lose 25% of oracle quality; under strong noise, every model loses a quarter of its quality within fewer than 2.5 distractors. **(d)** Marginal damage (ΔNR per added distractor) by interval: the first few distractors carry nearly all the cost, after which additional context is approximately free.

dence is retrieved, mean RS remains substantially below the clean-reference score of 10, ranging from 6.4 on Mental Wellness to 7.9 on Finance Guidance. A non-trivial fraction of these high-recall instances still fall below the DR threshold of 6, including 53% on Mental Wellness and 29% on Document Analysis. These results suggest that retrieval is necessary but insufficient for strong performance on UTILMEM. Once relevant evidence is retrieved, systems must still integrate information across sessions, distinguish evidence from semantically similar distractors, and compose coherent long-form responses. The remaining gap at high recall reflects failures in these downstream utilization steps rather than retrieval misses alone.

5.3 Sensitivity to Retrieval Interference

We isolate the effect of retrieval noise by fixing recall at 1.0 and progressively adding distractor sessions to the oracle evidence set. We vary three factors: the number of distractors k, distractor difficulty, and generator size. Strong noise uses semantically similar distractors from the same user trajectory, while weak noise uses unrelated cross-domain sessions. We shuffle the oracle and distractor sessions together so position is not a confound. The $k = 0$ condition is a protocol normalization anchor. Rather than re-conditioning the generator on oracle-only context, which collapses to a near-identical regeneration of a_{oracle}, we obtain $\hat{a}$ by having the judge lightly paraphrase a_{oracle} with surface-level edits that preserve all facts, structure, and length, so $k = 0$ measures the attainable RS ceiling under our protocol. All other settings follow the protocol above unchanged. Figure 4 reports the results. **1) Increasing retrieval depth is monotonically lossy under retrieval noise.** Figure 4 (a,b) shows that answer quality consistently degrades as more distractors are added, even though all relevant evidence remains in context. Under strong noise, NR drops from 100 to 47.4, 56.0, and 57.2 at $k = 40$ for Qwen3-30B, Qwen3-235B, and Qwen3-Max respectively, while DR rises to 73.6%, 57.6%, and 53.2%. Weak noise is less harmful but still causes substantial degradation, with NR remaining only around 57 to 60 at $k = 40$. These results suggest that adding more retrieved context does not reliably improve utilization performance once distractors are introduced. **2) Distractor difficulty matters more than context length alone.** The gap between strong and weak noise remains large across all generators and values of k. In Figure 4 (a), strong-noise curves degrade substantially faster than weak-noise curves, particularly in the low-k regime. At $k = 10$, the NR gap between weak and strong noise reaches 18.6 for Qwen3-30B, 10.4 for Qwen3-235B, and 8.2 for Qwen3-Max. Figure 4 (b) shows the same pattern for DR. The dominant failure mode is therefore not longer context length itself, but semantically plausible distractors that overlap with the target evidence in entities and vocabulary. **3) Larger generators improve absolute robustness but do not remove sensitivity to noise.** Across both noise settings, larger generators consistently achieve higher NR and lower DR. For example, under strong noise at $k = 10$, Qwen3-Max reaches 63.1 NR compared to 59.2 for Qwen3-235B and 48.2 for Qwen3-30B. However, all generators exhibit similar degradation trends as distractors accumulate. The relative improvement from scaling the generator is modest compared to the overall loss caused by adversarial

noise. **4) Most quality degradation comes from the first few distractors.** Figure 4 (c,d) shows that the damage is concentrated in the early retrieval regime. Under strong noise, $k_{75\%}$ is only 1.5 for Qwen3-30B, 1.7 for Qwen3-235B, and 2.2 for Qwen3-Max, meaning that only a few adversarial distractors are sufficient to remove 25% of the clean-context reference score. Figure 4 (d) further shows that marginal NR loss peaks within roughly $k \in [2, 8]$ and rapidly saturates afterward. Beyond $k \approx 10$, additional distractors contribute comparatively little additional degradation because the context is already heavily contaminated. Overall, the results identify retrieval precision as a larger source of degradation than recall or generator scale in this controlled interference study. Once semantically similar distractors enter the context, downstream evidence integration degrades rapidly even when all relevant evidence is present.

6 Conclusion

We introduced UTILMEM, a diagnostic benchmark that systematically operationalizes an essential but under-evaluated requirement of long-term conversational memory: using distributed evidence to produce grounded, task-oriented outputs under retrieval interference. Across the systems and configurations tested, factual-recall strength does not reliably transfer to utilization quality; high recall can coexist with integration failures, and a small number of semantically similar distractors causes substantial degradation. Systems preserving more turn-level evidence also outperform the compression-heavy systems tested on these detail-sensitive tasks, although the benchmark's evidence-sensitivity filter contributes a selection effect and the result should not be generalized to all memory use cases. These findings expose reproducible failure modes. We hope UTILMEM supports work that combines human-grounded validation with causal decomposition of writing, retrieval, and generation failures, and that tests interventions for preserving evidence while suppressing functionally irrelevant memories.

7 Limitations

UTILMEM relies on LLM-based construction and evaluation. Conversational rewriting, question generation, distractor synthesis, oracle-answer generation, filtering, and final scoring may inherit stylistic regularities, templated structures, biases, or reasoning preferences from the models used. The oracle-evidence answer is a shared protocol reference, not a human optimum. Long-form planning, finance, and mental-wellness questions can admit multiple valid organizations and emphases, so absolute scores depend on the stability of the reference and rubric. The two GPT-family judges agree strongly in ranking and near-score agreement (Appendix E.3), but this does not establish judge–human validity or cross-family independence. Larger human evaluation analyses remain important future work. The evidence-sensitivity procedure verifies degradation along sampled sequential removal paths. It does not exhaustively prove that every evidence session is individually indispensable, that all subsets yield monotonic degradation, or that retained evidence sets are minimal. Accordingly, the reported findings is scoped to UTILMEM-style evidence utilization and the evaluated configurations, not to long-term memory ability in general.

Acknowledgment

This research was supported in part by the National Science Foundation under Grant No. 2452367.

References

Haonan Bian, Zhiyuan Yao, Sen Hu, Zishan Xu, Shaolei Zhang, Yifu Guo, Ziliang Yang, Xueran Han, Huacan Wang, and Ronghao Chen. 2026. RealMem: Benchmarking LLMs in real-world memory-driven interaction. *arXiv preprint arXiv:2601.06966*.

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

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.

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*.

Xingjian Diao, Chunhui Zhang, Weiyi Wu, Zhongyu Ouyang, Peijun Qing, Ming Cheng, Soroush Vosoughi, and Jiang Gui. 2025. Temporal working memory: Query-guided segment refinement for enhanced multimodal understanding. *arXiv preprint arXiv:2502.06020*.

Yiming Du, Wenyu Huang, Danna Zheng, Zhaowei Wang, Sebastien Montella, Mirella Lapata, Kam-Fai Wong, and Jeff Z. Pan. 2025. Rethinking memory in AI: Taxonomy, operations, topics, and future directions. *arXiv preprint arXiv:2505.00675*.

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From local to global: A Graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*.

Jizhan Fang, Xinle Deng, Haoming Xu, Ziyan Jiang, Yuqi Tang, Ziwen Xu, Shumin Deng, Yunzhi Yao, Mengru Wang, Shuofei Qiao, Huajun Chen, and Ningyu Zhang. 2025. LightMem: Lightweight and efficient memory-augmented generation. *arXiv preprint arXiv:2510.18866*.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. LightRAG: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. HippoRAG: Neurobiologically inspired long-term memory for large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Chuanrui Hu, Xingze Gao, Zuyi Zhou, Dannong Xu, Yi Bai, Xintong Li, Hui Zhang, Tong Li, Chong Zhang, Lidong Bing, and Yafeng Deng. 2026a. EverMemOS: A self-organizing memory operating system for structured long-horizon reasoning. *arXiv preprint arXiv:2601.02163*.

Yuanzhe Hu, Yu Wang, and Julian McAuley. 2026b. Evaluating memory in LLM agents via incremental multi-turn interactions. In *International Conference on Learning Representations (ICLR)*.

Jiazheng Kang, Mingming Ji, Zhe Zhao, and Ting Bai. 2025. Memory OS of AI agent. *arXiv preprint arXiv:2506.06326*.

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

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhiyu Li, Shichao Song, Chenyang Xi, Hanyu Wang, Chen Tang, Simin Niu, Ding Chen, Jiawei Yang, Chunyu Li, Qingchen Yu, et al. 2025. MemOS: A memory OS for AI system. *arXiv preprint arXiv:2507.03724*.

Xinnian Liang, Bing Wang, Huijia Huang, Shuangzhi Wu, Peihao Wu, Lu Lu, Zejun Ma, and Zhoujun Li. 2023. SCM: Enhancing large language model with self-controlled memory framework.

Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. 2025. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*.

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

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

Hunter McNichols, Fareya Ikram, and Andrew Lan. 2025. The StudyChat dataset: Student dialogues with ChatGPT in an artificial intelligence course. *arXiv preprint arXiv:2503.07928*.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. 2025. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. 2024. Nomic embed: Training a reproducible long context text embedder. *arXiv preprint arXiv:2402.01613*.

Charles Packer, Vivian Fang, Shishir G. Patil, Kevin Lin, Sarah Wooders, and Joseph E. Gonzalez. 2023. MemGPT: Towards LLMs as operating systems. *arXiv preprint arXiv:2310.08560*.

Egor Pakhomov, Erik Nijkamp, and Caiming Xiong. 2025. ConvoMem benchmark: Why your first 150 conversations don't need RAG. *arXiv preprint arXiv:2511.10523*.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*.

Peijun Qing, Xingjian Diao, Chiyu Ma, Saeed Hassanpour, and Soroush Vosoughi. 2026a. Tailoring memory granularity for multi-hop reasoning over long contexts. In *Findings of the Association for Computational Linguistics: EACL 2026*.

Peijun Qing, Puneet Mathur, Nedim Lipka, Varun Manjunatha, Ryan Rossi, Franck Dernoncourt, Saeed Hassanpour, and Soroush Vosoughi. 2026b. Cluster-r1: Large reasoning models are instruction-following clustering agents. *arXiv preprint arXiv:2603.23518*.

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, Wei Wei, and Yujia Bao. 2025. From isolated conversations to hierarchical schemas: Dynamic tree memory representation for LLMs. In *International Conference on Learning Representations (ICLR)*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning (ICML)*.

Solomonk. 2022. Reddit mental health posts dataset. Hugging Face dataset.

Haoran Tan, Zeyu Zhang, Chen Ma, Xu Chen, Quanyu Dai, and Zhenhua Dong. 2025. MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. In *Findings of the Association for Computational Linguistics: ACL 2025*.

The LangChain Team. 2025. LangMem SDK for agent long-term memory. https://www.langchain.com/blog/langmem-sdk-launch.

Akhil Theerthala. 2025. Reddit personal finance q&a dataset. Hugging Face dataset.

Yu Wang and Xi Chen. 2025. MIRIX: Multi-agent memory system for LLM-based agents. *arXiv preprint arXiv:2507.07957*.

Yu Wang, Chi Han, Tongtong Wu, Xiaoxin He, Wangchunshu Zhou, Nafis Sadeq, Xiusi Chen, Zexue He, Wei Wang, Gholamreza Haffari, Heng Ji, and Julian McAuley. 2024. Towards lifespan cognitive systems. *arXiv preprint arXiv:2409.13265*.

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 *International Conference on Learning Representations (ICLR)*.

Wujiang Xu, Zujie Liang, Kai Mei, Hang Gao, Juntao Tan, and Yongfeng Zhang. 2025. A-Mem: Agentic memory for LLM agents. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Zeyu Zhang, Xiaohe Bo, Chen Ma, Rui Li, Xu Chen, Quanyu Dai, Jieming Zhu, Zhenhua Dong, and Ji-Rong Wen. 2024. A survey on the memory mechanism of large language model based agents. *arXiv preprint arXiv:2404.13501*.

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

A Benchmark Statistics

This appendix summarizes the structural properties of the final UTILMEM benchmark used in our experiments. We characterize the benchmark along three axes: how many evidence sessions each sample requires (Figure 5), how those evidence sessions are surrounded by in-domain strong distractors and cross-domain weak distractors (Figure 6), and the resulting per-sample token budget that a memory system must operate over (Figure 7). We compare against LONGMEMEVAL-S (Wu et al., 2025), which is the closest prior benchmark in scale and motivation.

Evidence session counts. Figure 5 shows the distribution of ground-truth evidence sessions per sample. The five UTILMEM domains have medians of 6, 3, 4, 6, and 5 for Learning Support, Finance Guidance, Mental Wellness, Fitness Coaching, and Document Analysis respectively, all materially higher than the median of 2 in LONGMEMEVAL-S. The mass of the distribution sits between 3 and 7 designated evidence sessions per sample, increasing the multi-session evidence burden relative to localized recall. Retained questions pass the sampled degradation test in Section 3.3; these counts alone do not prove that every session is individually indispensable.

Haystack composition. Figure 6 decomposes each sample's history into three categories: evidence sessions (GT), strong distractors synthesized to be semantically adjacent to the evidence (§3.4), and weak distractors imported from the other four domains within the same instance. Across domains, evidence sessions account for between 1.9 and 5.9 sessions per sample, while the surrounding strong and weak distractors raise the average per-sample session count to roughly 60–70. On the token axis, Document Analysis carries the largest per-session footprint, with evidence alone averaging 46K tokens per sample owing to the length of 10-K filing items, while the other four domains contribute between 4K and 10K tokens of evidence per sample. By construction, the strong-distractor band dominates the haystack in every domain, which is the design choice that makes UTILMEM a stress test of retrieval robustness rather than of recall.

Total token budget. Figure 7 reports the per-sample total token budget after the multi-domain haystacks are concatenated. The mean per-sample budget in UTILMEM is 120K tokens, closely matched to the 122K mean of LONGMEMEVAL-S. This alignment reduces raw context length as an obvious explanation for the performance gap and makes evidence multiplicity (Figure 5) and distractor adversariality (Figure 6) more directly comparable. It does not eliminate other benchmark and implementation differences or provide a causal decomposition.

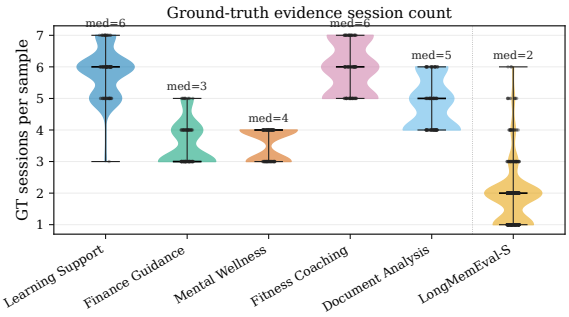

Figure 5: Distribution of ground-truth evidence-session counts per sample, by domain. Medians are annotated above each violin. The five UTILMEM domains all sit materially above LONGMEMEVAL-S (median 2), and the bulk of mass between 3 and 7 designated evidence sessions increases the evidence aggregation burden.

B Models Used Across Benchmark Construction and Evaluation

Table 3 summarizes the models used at each stage of UTILMEM. During benchmark construction, temperatures are set to 0.7 for conversational synthesis and distractor generation to encourage diversity and natural variation. During evaluation, QA generation and judge scoring use temperature 0.

C Per-Domain Synthesis Details

This appendix describes how the four-step synthesis pipeline of Section 3 is instantiated for each of the five source datasets. Each pipeline produces a bundle $B = \{s_1, \ldots, s_K\}$ of multi-turn sessions representing a single user's history, together with the metadata consumed by the downstream question-generation and haystack-construction stages.

Learning Support (StudyChat). StudyChat (McNichols et al., 2025) provides per-turn rows from an undergraduate AI course, each annotated with a user ID, a chat ID, and a topic label from seven course topics. For each (user, topic, chat) triple we retain only the final-turn snapshot, whose message list contains the complete

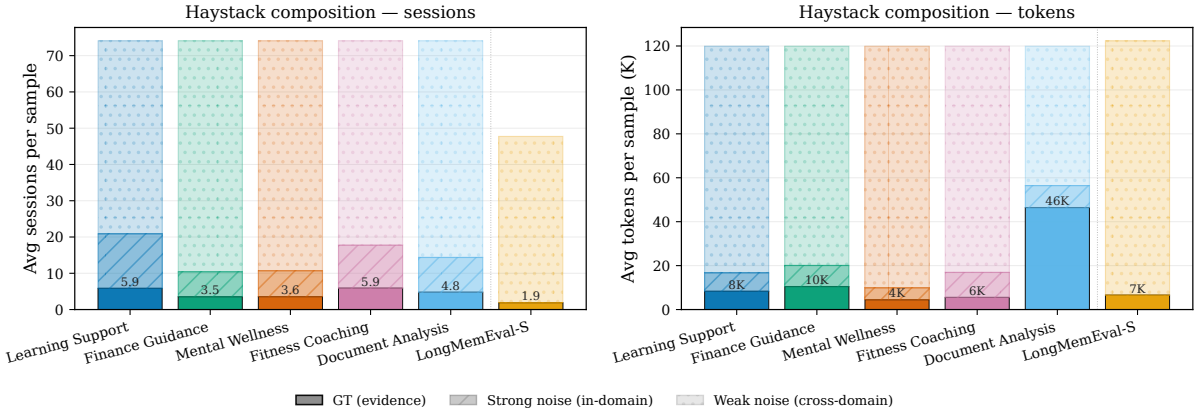

Figure 6: Haystack composition per sample, decomposed into evidence sessions (GT), in-domain strong distractors, and cross-domain weak distractors. **Left:** average number of sessions per sample. **Right:** average token count per sample, in thousands. Strong distractors dominate every domain by construction, making the benchmark a test of retrieval robustness rather than of recall. Document Analysis carries the largest evidence footprint on the token axis because 10-K filing items are individually long.

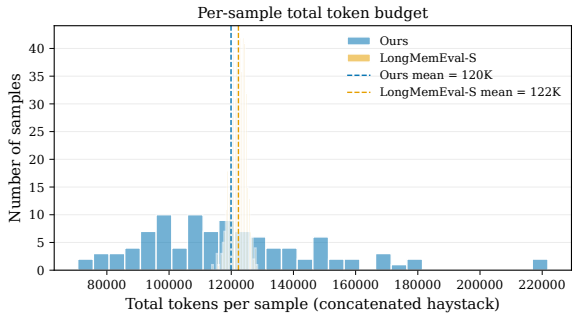

Figure 7: Per-sample total token budget after concatenating the multi-domain haystack. The mean budget in UTILMEM (120K tokens) is closely matched to that of LONGMEMEVAL-S (122K tokens), reducing raw context length as an obvious explanation while leaving other benchmark differences uncontrolled

.

dialogue. A bundle is formed by selecting one user, sampling K topics covered by that user, and drawing one chat per sampled topic uniformly at random. Persona consistency is enforced by construction, since all sessions in a bundle share the same user ID, and bundles whose chat-ID set was previously accepted are rejected.

Finance Guidance (Personal-Finance-Queries). Personal-Finance-Queries (Theerthala, 2025) consists of Reddit-style QA rows labelled with a finance category. We restrict to five categories spanning debt, investing, budgeting, insurance, and savings. A bundle is formed by sampling K categories and drawing M QA pairs per category. An LLM judge then inspects all sampled queries jointly and rejects bundles with conflicting demographics, life stages, or mutually exclusive financial circumstances. Surviving queries are rewritten into natural conversational form, and each rewritten pair is expanded into a T-turn dialogue between an agent model and a user model. Bundles whose source row-ID set was previously accepted are rejected before any LLM calls.

Mental Wellness (Reddit Mental Health Posts). This pipeline ingests posts from a curated set of mental health subreddits including *ADHD*, *Aspergers*, *depression*, *OCD*, and *ptsd*, with each post body treated as the seed for one session. A bundle is formed by sampling K subreddits and drawing one post per subreddit. An LLM judge then screens for persona consistency across the sampled bodies, rejecting bundles with incompatible self-descriptions. Accepted post bodies are rewritten into conversational queries via a batched LLM call, after which an agent model and a user model alternate for T turns per query, using distinct backbones for the two roles to reduce stylistic collapse.

Fitness Coaching (Fitness Q&A). The fitness pipeline targets longitudinal coaching scenarios in which a user returns to an AI fitness coach over days, weeks, or months. A judge LLM first decides whether each seed question is inherently longitudinal, and questions that are not are either rewritten into longitudinal form by a second judge call or discarded. For each accepted topic, an agent model and a user model alternate for T turns to produce a single session, and this step is then repeated for

Stage	Role	Model
Benchmark construction		
Multi-session synthesis	Assistant role (agent)	Qwen3-235B-A22B-Instruct-2507
	User role (simulated)	Qwen3-235B-A22B-Instruct-2507
	Reddit-style rewriter	Qwen3-235B-A22B-Instruct-2507
	Topic-screening judge	GPT-5.2-1211-global
Question generation and filtering	Question generator	GPT-5.2-1211-global
	Target model (counterfactual generation)	Qwen3-235B-A22B-Instruct-2507
	Pairwise sensitivity judges (2×)	GPT-5.2-1211-global, GPT-5.4-0305-global
Distractor synthesis	Direction generator	Qwen3-Max
	Distractor conversation synthesis	Qwen3-Max
Evaluated systems		
Memory architecture	Extraction / compression LLM	Qwen3-235B-A22B-Instruct-2507
	Embedding backbone (default)	Nomic-embed-text-v1.5
Evaluation		
QA and judging	QA generator (frozen across systems)	Qwen3-235B-A22B-Instruct-2507
	QA generators (model-scaling study)	Qwen3-30B-A3B-Instruct-2507, Qwen3-235B-A22B-Instruct-2507, Qwen3-Max
	Noise-robustness judges (2×)	GPT-5.2-1211-global, GPT-5.4-0305-global

Table 3: Models used at each stage of UTILMEM. Temperatures are set to 0.7 during benchmark data synthesis and 0 during QA generation and evaluation.

S sessions per bundle. Each new session is conditioned on the full prior history together with a time-gap prompt instructing the user model to reference progress or setbacks since the previous session, and is assigned a realistic 2026 timestamp by advancing from the previous session by a duration matching its time-gap description. The resulting per-session timestamps form a coherent timeline that the haystack-construction stage later uses as the temporal anchor for the shared history $\mathcal{H}$.

Document Analysis (S&P 500 10-K Filings). Each 10-K filing is decomposed into standardized items such as *Business*, *Risk Factors*, *Properties*, and *MD&A*, with items below a minimum character threshold discarded. A bundle is formed by drawing a single filing (one company–year pair) and sampling K items from it. Persona consistency is trivially satisfied since all sessions in a bundle pertain to the same filing. For each sampled item, an agent model playing a financial-analyst assistant and a user model playing the client alternate for T turns, producing a focused multi-turn dialogue grounded in the item's text. Bundles whose (filing, items) combination was previously accepted are rejected.

D Representative Benchmark Instances Across Five Domains

To make the benchmark tasks concrete, we present three questions from distinct bundles in each domain. The "integration cues" are editorial annotations that summarize the information types named in the query.

Learning Support: Cross-Assignment Synthesis

Example L1
Query. Based on all my AI coursework conversations, can you make one study guide that brings together the binary tree implementation help, the chronic kidney disease dataset features and labels, the regression and cross-validation notebook tasks, the MLP hyperparameter tuning experiments, and the character-level RNN assignment with text generation?
Integration cues. Tree implementation; dataset schema; regression and cross-validation; MLP tuning; character-level RNN.

Example L2
Query. Can you give me one combined recap of my AI coursework progress that covers the documentary reflection on automation and jobs, BFS in graphs, formatting and summarizing dataset tables, creating linear regression examples, understanding accuracy score versus model score, building character n-gram counts, and preparing text for an RNN?
Integration cues. Social implications of AI; graph search; tabular analysis; regression; evaluation metrics; n-grams; RNN preprocessing.

Example L3
Query. Across all my AI coursework conversations, can you give me a combined notes sheet that lists the main algorithms, libraries, datasets, and reporting tasks I discussed, including trie-based autocomplete with BFS/DFS/UCS, LLM data conversion safeguards, pandas and scikit-learn for slime regression and kidney disease classification, MLP tuning settings, and n-gram probability-based next-character prediction?
Integration cues. Multiple search algorithms; trustworthy data conversion; two modeling tasks; hyperparameter settings; probabilistic sequence prediction.

Finance Guidance: Multi-Constraint Planning

Example F1
Query. Can you help me put together a single financial plan for my personal financial situation that covers cleaning up my collections, dealing with the disputed hospital charge, saving for my future move, making a smart car decision, and thinking through rent versus buying?
Integration cues. Credit cleanup; disputed medical debt; near-term liquidity; vehicle costs; housing choice.

Example F2
Query. Can you give me a full-picture plan for what I should do before and after leaving the military, using everything we discussed about housing, emergency savings, the 7.6% car loan, my monthly expenses, and the possibility of taking a lower-paying but higher-growth job?
Integration cues. Career transition; income uncertainty; debt cost; emergency reserves; housing timing.

Example F3
Query. Looking at my finances as a whole, what are all the key deadlines, decisions, and follow-up steps I should keep track of for the car situation, my savings accounts and possible investments for a house, the settlement notice, and planning for my kids?
Integration cues. Debt decision; savings vehicles; home goal; legal/administrative deadline; children's savings.

Mental Wellness: Grounded Support Across Evolving Concerns

Example M1
Query. Can you explain, using examples from my own experiences that I described across all sessions, how trauma-related numbness/sadness, depression-related brain fog, ADHD focus problems, relationship stress around reassurance, and OCD-like intrusive thoughts/compulsions can overlap—while staying grounded in what I actually told you and not diagnosing me?
Integration cues. Symptom overlap; personal evidence; relationship context; explicit non-diagnosis constraint.

Example M2
Query. Can you create a structured exposure plan for my contamination OCD that fits with my real-life constraints I described (limited clothing, fear of contaminating the washing machine, living in a developing country, folliculitis worries), and also add trauma-sensitive steps for days when my nervous system is already activated by neighborhood stress like helicopters, police activity, and reminders of the riots?
Integration cues. Exposure planning; resource constraints; health concern; environmental trauma triggers.

Example M3
Query. Can you write a therapist-ready "starting point" paragraph for me that combines everything I've shared across our conversations—my PTSD and cannabis-triggered flashbacks, my difficulty setting boundaries with my husband, my insomnia and sleep guilt tied to medication changes and comparison to my spouse, my repeated text checking, and my self-harm urges when good things happen—so I don't freeze or downplay it in the first session?
Integration cues. Trauma; boundaries; medication and sleep; compulsive checking; safety-relevant disclosure goal.

Fitness Coaching: Longitudinal Progression and Adaptation

Example C1
Query. Based on everything we've discussed about my workouts and knee rehab, can you lay out the full progression of exercises, sets, reps, and frequency I've used from the start up to my current routine?
Integration cues. Exercise sequence; dosage; frequency; rehabilitation progression over time.

Example C2
Query. Considering all my conversations about my workouts and daily routine, what does my current best-practice plan look like for creatine dosing, meal timing, hydration, sleep habits, movement during work, and handling both morning and later-day training sessions?
Integration cues. Supplement timing; nutrition; recovery; desk-work constraints; schedule-dependent training.

Example C3
Query. Across all my conversations about sprint training and recovery, can you pull together every symptom, lab value, nutrition change, and training adjustment into one clear progress update I could share with my coach or doctor?
Integration cues. Symptoms; bloodwork; supplementation and nutrition; training changes; audience-aware summary.

Document Analysis: Cross-Sectional 10-K Synthesis

Example D1
Query. Can you summarize all the places in FIRST SOLAR, INC.'s 2011 10-K where management relies heavily on estimates and timing (like percentage-of-completion, warranty/remediation accruals, recycling liability funding, and fair value/derivatives), and then tie that summary to the company's statement that controls are effective (including the ERP implementation) and the auditor's opinion?
Integration cues. Accounting estimates; timing assumptions; liabilities; internal controls; external audit.

Example D2
Query. From everything we reviewed about ESSEX PROPERTY TRUST INC's 10-K, can you draft a struc-

tured "how to read this filing" guide for me that walks in order through: (1) the property portfolio and how occupancy is defined (financial vs physical), (2) the big geographic concentration point (California share of rental revenues), (3) the interest-rate and debt/derivatives picture (fixed vs variable amounts and hedges), and (4) the reporting assurance layer (KPMG audit reports and management's controls conclusions)—so each step references the exact facts we discussed?
Integration cues. Operating definitions; concentration risk; financing and hedges; reporting assurance.

Example D3
Query. Can you summarize, in one place, how C H Robinson's liquidity and capital allocation story fits together across the filing: operating cash flow trends (2011–2013), the big 2013 income-tax payment tied to the T-Chek gain, acquisition cash outlays for Apreo and Phoenix, ongoing IT and building capex in Eden Prairie, dividends, share repurchases (including the 2013 ASR mechanics), and how the revolver and senior notes support these uses of cash?
Integration cues. Cash-flow history; tax event; acquisitions and capital expenditure; distributions; debt financing.

E Evaluation Reference and Validation Analyses

We treat the oracle-evidence answer a_{oracle} as a fixed protocol reference and assign it score 10 by definition. It is a model output rather than a human-authored gold answer, so this convention supplies a shared normalization anchor; it does not establish a human optimum or an absolute upper bound on answer quality.

Causal isolation. The generator and decoding configuration are held fixed (T=0) across the oracle and memory-system conditions. The controlled input is therefore the context supplied to that generator: a_{oracle} is produced from the designated evidence $\mathcal{E}$, whereas $\hat{a}$ is produced from the context returned by a memory pipeline. This design controls generator variation and focuses the comparison on the consequences of each pipeline's stored and retrieved context. It does not, however, prove that every residual gap is caused by a single architectural mechanism; writing prompts, representation choices, indexing, and retrieval policies can all contribute.

Sufficiency of the oracle evidence. The designated set $\mathcal{E}$ contains the sessions used to construct the question and reference. Conditioning the frozen generator on this clean set provides a controlled evidence-complete reference under the benchmark protocol. The sampled evidence-sensitivity filter reduces shortcut cases, but it neither proves that every session is individually indispensable nor guarantees that the generated reference exhausts every valid organization or emphasis. A different grounded response may therefore be equally useful, and the rubric explicitly permits information supported by $\mathcal{E}$ even when it is absent from a_{oracle}.

Robustness of the reference. Every evaluated system is compared with the same per-instance reference, which makes relative comparisons better controlled than independently generated system-specific references. Reference imperfections can nevertheless affect particular systems or answer styles differently, especially for open-ended planning and analysis. We reduce single-judge variance by averaging two judge calls and report their agreement below, while treating human-grounded validity as a remaining limitation.

E.1 Development-Time Human Verification

Prior to large-scale synthesis, two computer science graduate students manually inspected 50 development instances, with 10 instances sampled from each of the five domains. The inspection considered whether each question required evidence across sessions, whether synthesized distractors remained functionally irrelevant despite surface similarity, whether the oracle answer was grounded and sufficiently complete, and whether the judge output was consistent with the evaluation rubric. Feedback from this inspection was used to refine the prompts, rubric, and model allocation before the final generation run. The development log records full agreement between the two annotators on the inspected instances, suggesting that the resulting pipeline did not exhibit obvious construction failures in this sample.

E.2 Human-Checked Distractor Examples

The following development examples illustrate the distinction between semantic similarity and functional relevance that guided the manual checks.

Human Check H1 — Fitness: Lexical Match, Functional Mismatch

Target question. Can you review my fitness journey so far and explain how my body responded as my long runs increased, from the calf lock-ups early on to the later GI issues and taper fatigue?

Semantically similar distractor. "Ran a 2.5-hour field test this morning—the full '2:30 run'—to validate the new terrain navigation stack ... I am seeing weird IMU drift and voltage sags ..."

Shared surface cues. **2.5-hour run**, **fatigue-like behavior**, **fueling**, and adjusting the next "long test."
Why it is irrelevant. The session describes a **rover/navigation system**, not human endurance training. It supplies no evidence about calf cramps, sodium response, carbohydrate intake, gastrointestinal symptoms, taper fatigue, or half-marathon preparation.

Human Check H2 — Mental Wellness: Shared Symptoms, Wrong Subject

Target question. Can you create a "shutdown protocol" I can practice outside therapy that uses the same grounding ideas we discussed for OCD spikes and the overwhelm/fear-of-death moments, and also includes a pre-planned minimal signal plan for therapy, since I cannot gesture or write when I freeze, plus a simple way to record what happened afterward?

Semantically similar distractor. "My rescue dog is great at home ... but whenever we are in certain places, like the vet's office ... he completely shuts down. His body freezes, he will not look at me, move, or respond to treats or cues ..."

Shared surface cues. **Freezing**, **shutdown**, triggers, inability to communicate, and desensitization.
Why it is irrelevant. The session concerns **dog training**, not the user's therapy shutdown. It provides no valid evidence about the user's OCD spikes, fear-of-death episodes, therapy dynamics, grounding practices, or personal signal plan.

E.3 Inter-Judge Consistency

We measured consistency between GPT-5.2-1211-global and GPT-5.4-0305-global on all 1,717 NaiveRAG evaluation outputs. Each judge independently assigned an integer score on the same 1–10 rubric; no score-parsing failures occurred. Table 4 reports both association and absolute agreement statistics.

Pearson and Spearman correlations above 0.84 indicate strong agreement in raw-score movement and rank ordering. Exact agreement on the fine-grained scale is moderate, but 82.94% of pairs differ by at most one point and 93.65% by at most two. These results support the reproducibility of relative scoring between the two GPT-family judges.

E.4 Qualitative Mem0 Failure Cases

We inspected raw Mem0 evaluation records to determine whether its low aggregate scores arose from empty retrieval or malformed judge outputs. The two cases below have non-empty retrieved-memory lists and valid scores from both judges. They therefore provide case-level evidence of lossy or distractor-attracted retrieval rather than an empty-store or parsing failure; they are not a complete implementation audit.

Failure Case E1 — Learning Support: Creative-Writing Leakage

Question. Can you summarize everything I've worked on across my AI assignments, including the buildTree coding question, the chronic kidney disease data description, the slime linear and polynomial regression notebook, the neural network tuning results, and the RNN autocomplete project?

Selected retrieved memories (non-empty).
- "Planning to show Dr. Lin tapping each line of the report ..."
- "Revising Chapter 7 of *Neural Complete* ..."
- "Wrote a short story titled *Equation of a Slime* ..."

Model-output excerpt. "You wrote and structured a creative short story titled *Equation of a Slime*, featuring Dr. Aris Thorne, a biochemist with chronic kidney disease. The narrative is framed as a dataset ..."
Judge outcome. Average score **1.5**; judge scores **[1, 2]**; primary tags `hallucination` and `mixed`.
Diagnosis. Retrieval succeeds mechanically but is attracted to creative-writing memories sharing **CKD**, **slime**, and machine-learning vocabulary. The answer follows this distractor narrative and omits the requested tree implementation, regression, tuning, and RNN evidence.

Failure Case E2 — Finance: Game Context Causes Refusal

Question. Can you create a practical plan for how my wife and I should manage cash flow and liquidity across all these issues, including the money tied up in the car overpayment, paying for my parents' visitor insurance, and keeping enough accessible while deciding between VTIP, a CD, or a short-term bond fund?

Selected retrieved memories (non-empty).
- "Thinking of game cash as an emergency fund against random fees and rent spikes."
- "Prioritizing holding 30–40% of game cash as emergency liquidity in early rounds."
- "Using a ledger to track cash balance, properties owned, estimated rent income, and cash flow."

Model-output excerpt. "All references to financial concerns ... are framed within the context of a **Monopoly-style game** ... Therefore, **no practical real-life financial plan can be created from these memories**."
Judge outcome. Average score **1.0**; judge scores **[1, 1]**; primary tags `refusal` and `contradiction`.
Diagnosis. The retrieved memories share the language of liquidity and cash-flow management but belong to the wrong functional setting. Their presence causes the model to reject the real-world task instead of planning around the overpayment, visitor insurance, and low-risk investment options.

F Noise-Robustness Judge Prompt

Figure 8 shows the full prompt used by both judge models in the evaluation protocol of Section 4. For each instance the prompt is instantiated with the oracle context $\mathcal{E}$, the question q, the reference answer a_{oracle}, and the system-retrieved answer $\hat{a}$.

Items	Pearson	Spearman	QWK	Exact	Within 1	Within 2	Mean $J_2 - J_1$	Mean $\lvert J_2 - J_1 \rvert$
1,717	0.848	0.843	0.832	42.98%	82.94%	93.65%	-0.391	0.824

Table 4: Consistency between the two judge calls on NaiveRAG outputs. QWK denotes quadratic-weighted Cohen's κ. Judge 2 is GPT-5.4-0305-global and is stricter by 0.391 points on average.

The judge proceeds in two steps. Step 1 classifies $\hat{a}$ along five sub-dimensions into severity levels. Step 2 maps these severities onto an integer score $s \in [1, 10]$ via explicit anchors that cap the score at the level of the worst dimension.

G Question-Generation Prompt

Figure 9 shows the system prompt shared across all five domains for the question-generation stage of Section 3.2. The system prompt is concatenated with a domain-specific template that injects the full text of all sessions in the bundle, and the same target count of ten candidate questions per bundle is used for every domain. A representative per-domain template, for the Learning Support domain, is shown in Figure 10. Templates for the remaining four domains share the same structure and differ only in their domain anchors and content guidance.

H Evidence-Sensitivity Judge Prompt

Figure 11 shows the pairwise comparison prompt used by both judges in the evidence-sensitivity filtering stage of Section 3.3. For each candidate question, this prompt is instantiated twice per judge model, once comparing $\hat{a}^{(1)}$ against $\hat{a}^{(2)}$ and once comparing $\hat{a}^{(2)}$ against $\hat{a}^{(3)}$. A question is retained only when all four resulting verdicts are positive.

I Distractor-Synthesis Prompts

The two-stage distractor synthesis procedure of Section 3.4 uses two LLM calls per evidence session. Figure 12 shows the direction-generation prompt that proposes N adversarial directions from the ground-truth session and a curated family list. Figure 13 shows the conversation-synthesis prompt that instantiates one direction into a multi-turn distractor conversation. Both prompts are shared across all five domains and differ only in the family list injected into {family_descriptions}.

Noise-Robustness Judge Prompt

You are an expert judge evaluating how well a memory-augmented QA system tolerates retrieval noise.

 # Setup
The same QA model (temperature=0, deterministic) was asked the same question twice: once with ONLY oracle context (the relevant past sessions), and once with oracle context PLUS distractor sessions (retrieval noise). Because the QA model is deterministic and the only difference is the injected noise, the Reference Answer below is treated as a strong baseline (score = 10). Your job is to score the Noisy Answer on how faithfully it preserves the quality of the Reference Answer.

 # Context (Ground-Truth Oracle Sessions)
{context}

 # Question
{question}

 # Reference Answer (Oracle context, score = 10 by definition)
{answer_oracle}

 # Noisy Answer (Oracle + distractor context, TO BE EVALUATED)
{answer_noisy}

 # Step 1: Assess each of the five sub-dimensions
For each dimension below, classify the Noisy Answer into exactly one severity level: intact (no meaningful difference), mild (a small, bounded issue a reader would still trust), clear (a meaningful issue that affects usefulness), or severe (a failure that would mislead the reader).

1. **Factual Fidelity.** Are the claims consistent with either the Context or the Reference Answer? A claim absent from the Reference but supported by the Context is acceptable. Classify as severe when the Noisy Answer contradicts the Context or the Reference on an important claim.
2. **Completeness Preservation.** Does the Noisy Answer cover a comparable level of question-relevant information from the Context as the Reference does? Judge relative to what the question actually asks. Classify as clear or worse when the Noisy Answer omits or covers less question-relevant information.
3. **Hallucination Absence.** Does the Noisy Answer add information that is NOT in the Reference AND NOT supported by the Context? Content grounded in the Context but absent from the Reference is NOT a hallucination. Penalise fabricated facts, invented specifics, or distractor-sourced content.
4. **Semantic Equivalence.** Would a reader reach a similar conclusion or take a similar action from both answers? Classify as clear or worse when the noise changed the overall meaning, recommendation, or practical takeaway.
5. **Ungrounded Inference.** Does the Noisy Answer infer, assume, or speculate beyond what the Context supports? Distinguish from hallucination: hallucination = fabricated facts, ungrounded inference = fabricated reasoning or assumptions presented as fact.

 # Step 2: Assign the final integer score (1–10)
Use the severity classifications from Step 1. When an answer sits between two anchors, choose the LOWER score, since a severe failure on a single dimension caps the score regardless of strength elsewhere.

- **10.** All five dimensions intact. Factually identical in substance, complete, hallucination-free, same meaning. Differences are purely surface-level.
- **9.** Up to two dimensions are mild, the rest intact. No clear or severe issues, no contradiction, no hallucination, no change in conclusion.
- **8.** Three or more dimensions are mild, but NO dimension is clear or severe. Fully trustworthy, with only minor bounded drift.
- **7.** One dimension is clear, any number of other dimensions may be mild. A single noticeable issue, but core facts and overall conclusion intact.
- **6.** One dimension is clear with bounded impact (a secondary fact missing or a single unsupported inference). Answer remains mostly useful.
- **5.** Two dimensions are clear. Partially reliable.
- **4.** Two or more dimensions are clear AND at least one is severe: a key fact missing, one significant hallucination, or the conclusion has shifted.
- **3.** Multiple severe failures: large omissions, notable hallucinations, or distractor content visibly bleeding into the answer.
- **2.** Mostly wrong. Contradicts the Reference on important claims, distractor content dominates, or the conclusion is fundamentally different.
- **1.** Completely wrong, refuses to answer, or contradicts the core facts entirely.

The final score MUST be an integer in [1, 10] and MUST be consistent with the Step 1 severities.

 # Failure-Mode Tag
Pick the SINGLE best tag for the PRIMARY source of degradation (use "none" only when the score is 10): none, style_only, omission, hallucination, contradiction, ungrounded_inference, meaning_shift, refusal, mixed.

Respond with ONLY a JSON object containing analysis (2-4 sentence comparison), the five sub_dimensions severity labels, the integer score, and the failure_mode tag.

Figure 8: Noise-robustness judge prompt. Two independent judges instantiate this template per instance, and their scores are averaged to obtain the per-instance score s. The severity-then-anchor two-step design ensures that a single severe failure caps s, so that catastrophic hallucinations or contradictions cannot be offset by strength on unaffected dimensions.

Question-Generation System Prompt

You are an expert at creating realistic evaluation questions that a real user would ask an AI assistant after having multiple past conversations.

CRITICAL REQUIREMENTS FOR EVERY QUESTION:

1. REQUIRES ALL SESSIONS (HARD REQUIREMENT): Each question MUST be constructed so that a complete, correct answer needs information from EVERY single provided conversation session. If a question could be answered well using only one session, or only a few sessions, the question is USELESS — reject it and write a different one.

2. DETERMINISTIC AND NON-JUDGMENTAL (HARD REQUIREMENT): The question must have an answer determined by the conversation content itself, not by the model's subjective judgment. Do not ask the model to diagnose, rank, judge, or infer anything that is not directly stated in the conversations. A good question asks for organization, integration, listing, or planning over all sessions — operations that yield stable, reproducible answers grounded directly in conversation content.

3. NO INTERNAL LABELS: NEVER use category names, topic IDs, subreddit names, section numbers, or any organizational labels. The user thinks in terms of the actual content they discussed.

4. USE GENERAL CONTENT KEYWORDS: Questions should contain keywords that capture the general semantics of the conversations and span all sessions. This is critical for retrieval.

5. USE DOMAIN-LEVEL ANCHORS: Include broad domain phrases that span all sessions, such as "across all my conversations", "based on everything we discussed about the AI course assignments".

6. NATURAL USER VOICE: Write as if a real user is asking ("What should I focus on?", "Summarize my progress"), not in analytic terms ("Analyze the intersection of category X and category Y").

GOOD EXAMPLE (study-aid domain): "Summarize my study progress based on everything we discussed about the AI course assignments."

BAD EXAMPLE: "What recurring mistakes did I make across the AI course assignments?" — the conversations do not label anything as a mistake, so different runs would guess differently.

When you write a candidate question, mentally check both hard requirements: does the answer literally need content from EVERY session, and will two runs of the same QA model produce substantively the same answer? Only emit questions that pass both.

Figure 9: System prompt shared across all five domains for evidence-grounded multi-session question generation.

Per-Domain Template (Learning Support)

Below are the COMPLETE conversation logs from a student's AI course tutoring sessions. The student discussed various AI topics across multiple sessions with an AI tutor.

Read ALL sessions carefully. Then generate exactly 10 open-ended questions that the student might realistically ask the AI assistant later, where answering properly REQUIRES information from ALL sessions.

RULES:

- Use domain keywords like "AI assignments", "my AI coursework". The DOMAIN-LEVEL ANCHORS must indicate these conversations are about AI course assignments.
- Use content keywords that capture the general semantics of the conversations (e.g., specific algorithms, concepts, techniques the student discussed).
- NEVER use topic IDs, category labels, or session numbers.
- Write as the student would naturally ask.

--- CONVERSATION SESSIONS ---
{full_sessions}
--- END SESSIONS ---

Output ONLY a JSON list of 10 question strings. No other text.

Figure 10: Per-domain template for the Learning Support domain. Templates for Finance Guidance, Mental Wellness, Fitness Coaching, and Document Analysis are structurally identical and differ only in the domain anchors and content guidance bullets.

Pairwise Evidence-Sensitivity Judge Prompt

```
You are an expert judge evaluating two AI assistants' answers to a memory-based question.

Two answers were generated with different amounts of conversation context:

  • Answer A: generated with MORE context.
  • Answer B: generated with LESS context ({degraded_level}).

# Context (Ground-truth sessions that Answer A had access to)
{context}

# Question
{question}

# Answer A (more context)
{answer_a}

# Answer B (less context)
{answer_b}

# Task
Is Answer B worse than Answer A?

  • "Yes" = B is noticeably worse, less correct, less complete, or contains more
    hallucination or irrelevant information.
  • "No" = B is roughly as good as A despite less context.

Respond with ONLY a JSON object:
{
    "analysis": "<brief reasoning>",
    "answer_b_is_worse": true or false
}
```

Figure 11: Pairwise judge prompt used in evidence-sensitivity filtering. Two independent judge models render verdicts on $(\hat{a}^{(1)}, \hat{a}^{(2)})$ and $(\hat{a}^{(2)}, \hat{a}^{(3)})$; a question is retained only if all four verdicts agree that the lower-context answer is worse.

Distractor Direction-Generation Prompt

You are designing distractor conversations for an AI memory retrieval benchmark.

Below is a GROUND-TRUTH conversation session between a user and an AI assistant. Your task: generate exactly {n_directions} "distractor direction" descriptions.

Each direction describes a DIFFERENT type of conversation that:

1. SHARES many keywords, phrases, and content from the GT session below.
2. Is OBVIOUSLY about a completely different domain, intent, or scenario.
3. Would REALISTICALLY be retrieved by a semantic search engine when given questions about the GT content (because of content overlap).

Here are some distractor family ideas to draw from (pick diverse ones, do not repeat the same family, and adapt them to the specific content in the GT):

{family_descriptions}

--- GROUND-TRUTH SESSION ---
{gt_session}
--- END GT SESSION ---

For each direction, output:

- domain: what the distractor conversation is about.
- intent: the conversational purpose.
- distinguisher: what makes it obviously NOT about the GT topic.
- shared_content: list of 5-10 keywords/phrases/content from the GT session that the distractor should naturally incorporate.
- instruction: a detailed instruction for generating the distractor conversation, specifying the domain, the roles, and what content to weave in.

Output ONLY a JSON list of {n_directions} direction objects. No other text.

Figure 12: System prompt for the first stage of distractor synthesis. The {family_descriptions} placeholder is filled with a domain-specific list of candidate distractor families (e.g., for Learning Support: science-fiction storytelling, philosophy discussion, creative-writing workshop, psychology tutoring, music theory, cooking, history/sociology).

Distractor Conversation-Synthesis Prompt

```
You are generating a distractor conversation for an AI memory retrieval benchmark.

CONTEXT: Below is a real conversation session (the "ground truth") between a user and an
AI assistant. You must generate a NEW conversation that shares keywords, content, and
information with this GT session but is about a DIFFERENT topic.

--- GROUND-TRUTH SESSION (for keyword reference only) ---
{gt_session}
--- END GT SESSION ---

DIRECTION: {instruction}

REQUIREMENTS:

  1. Generate a multi-turn conversation with exactly {num_turns} exchanges (each exchange =
     one user message + one assistant message).
  2. Naturally weave in these keywords/phrases and content from the GT session:
     {shared_keywords}.
  3. The conversation MUST be about "{domain}" — fill each turn with domain-specific
     terminology, concepts, and details so that a human reader immediately knows this
     is about {domain}, NOT about the GT topic.
  4. The conversation should feel natural and coherent — not forced keyword stuffing. The
     content should appear organically in the new context.
  5. Most importantly, the overall conversation token length should be the same as the GT
     session.

OUTPUT FORMAT:
A JSON list of message objects, alternating user and assistant:

[
  {"role": "user",      "content": "..."},
  {"role": "assistant", "content": "..."},
  ...
]

Output ONLY the JSON list. No other text.
```

Figure 13: System prompt for the second stage of distractor synthesis. Each direction produced in the first stage is instantiated M times with independent turn counts $\tau \sim \text{Uniform}[\tau_{\min}, \tau_{\max}]$, yielding $N \cdot M$ distractor conversations per evidence session.