MemTrace: Probing What Final Accuracy Misses in Long-Term Memory

Anonymous ACL submission

Abstract

LLM agents increasingly maintain long-term memory of user facts across sessions. Yet such memory is usually evaluated by aggregating accuracy over question rows or episodes. Because it scores question rows independently, even when several probe one fact, it cannot show how that fact behaves as conditions change. We introduce MemTrace, a benchmark whose unit of measurement is the knowledge point—a single typed fact about the user, not an individual question. MemTrace probes each fact along three controlled dimensions: memory age (how many sessions ago it appeared in the history), question type (current, earlier, or trajectory of change), and evidence condition (present, missing, or contradicted by a false premise). Evaluating 13 memory-system configurations across four paradigms, we find that similar pooled accuracy hides different failures: recovering a fact's current and earlier states does not imply tracking how it changed, and safe abstention does not imply correcting a false premise. The dominant bottleneck is evidence use, not retrieval: when systems fail, the evidence was retrievable 10× more often than it was missing—so improving memory depends on using reachable evidence, not on more storage or retrieval. The code is anonymously available at: https://anonymous.4open.science/r/MemTraceBench-2FB4.

1 Introduction

Large language models (LLMs) are moving from single-turn assistants to persistent agents that interact with users across many sessions (Deshpande et al., 2025; Shen et al., 2026; Deng et al., 2026). In this setting, memory is not only a matter of recalling isolated facts. A useful system (Hu et al., 2026; Huang et al., 2026) is expected to remember user-specific information, update it when the user changes state, and keep its answers coherent as goals and preferences evolve (Deshpande et al., 2025; Zhang et al., 2026; Chen et al., 2026b). This need is now being addressed by several lines of work, including long-context models that keep more history in the input (Gemini Team, Google DeepMind, 2025; OpenAI, 2025; Qwen Team, 2025b), retrieval-augmented systems that retrieve evidence at inference time (Asai et al., 2024; Jimenez Gutierrez et al., 2024), external-memory systems that maintain persistent memory stores (Liu et al., 2026; Zhong et al., 2023), and agentic-memory architectures that use policies or agents to manage memories across interactions (Wang and Chen, 2025; Yue et al., 2026).

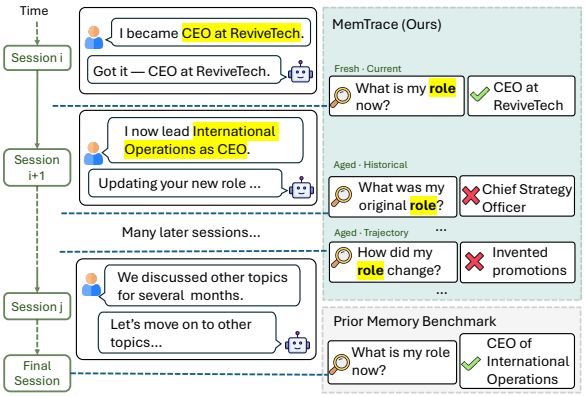


Figure 1: A pooled QA view can mark one final answer as correct while hiding failures on the same knowledge point. In this example, a system answers the current role correctly but fails when the same role fact is probed as aged historical and trajectory questions.

Many current long-term memory benchmarks aggregate accuracy over question rows or interaction episodes (Yen et al., 2025; Tavakoli et al., 2026), which hides variation across different queries. This aggregation hides where a system fails, but it is only the surface problem. The deeper one is the unit of measurement: accuracy is scored per question row or per interaction episode, so questions that probe one underlying fact are treated as independent items. As a result, no score over that unit can hold a fact fixed and ask how it behaves as conditions change. These conditions matter in practice:

a system may recall a fact when recent but forget it after many sessions, answer its current state but not how it changed, or refuse an unmentioned fact yet accept a false premise about it. Capturing such contrasts requires fixing the fact and varying the conditions around it. Figure 1 gives an example.

Distinguishing these failures is crucial because they reflect how memory systems are used in practice. Users ask about the current state of a fact, refer back to earlier states, and ask how a fact has changed over time. They also ask questions whose evidence is absent, where the system should abstain, or whose premise conflicts with remembered information, where the system should reject the premise. Benchmarks that aggregate across these dimensions cannot show which behavior breaks, even when two systems have similar pooled scores.

We introduce MemTrace, a benchmark whose unit of measurement is the *knowledge point*: a single typed fact about the user, rather than the individual question. For each fact, MemTrace constructs repeated probes along controlled dimensions. *Memory age* measures how many sessions have passed since the fact first appeared in history. *Question type* asks for the current state, an earlier state, or the trajectory of change over time. *Evidence condition* controls whether the relevant evidence is present, missing, or contradicted by a false premise. Together, these dimensions instantiate MemTrace as a benchmark of 835 typed knowledge points from 20 users, expanded into 15,422 question rows and over 200,000 scored answers.

Each user fact in MemTrace is repeatedly probed, which lets us ask whether memory persists as sessions accumulate, whether systems track the state and evolution of a fact, and whether they behave safely when evidence is missing or conflicting. We also diagnose whether remaining errors come from unreachable evidence or from reachable evidence that is not used. We evaluate 13 memory-system configurations across four paradigms on MemTrace; our key findings are as follows: *(1) Performance varies systematically across memory age, question type, and evidence condition*. Long-context systems answer recent facts well, but lose accuracy as facts age, especially on trajectory questions. RAG systems, including graph-based retrieval, handle current and earlier-state questions better than questions that require tracking change over time. Some external-memory systems decline almost all questions about facts that were never mentioned, yet rarely answer correctly when the prompt contains a false premise. *(2) Across systems, the dominant remaining gap is evidence use rather than retrieval:* when systems fail, the evidence was already retrievable about 10× more often than it was missing. Our contributions are:

- We introduce **MemTrace**, a knowledge-point benchmark built around three probing dimensions central to long-term memory evaluation: memory age, question type, and evidence condition. These dimensions test retention, varied fact queries, and safe behavior under missing or conflicting evidence.
- We evaluate 13 memory-system configurations across four paradigms and show that systems with similar pooled scores fail differently. In particular, trajectory questions expose a broad weakness: systems that recover current or earlier states of a fact can still fail when asked how it changes over time.
- We provide diagnostic analysis of where memory failures arise. Across systems, the main bottleneck is evidence use rather than retrieval: failures come 10× more often from unused evidence than from unreachable evidence.

2 Related Work

Memory Architectures. Persistent LLM agents have motivated several memory architectures. Long-context models read prior interactions directly from the prompt. Retrieval-augmented systems index and retrieve external evidence before generation (Lewis et al., 2020; Robertson and Zaragoza, 2009; Jimenez Gutierrez et al., 2024). Explicit or agentic memory systems maintain dedicated stores and policies for writing, updating, and retrieving memories. Some systems organize memory as an explicit agent state (Park et al., 2023; Packer et al., 2023; Chhikara et al., 2025); others study lightweight memory stores (Liu et al., 2026; Shu et al., 2026; Xu et al., 2025) or memory-management agents (Wang and Chen, 2025; Yue et al., 2026). This architectural diversity makes final-answer accuracy an incomplete way to compare memory systems.

Memory Benchmarks. Memory evaluation spans long-context stress tests (Bai et al., 2024; Zhang et al., 2024; Hsieh et al., 2024), newer long-context suites (Bai et al., 2025), and long-term conversational memory (Xu et al., 2022; Maharana et al., 2024; Wu et al., 2025). Other

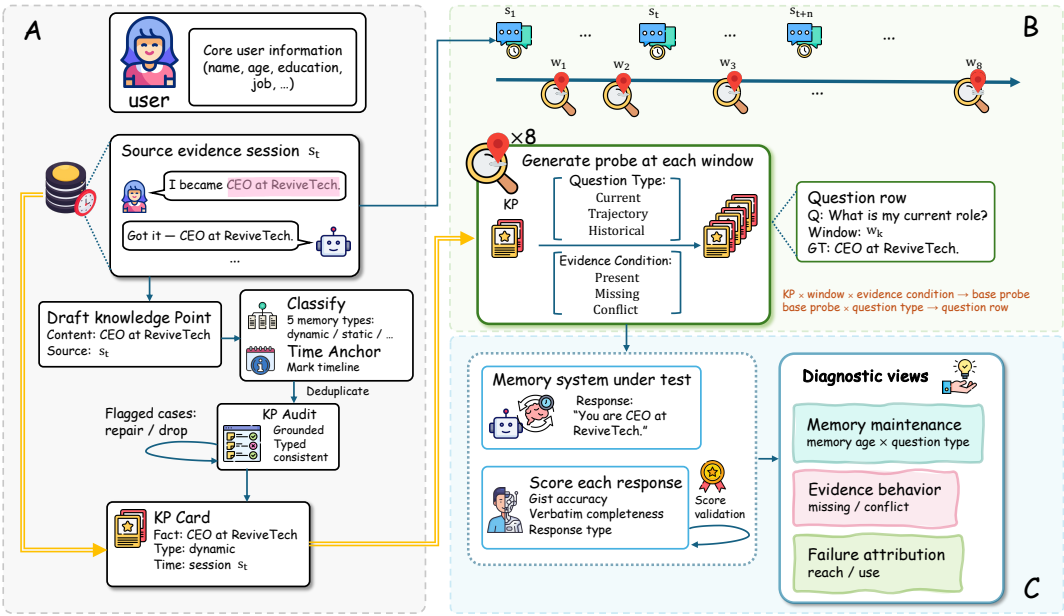


Figure 2: Construction and evaluation schematic for MemTrace. **(A)** Source sessions are converted into typed knowledge points with session anchors and quality checks. **(B)** Probe construction pairs each knowledge point with a memory window and evidence condition, then expands the base probe by question type into question rows. **(C)** Memory-system responses are scored and summarized into diagnostic views of memory maintenance, evidence-condition behavior, and failure attribution.

benchmarks focus on personalization (Jiang et al., 2025; Bian et al., 2026), agent memory (Hu et al., 2025; Tavakoli et al., 2026), dynamic profiles and evolving preferences (Li et al., 2026; Uddin et al., 2026), stale or hallucinated memories (Chao et al., 2026; Tan et al., 2025), and missing or conflicting evidence (Chen et al., 2026a; Ai et al., 2026). These benchmarks broaden memory-system evaluation, but usually score question rows or interaction outcomes and aggregate across them. MemTrace uses knowledge points as the analysis unit. Each fact is probed across windows, question types, and evidence conditions, so behavior is measured on the knowledge point rather than averaged across independent question rows.

3 The MemTrace Benchmark

Different from existing benchmarks that mainly evaluate isolated question rows or aggregate corpus-level accuracy, MemTrace uses the *knowledge point* as the unit of measurement. A knowledge point is a typed fact about the user. For each knowledge point, we construct repeated probes that vary memory age, question type, and evidence condition while keeping the underlying fact fixed. This design allows us to test three behaviors that pooled QA scores usually merge: whether a fact remains usable as sessions accumulate, whether the system can answer current, historical, and trajectory questions about the same fact, and whether it behaves appropriately when evidence is present, missing, or contradicted. Figure 2 shows the full construction and evaluation flow: source sessions are converted into knowledge points, expanded into controlled probes, scored through memory-system responses, and summarized as diagnostic views. Table 1 positions this protocol against representative long-term and personalized-memory benchmarks (Maharana et al., 2024; Wu et al., 2025; Jiang et al., 2025; Chen et al., 2026a; Hu et al., 2025).

3.1 Data Source and Knowledge Points

To evaluate a system along controlled dimensions, a per-fact protocol requires source data with two properties. First, facts must be grounded in multi-session user histories and anchored to specific sessions, so memory age can be defined. Second, the data must include distractor items for constructing missing-evidence and conflict probes, so evidence conditions can be tested. HaluMem-Medium (Chen et al., 2026a) provides both properties. HaluMem evaluates whether memory systems hallucinate during extraction, updating, and question answering. However, its examples are not organized around a fixed knowledge point queried under multiple conditions. We therefore use HaluMem-Medium as source data and transform its histories, memory points, and diagnostic questions into knowledge-point probes. Substantive knowledge points are static facts, dynamic facts with earlier and updated

Feature	LoCoMo	LongMemEval	PersonaMem	HaluMem	MemoryAgentBench	**MemTrace**
General long-memory axes						
Multi-session users	✓	✓	✓	✓	✓	✓
Personal facts/profiles	✓	○	✓	✓	○	✓
Updates / change	○	✓	✓	✓	✓	✓
Conflict / missing evidence	–	○	–	✓	✓	✓
Knowledge-point protocol and attribution						
Knowledge point as unit	–	–	–	–	–	✓
Same fact across age windows	–	–	–	○	–	✓
Same fact across question types	–	–	–	–	–	✓
Evidence-condition probes	–	○	–	✓	✓	✓
Failure attribution: reach vs. use	–	–	–	–	–	✓

Table 1: Benchmark-design comparison. Rows are protocol axes, not quality judgments; ✓ = central; ○ = related; – = not part of the main protocol.

Knowledge point group	Count	%
Static	348	41.7
Dynamic	213	25.5
Preference	74	8.9
Conflict distractor	100	12.0
Boundary distractor	100	12.0
Total	835	100.0

Table 2: Distribution of knowledge points in MemTrace: static, dynamic, and preference are substantive facts; conflict and boundary are distractors.

states, or preference facts. Conflict and boundary distractors come from HaluMem's diagnostic questions and are reformulated for the per-fact protocol.

The benchmark covers 20 users and 835 typed knowledge points, expanded into 5,677 base probes, 15,422 question rows, and 200,453 scored answers across 13 memory-system configurations spanning four paradigms (Appendix A). As shown in Table 2, 635 knowledge points are substantive user facts and 200 are distractor knowledge points used for conflict and boundary probes. Each user contributes 35–48 knowledge points (mean 41.8).

3.2 Probe Construction

Each probe in MemTrace reflects how long-term memory systems are queried in multi-session conversations. Rather than treating evaluation as a flat question-answering task, we fix the knowledge point and evaluate it along three dimensions that aggregate QA scores usually merge: how long the system can retain it (*Memory age*), whether the system can flexibly utilize it across different question contexts (*Question type*), and whether the system can handle it safely under unmentioned or misleading conditions (*Evidence condition*).

Formally, a *base probe* pairs a knowledge point, memory window W_w, and evidence condition, then expands into question rows. A memory window is an evaluation checkpoint; its *prefix* is the conversation history shown to the system. If W_w ends after session t_{eval}, the system receives sessions with indices $< t_{\text{eval}}$. This differs from boundary probes, which query unmentioned facts.

Memory age. To track how memory persists as the conversation grows, we evaluate each user's history at eight chronological checkpoints (rather than on disjoint session blocks) to map a continuous retention trace. The evaluation checkpoints are chosen to span from early to saturated histories, with extra windows anchored near dynamic updates to probe pre- and post-update states. A knowledge point is only probed after its initial appearance. If it first appears in session t_{source}, its memory age at W_w is $t_{\text{eval}} - t_{\text{source}}$ sessions.

Question type. Question type tests whether the same fact supports different uses. Users ask beyond the current state of a fact. Each substantive knowledge point is instantiated as a Current question (what is the current state?), a Historical question (what was the state at an earlier point?), and a Trajectory question (how has the state changed over time?); distractor knowledge points, which drive the evidence-condition probes described next, carry only Current and Historical questions.

Evidence condition. Evidence condition tests how systems respond when evidence is absent or conflicting. Real conversations include queries about unmentioned facts (missing evidence) and queries whose premise contradicts memory (conflicting evidence). The standard probes provide the evidence-present case. We therefore expand distractor knowledge points into two additional probe families: *boundary probes*, which query unmentioned facts, and *conflict probes*, which assert a false premise that contradicts memory. Both families are adapted from HaluMem's diagnostic taxonomy and aligned with the per-fact unit.

System	Current			Historical			Trajectory			Overall		
	Fresh	Saturated	ΔForget	Fresh	Saturated	ΔForget	Fresh	Saturated	ΔForget	Fresh	Saturated	ΔForget
Long Context												
Gemini-3-Flash	66.0	39.7	26.2	59.6	47.1	12.5	31.5	11.0	20.5	52.3	32.6	19.7
Qwen3.5-35B	68.8	32.7	36.1	66.2	44.2	21.9	**49.0**	6.7	42.3	**61.3**	27.9	33.5
GPT-5-nano	67.8	34.8	33.0	62.4	41.9	20.5	38.4	6.5	31.9	56.2	27.7	28.5
RAG												
HippoRAG-v2	**69.7**	**45.4**	24.4	**67.6**	**50.9**	16.7	33.0	13.4	19.7	56.8	**36.5**	20.2
text-emb-3-small	65.7	43.4	22.3	59.3	47.5	11.8	19.7	9.4	10.3	48.2	33.4	14.8
Qwen3-Emb	61.2	41.1	20.1	59.2	48.2	11.0	15.7	8.6	7.1	45.4	32.6	12.7
BM25	61.2	39.9	21.3	56.2	46.9	9.3	18.3	9.2	9.1	45.2	32.0	13.2
External Memory												
SimpleMem	60.4	43.8	16.6	58.0	44.8	13.2	40.1	17.3	22.8	52.9	35.3	17.5
SimpleMem (gpt-4.1-mini)	63.7	40.9	22.8	65.7	41.2	24.6	42.7	13.4	29.4	57.4	31.8	25.6
REMem	62.1	43.1	18.9	53.7	43.1	10.6	24.1	12.6	11.5	46.6	33.0	13.7
AMem	55.3	42.0	13.3	47.7	39.3	8.5	15.6	10.9	4.7	39.5	30.7	8.8
Mem0	50.1	31.5	18.7	37.5	29.7	7.7	9.6	8.2	1.5	32.4	23.1	9.3
Agentic Memory												
Mem-T	58.9	40.4	18.5	59.8	47.3	12.5	47.2	**19.8**	27.4	55.3	35.8	19.5
MIRIX	56.4	26.4	30.0	40.5	23.8	16.7	12.4	4.8	7.6	36.4	18.3	18.1
MIRIX (gpt-4.1-mini)	53.2	36.2	17.0	69.3	46.3	23.0	29.3	16.1	13.2	50.6	32.9	17.7

Table 3: **Memory maintenance by question type.** Fresh averages Gist over checkpoints W1 and W2, Saturated averages Gist over checkpoints W7 and W8, and ΔForget is their gap computed before rounding, all in percentage points. Score cells are shaded within each column; bold marks each Fresh/Saturated leader among the main rows. Gray rows are paper-native backbone sensitivity rows and are not counted as additional main systems.

3.3 Metrics

A single accuracy metric would conflate distinct response behaviors. Each response is therefore assigned a score tuple (g, v, r). Here, g is binary *Gist accuracy*, which captures semantic correctness; v is continuous *Verbatim completeness* in $[0, 1]$, which checks whether canonical answer tokens appear; and r is *Response type*, which supports diagnostics over correct answers, abstentions, and hallucinations. GPT-4o is the primary judge; Gemini-3-Flash checks reliability on a stratified 200-probe sample under the same rubric (Appendix D).

3.4 Diagnostic Views

We report three diagnostic views instead of one pooled score: memory maintenance, evidence-condition behavior, and reach/use attribution.

Memory maintenance. For each knowledge point and question type, we aggregate Gist accuracy across all eight windows to ask whether memory persists as the conversation grows. We report this in two ways. The first is the full retention trace $W_1, \ldots, W_8$, visualized as a per-system retention curve. The second is three scalar summaries: Fresh accuracy (W_1, W_2), measuring recent access; Saturated accuracy (W_7, W_8), measuring aged-memory retention; and ΔForget, the gap between the two. ΔForget is a diagnostic gap, not a standalone leaderboard: a small value can mean either stable retention or uniformly low endpoint accuracy.

Evidence-condition behavior. For each system, we aggregate response-type rates over distractor probes: correct boundary refusal, correct conflict resolution, and hallucination on both. Safe abstention and false-premise rejection differ, but pooled accuracy treats both as wrong.

Failure attribution. When a system answers incorrectly, we ask whether the gold evidence was unreachable or reachable but unused. This diagnostic runs on dedicated samples rather than the full benchmark and has three components.

Oracle on hard failures. We supply gold memory evidence directly to probes that all systems respond incorrectly and then re-score. The lift over the production baseline tests whether these probes can be solved when the needed evidence is explicit. This open-book check is not a retrieval setting; it only checks whether the hard probes are unanswerable.

Reach-vs.-use replay. On a broader sample, a simple Text-emb-3-small retriever checks whether it can reach the source session containing the gold evidence. We combine this reach signal with the original production answer to separate reach misses from reached-but-unsolved cases.

Oracle on retriever-reached but unsolved probes. We then rerun the open-book oracle on the retriever-reached but unsolved cases. If accuracy rises, the error is recoverable with explicit evidence; if not, the case may remain unsolved even with evidence.

4 Experiment and Findings

This section evaluates MemTrace along the diagnostic views defined in section 3.4. We then report memory maintenance across sessions and question types, behavior under missing and conflicting evi-

dence, and failure attribution by retrieval reach and evidence use. We close with the main implications for long-term memory system design.

4.1 Experiment Setting

We evaluate 13 memory-system configurations across four paradigms. **(a) Long-context configurations** include Qwen3.5-35B (Qwen Team, 2025b), Gemini-3-Flash (Gemini Team, Google DeepMind, 2025), and GPT-5-nano (OpenAI, 2025); they read visible history directly in the model context. **(b) RAG configurations** include BM25 (Robertson and Zaragoza, 2009), Text-emb-3-small (OpenAI, 2024), Qwen3-Emb (Qwen Team, 2025a), and HippoRAG-v2 (Jimenez Gutierrez et al., 2024); they retrieve memory evidence before generation. **(c) External-memory configurations** include Mem0 (Chhikara et al., 2025), SimpleMem (Liu et al., 2026), REMem (Shu et al., 2026), and AMem (Xu et al., 2025); they maintain memory stores. **(d) Agentic-memory configurations** include MIRIX (Wang and Chen, 2025) and Mem-T (Yue et al., 2026); they use policy-driven or multi-agent memory management. Appendix B gives the full configuration for all 13 systems.

For long-context models, each probe is answered by prompting the model with the visible session prefix for that window. For the other configurations, the memory state for each checkpoint is built or updated only from the same visible prefix; future sessions are not available to storage, retrieval, or generation. The system then returns retrieved or stored memory evidence for the probe. We pass that evidence to a shared gpt-4o-mini answer generator with the same answer prompt, so the main comparison changes the memory mechanism while holding the final generator fixed. We also report paper-native gpt-4.1-mini sensitivity rows for SimpleMem and MIRIX where the original systems specify that backbone. Unless otherwise stated, results use Gist accuracy as the correctness signal. The following subsections use this setup to test the three diagnostic views introduced in Section 3.4.

4.2 Memory Maintenance Across Sessions and Question Types

Pooled accuracy hides when and how memory fails. A pooled final-accuracy score removes the two variables that define long-term memory use: when a knowledge point is queried and what the question asks the system to do with it. MemTrace therefore reports Fresh access, averaged over W1 and W2; Saturated retention, averaged over W7 and W8; and ΔForget, their gap. We compute these summaries separately for current, historical, and trajectory questions. Table 3 reports these summaries for all main configurations, and Figure 3 shows the traces from W1 through W8. These results show why a single leaderboard is lossy. On overall Saturated Gist, HippoRAG-v2 has the strongest endpoint (36.5%), but the leader changes by question type: HippoRAG-v2 leads saturated current and historical questions (45.4% and 50.9%), while Mem-T leads saturated trajectory questions (19.8%). Trajectory is therefore not just harder lookup; it asks for a different memory behavior. Long-context models fail differently: Qwen3.5-35B and GPT-5-nano have high Fresh trajectory scores (49.0% and 38.4%), but fall to 6.7% and 6.5% when saturated. Appendix C reports the corresponding bootstrap rank checks in Table 11. This trajectory drop suggests that relevant evidence alone is not enough. A trajectory answer must connect multiple states of the same knowledge point, identify their temporal order, and express the update. Long-context models receive the visible conversation history, but as it grows, old and new mentions compete with later sessions; the model may answer a local current or historical query while failing to organize the temporal relation. RAG systems show a complementary limitation: HippoRAG-v2 leads saturated current and historical questions, but reaches only 13.4% on saturated trajectory. In both cases, the hard part is using multiple states as a temporal update trace, not recalling one state. A small ΔForget can still hide low performance at both endpoints, so we interpret the gap together with Fresh and Saturated.

4.3 Behavior Under Missing and Conflicting Evidence

Abstaining is not resolving. Reliable memory systems must handle missing evidence and false premises, not only recall supported facts. Boundary probes ask about facts absent from the history and should be refused. Conflict probes contain a false premise and require correction from memory. Table 4 and Figure 4 compare the two axes. These two behaviors are empirically separate. Mem0, AMem, and REMem are boundary-safe: their boundary abstention rates are 99.3%, 97.4%, and 94.0%, but they rarely resolve conflict probes. Their conflict Gist scores are 14.6%, 20.1%, and 35.1%. Their conflict failures are mostly abstentions rather than fabrications: conflict hallucination is 2.7%, 1.1%,

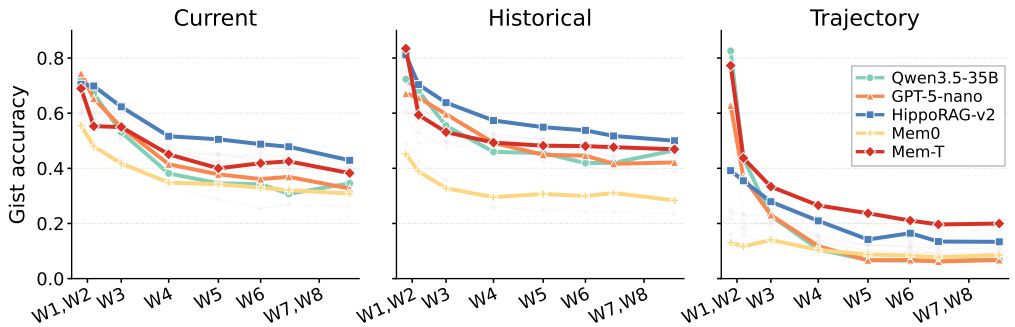


Figure 3: **Memory-age traces.** Gist traces from checkpoint W1 through W8 for the main configurations. Each panel fixes one question type: current state, earlier state, or trajectory of change. Gray lines show the remaining systems, and colored lines highlight representative configurations used in the text. Endpoints are summarized in Table 3.

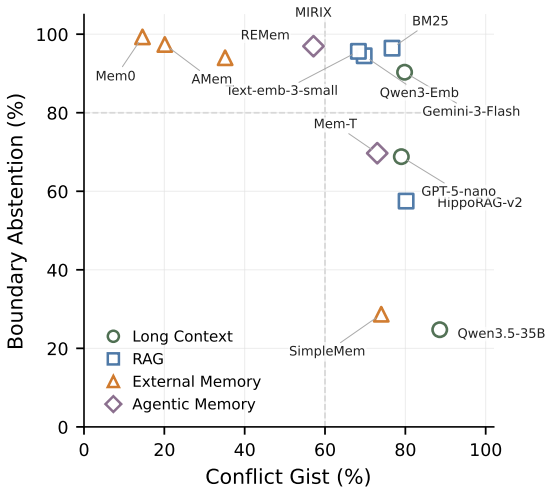


Figure 4: **Conflict Gist versus boundary abstention by system.** The two axes separate missing-evidence refusal from false-premise resolution; dashed guides mark the high-boundary and high-conflict regions.

	Conflict			Boundary		
System	Gist	Abst.	Hallu.	Gist	Abst.	Hallu.
Long Context						
Qwen3.5-35B	88.6	8.4	3.9	89.6	24.8	10.4
Gemini-3-Flash	79.8	15.2	6.4	96.0	90.3	3.8
GPT-5-nano	79.0	16.9	5.7	95.1	68.8	4.9
RAG						
HippoRAG-v2	80.2	11.7	11.1	94.4	57.5	5.2
BM25	76.7	20.3	3.6	96.6	96.4	3.4
Qwen3-Emb	69.8	23.1	8.0	95.8	94.5	4.2
Text-emb-3-small	68.3	25.9	6.9	96.4	95.6	3.6
External Memory						
SimpleMem	74.0	13.9	12.6	91.6	28.7	7.6
REMem	35.1	61.8	4.9	94.5	94.0	3.6
AMem	20.1	78.3	1.1	97.4	97.4	2.6
Mem0	14.6	82.5	2.7	99.3	99.3	0.7
Agentic Memory						
Mem-T	73.0	18.3	11.4	91.1	69.7	8.6
MIRIX	57.1	27.5	18.1	97.1	96.9	2.9

Table 4: **Conflict and boundary profile.** For both conflict and boundary probes we report Gist accuracy, abstention rate (Abst.), and hallucination rate (Hallu.). On conflict probes, Gist measures whether the system resolves a false premise using memory; on boundary probes, abstention measures safe refusal when the requested fact is missing. All values are percentages.

and 4.9%. This pattern suggests an evidence-use failure at the memory-to-answer interface, not a lack of stored memory. Under the main answer backbone, boundary probes can be handled by abstention when support is missing, while conflict probes require the generator to use memory against the user's false premise. Relevant memory evidence is present, but the system must first detect that the premise contradicts memory and then convert the memory into a correction. Thus, these systems appear safe on missing evidence while often treating contradicted evidence as if it were absent. Appendix E.2 reports answer-backbone ablations showing that conflict resolution is sensitive to how memory evidence is passed to the generator.

4.4 Failure Attribution: Retrieval Reach vs. Evidence Use

Reaching evidence is not the same as using it. We analyze failure origin in three steps. First, we check whether hard probes can be answered when gold evidence is given directly. Second, we check whether a simple retriever can reach the gold evidence. Third, for cases where evidence is reachable but the system still fails, we give the gold evidence and test whether the answer can be recovered.

We first analyze 120 hard probes: 40 all-systems-fail trajectory probes, 40 dynamic-changed trajectory probes, and 40 hard current/historical control probes. We supply gold memory evidence directly to the generator, an open-book setting that tests answerability rather than retrieval. Oracle Gist rises to 80–85% in every bucket, while production baselines stay between 0% and 33.8% (Figure 5). Thus many hard probes are answerable with the right evidence; we next ask whether that evidence was unreachable or reachable but unused.

To check whether evidence is reachable, we run

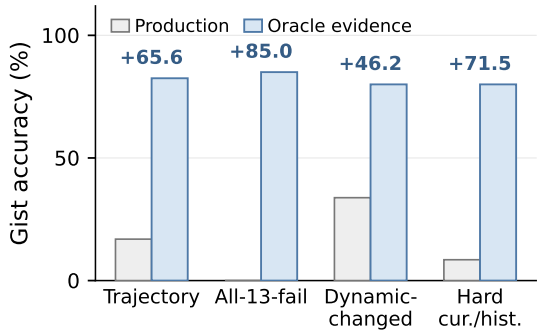


Figure 5: **Failure-attribution decomposition.** The oracle component gives the gold evidence directly to test whether hard probes are answerable; it is not a retrieval method. The retrieval replay splits failures into reach vs. reached-but-unsolved.

System	$R{=}1, U{=}0$	$P(U{=}0 \mid R{=}1)$	Oracle Gist
Long Context			
Qwen3.5-35B	232/279	83.2	81.5
Gemini-3-Flash	224/279	80.3	80.4
GPT-5-nano	225/279	80.6	80.9
RAG			
HippoRAG-v2	193/279	69.2	83.4
BM25	231/279	82.8	81.8
Qwen3-Emb	222/279	79.6	81.5
Text-emb-3-small	220/279	78.9	80.9
External Memory			
SimpleMem	208/279	74.6	82.7
REMem	211/279	75.6	83.9
AMem	237/279	84.9	82.3
Mem0	234/279	83.9	80.8
Agentic Memory			
Mem-T	220/279	78.9	82.3
MIRIX	246/279	88.2	81.7

Table 5: **Retriever-reached but unsolved failures.** The $R{=}1, U{=}0$ column gives the count of retriever-reached but unsolved probes over the 279 retriever-reached probes; $P(U{=}0 \mid R{=}1)$ and Oracle Gist are percentages. Oracle Gist re-supplies the gold memory evidence on each system's $R{=}1, U{=}0$ subset (per-row denominator) and re-scores with gpt-4o-mini backbone.

a 300-probe replay with a simple Text-emb-3-small retriever. We record two signals per probe: $R{=}1$ if the retriever reaches the gold session, and $U{=}1$ if the original production answer solves the probe. Of the 300 probes, 21 are reach misses (7.0%), 220 are retriever-reached but unsolved (73.3%), and 59 are solved (19.7%). Thus the dominant failure case is $R{=}1, U{=}0$ (Table 5). Extending this reach/use replay to all 13 main configurations gives the same picture: $P(U{=}0 \mid R{=}1)$ ranges from 69.2% for HippoRAG-v2 to 88.2% for MIRIX. We then ask whether these reached but unsolved probes become answerable with explicit evidence.

Finally, we rerun the open-book oracle on each main system's $R{=}1, U{=}0$ probes (Oracle Gist column, Table 5). All 13 systems recover to 80.4%–83.9%, a pooled lift of 81.8 percentage points over the 0% baseline. This confirms that most errors are not caused by unreachable evidence. Reaching the gold session is only a weak precondition: it does not guarantee that the relevant span, temporal relation, or conflict signal is exposed to the generator. The harder step is selecting and presenting reachable evidence in a form the generator can use, especially when the answer requires comparing states over time or correcting a false premise.

4.5 Summary and Implications

Taken together, the results suggest that long-term memory failure is not merely a storage or retrieval failure, but a temporally grounded evidence-use failure. Long-context systems read the visible history directly, but their aged-trajectory collapse is consistent with attention dilution and weak temporal organization: old and new mentions must compete with many irrelevant sessions before the model can form an update. RAG systems recover current and historical evidence more effectively, but retrieval exposes local evidence pieces rather than a coherent update path across states. External-memory systems can appear safe on missing evidence because conservative abstention is sufficient there, but conflict probes require active correction: rejecting the false premise while committing to the remembered fact. The reach/use audit connects these patterns: many wrong answers occur after relevant evidence is already reachable. Personalized memory systems therefore need more than larger context windows, stronger retrievers, or safer abstention; they need mechanisms that expose reachable evidence with temporal and conflict structure for multi-session queries.

5 Conclusion

MemTrace recasts long-term memory evaluation around the knowledge point: rather than pooling question rows, it holds a single user fact fixed and varies its memory age, the question asked about it, and the evidence available. This makes visible what an aggregate score cannot: when a fact decays over sessions, when a system recalls a state but not how it changed, and when it abstains safely yet fails to correct a false premise. Across 13 memory-system configurations and four paradigms, systems with similar pooled accuracy diverge along these axes, and most errors persist even when the relevant evidence is already reachable. Long-term memory is therefore limited less by storing or retrieving facts than by using reachable evidence during inference.

Limitations

MemTrace is derived from HaluMem-Medium and covers 20 users from a single source distribution, so its results should not be read as spanning all domains or interaction styles. Our main numbers compare end-to-end configurations rather than isolated memory mechanisms: long-context rows are provider-native models, while most other rows share a gpt-4o-mini generator but differ in their retrievers and embeddings (AMem, for instance, keeps its original all-MiniLM-L6-v2 embedding). Because backbone choice alone can shift a system across much of the leaderboard, cross-paradigm rankings should be read at the configuration level. The reach/use attribution further relies on a single Text-emb-3-small retriever and an open-book oracle over hard probes, so the roughly 10× imbalance is a directional finding whose magnitude depends on the retriever. Lastly, scoring relies on an LLM judge, with reliability checked by a second judge on a stratified sample.

References

Qingyao Ai, Yichen Tang, Changyue Wang, Jianming Long, Weihang Su, and Yiqun Liu. 2026. MemoryBench: A benchmark for memory and continual learning in LLM systems. *Preprint*, arXiv:2510.17281.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. *Preprint*, arXiv:2310.11511. ICLR 2024.

Yushi Bai, Xin Lv, Jiajie Zhang, Hongchang Lyu, Jiankai Tang, Zhidian Huang, Zhengxiao Du, Xiao Liu, Aohan Zeng, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. 2024. LongBench: A bilingual, multitask benchmark for long context understanding. *Preprint*, arXiv:2308.14508. ACL 2024.

Yushi Bai, Shangqing Tu, Jiajie Zhang, Hao Peng, Xiaozhi Cui, Xin Wang, Xin Lv, Shulin Cao, Jiazheng Xu, Lei Liu, Zhen Wang, Chaoyue Lv, Yichuan Zhang, Xu Liu, Xiao Liu, Yang Wang, Ge Zhang, Ka-Hei Wong, Pengcheng Han, and 8 others. 2025. LongBench v2: Towards deeper understanding and reasoning on realistic long-context multitasks. *Preprint*, arXiv:2412.15204. ACL 2025.

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. *Preprint*, arXiv:2601.06966.

Hanxiang Chao, Yihan Bai, Rui Sheng, Tianle Li, and Yushi Sun. 2026. STALE: Can LLM agents know when their memories are no longer valid? *Preprint*, arXiv:2605.06527.

Ding Chen, Simin Niu, Kehang Li, Peng Liu, Xiangping Zheng, Bo Tang, Xinchi Li, Feiyu Xiong, and Zhiyu Li. 2026a. Halumem: Evaluating hallucinations in memory systems of agents. *Preprint*, arXiv:2511.03506.

Tiantian Chen, Jiaqi Lu, Ying Shen, and Lin Zhang. 2026b. ES-MemEval: Benchmarking conversational agents on personalized long-term emotional support. *Preprint*, arXiv:2602.01885.

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

Xinle Deng, Yida Xue, Yijun Chen, Mingjun Mao, Ruobin Zhong, Buqiang Xu, Jizhan Fang, Haoming Xu, Tingwei Wu, Yajing Xu, Shumin Deng, Haofen Wang, Huajun Chen, and Ningyu Zhang. 2026. MobileMem: Evaluating long-horizon memory for language agents in real-world mobile environments. OpenReview. ICLR 2026 Lifelong Agent Workshop.

Darshan Deshpande, Varun Gangal, Hersh Mehta, Anand Kannappan, Rebecca Qian, and Peng Wang. 2025. Memtrack: Evaluating long-term memory and state tracking in multi-platform dynamic agent environments. *Preprint*, arXiv:2510.01353. NeurIPS 2025 SEA Workshop.

Gemini Team, Google DeepMind. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. 2024. RULER: What's the real context size of your long-context language models? *Preprint*, arXiv:2404.06654.

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

Yuyang Hu, Shichun Liu, Yanwei Yue, Guibin Zhang, Boyang Liu, Fangyi Zhu, Jiahang Lin, Honglin Guo, Shihan Dou, Zhiheng Xi, Senjie Jin, Jiejun Tan, Yanbin Yin, Jiongnan Liu, Zeyu Zhang, Zhongxiang Sun, Yutao Zhu, Hao Sun, Boci Peng, and 28 others. 2026. Memory in the age of ai agents. *Preprint*, arXiv:2512.13564.

Wei-Chieh Huang, Weizhi Zhang, Yueqing Liang, Yuanchen Bei, Yankai Chen, Tao Feng, Xinyu Pan, Zhen Tan, Yu Wang, Tianxin Wei, Shanglin Wu, Ruiyao Xu, Liangwei Yang, Rui Yang, Wooseong Yang, Chin-Yuan Yeh, Hanrong Zhang, Haozhen Zhang, Siqi Zhu, and 41 others. 2026. Rethinking memory mechanisms of foundation agents in the second half: A survey. *Preprint*, arXiv:2602.06052.

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

Bernal Jimenez Gutierrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. 2024. Hipporag: Neurobiologically inspired long-term memory for large language models. *Preprint*, arXiv:2405.14831.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Preprint*, arXiv:2005.11401. NeurIPS 2020.

Shuyue Stella Li, Bhargavi Paranjape, Kerem Oktar, Zhongyao Ma, Gelin Zhou, Lin Guan, Na Zhang, Sem Park, Lin Chen, Diyi Yang, Yulia Tsvetkov, and Asli Celikyilmaz. 2026. HorizonBench: Long-horizon personalization with evolving preferences. *Preprint*, arXiv:2604.17283.

Jiaqi Liu, Yaofeng Su, Peng Xia, Siwei Han, Zeyu Zheng, Cihang Xie, Mingyu Ding, and Huaxiu Yao. 2026. SimpleMem: Efficient lifelong memory for LLM agents. *Preprint*, arXiv:2601.02553.

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.

OpenAI. 2024. New embedding models and API updates (text-embedding-3-small). https://openai.com/index/new-embedding-models-and-api-updates/. Model card and API documentation.

OpenAI. 2025. Introducing GPT-5 (gpt-5-nano variant). https://openai.com/index/introducing-gpt-5/. Model release announcement and system card.

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

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. *Preprint*, arXiv:2304.03442. UIST 2023.

Qwen Team. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *Preprint*, arXiv:2506.05176.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389.

Yiting Shen, Kun Li, Wei Zhou, and Songlin Hu. 2026. Mem2actbench: A benchmark for evaluating long-term memory utilization in task-oriented autonomous agents. *Preprint*, arXiv:2601.19935.

Yiheng Shu, Saisri Padmaja Jonnalagedda, Xiang Gao, Bernal Jiménez Gutiérrez, Weijian Qi, Kamalika Das, Huan Sun, and Yu Su. 2026. REMem: Reasoning with episodic memory in language agent. *Preprint*, arXiv:2602.13530. ICLR 2026.

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*, pages 19336–19352. Association for Computational Linguistics.

Mohammad Tavakoli, Alireza Salemi, Carrie Ye, Mohamed Abdalla, Hamed Zamani, and J. Ross Mitchell. 2026. Beyond a million tokens: Benchmarking and enhancing long-term memory in llms. *Preprint*, arXiv:2510.27246.

Md Nayem Uddin, Kumar Shubham, Eduardo Blanco, Chitta Baral, and Gengyu Wang. 2026. From recall to forgetting: Benchmarking long-term memory for personalized agents. *Preprint*, arXiv:2604.20006.

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

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*.

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.

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

Howard Yen, Tianyu Gao, Minmin Hou, Ke Ding, Daniel Fleischer, Peter Izsak, Moshe Wasserblat, and Danqi Chen. 2025. HELMET: How to evaluate long-context language models effectively and thoroughly. *Preprint*, arXiv:2410.02694. ICLR 2025.

Yanwei Yue, Guibin Zhang, Boci Peng, Xuanbo Fan, Jiaxin Guo, Qiankun Li, and Yan Zhang. 2026. Mem-T: Densifying rewards for long-horizon memory agents. *Preprint*, arXiv:2601.23014.

Weizhi Zhang, Xiaokai Wei, Wei-Chieh Huang, Zheng Hui, Chen Wang, Michelle Gong, and Philip S. Yu. 2026. MemoryCD: Benchmarking long-context user memory of LLM agents for lifelong cross-domain personalization. *Preprint*, arXiv:2603.25973.

Xinrong Zhang, Yingfa Chen, Shengding Hu, Zihang Xu, Junhao Chen, Moo Hao, Xu Han, Zhen Thai, Shuo Wang, Zhiyuan Liu, and Maosong Sun. 2024. InfiniteBench: Extending long context evaluation beyond 100k tokens. *Preprint*, arXiv:2402.13718.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2023. MemoryBank: Enhancing large language models with long-term memory. *Preprint*, arXiv:2305.10250.

6 Use of AI Writing Assistance

Consistent with the policy on AI writing assistance, we used AI assistants only for code development and for polishing the wording and grammar of the manuscript. All scientific content—the claims, experimental design, results, and analyses—was conceived, written, reviewed, and verified by the human authors, who take full responsibility for it.

A Benchmark Details

The final benchmark index listed in the artifact manifest contains 20 users, 835 knowledge points, 5,677 base probes, and 15,422 question rows. Paper-facing loaders use the scoring protocol listed in the same manifest.

A.1 Scale and Workload

This subsection reports the benchmark scale and scoring workload referenced from Section 3.1.

Scale		*Workload*	
Users	20	Main cfgs.	13
KPs	835	Paradigms	4
Base probes	5,677	Scored answers	200,453
Question rows	15,422	Windows/user	8

Table 6: Benchmark scale and scoring workload. Base probes are knowledge-point/window probes; question rows are produced after expanding probes by question type and evidence condition. Scored system-answer rows sum the completed 13 main configurations under the paper-facing scoring pass.

A.2 Quality-Control Pipeline

MemTrace treats quality control as a construction check rather than a result metric. The pipeline first preserves grounded source sessions and source memory evidence, then transforms them into typed knowledge points and repeated window-level probes. Automated checks flag candidate grounding, leakage, temporal-anchor, type, and semantic-consistency defects; flagged cases are manually reviewed and corrected before the benchmark index is frozen. The check summary is reported below, and the independent second-judge scoring check is reported in Appendix D.

Audit dimension	Threshold	Pre-fix fail	Post-fix fail
Grounding	$< 5\%$	0.16%	**0.00%**
Answer leakage (token)	$< 10\%$	6.76%	**2.79%**
Answer leakage (semantic)	$< 10\%$	–	**0.00%**
Cross-question differentiation	$< 5\%$	0.00%	**0.00%**
Temporal anchor	$< 15\%$	38.74%	**1.42%**
Type correctness	$< 2\%$	0.00%	**0.00%**
Semantic consistency	$< 5\%$	10.24%	**0.00%**
Format completeness	0%	0.00%	**0.00%**

Table 7: Quality-control check applied during MemTrace construction. Pre-fix shows defects inherited from the HaluMem-Medium-derived draft; Post-fix is after manual review and correction.

B System Configurations

Table 8 lists the 13 main memory-system configurations used in the controlled comparison. The comparison unit is a complete configuration: memory construction, retrieval or access path, update policy, answer prompt, and answer generator. Long-context rows are provider-native; the remaining main rows use the shared gpt-4o-mini answer generator described in Section 4.1.

Mem-T is integrated from the released Mem-T-4B checkpoint with the published MOT-GRPO retrieval policy and hindsight-credit-assignment memory construction. Its hierarchical memory contains working, factual, experiential, and raw stores, all backed by ChromaDB. Mem-T's published in-domain training corpus is LoCoMo; MemTrace is derived from HaluMem-Medium with LoCoMo session lineage, so the released policy may have soft distribution overlap with our evaluation data. We use the frozen checkpoint and do not fine-tune on MemTrace probes.

C Full Results and Failure-Origin Checks

This appendix provides full numeric companions to the compact main-text tables and figures. Values use the main MemTrace adapter configurations unless stated otherwise. It also includes supporting views for endpoints, hallucination, conflict-by-window behavior, abstention precision, and qualitative examples.

System	Paradigm	Backbone	Memory unit	Retrieval	Update mech.	Conflict path	Temporal path
Qwen3.5-35B	Long-Context	native	full context	–	–	partial	partial
Gemini-3-Flash	Long-Context	native	full context	–	–	partial	partial
GPT-5-nano	Long-Context	native	full context	–	–	partial	partial
BM25	RAG	4o-mini	chunk	sparse (BM25)	append	no	no
Text-emb-3-small	RAG	4o-mini	chunk	dense	append	no	no
Qwen3-Emb	RAG	4o-mini	chunk	dense	append	no	no
HippoRAG-v2	RAG	4o-mini	event graph (KG)	hybrid (graph+dense)	append	partial	no
Mem0	External-Memory	4o-mini	fact triple	dense	consolidate	partial	no
SimpleMem	External-Memory	4o-mini	compressed note	dense	consolidate	partial	partial
REMem	External-Memory	4o-mini	event graph (gist+fact)	hybrid (agentic)	consolidate	yes	yes
AMem	External-Memory	4o-mini	note	dense (Chroma)	append	partial	no
MIRIX	Agentic-Memory	4o-mini	6-store scratchpad	hybrid	consolidate	partial	partial
Mem-T	Agentic-Memory	4o-mini	note (Chroma)	dense (RL policy)	consolidate	partial	partial

Table 8: System-design matrix for the 13 evaluated main configurations. Backbone disclosures are used when interpreting cross-system comparisons. Conflict path and temporal path are descriptive implementation labels: "yes" denotes an explicit mechanism or prompt path for that axis, "partial" denotes an indirect or general mechanism, and "no" denotes no dedicated axis-specific mechanism in our implementation. They are not standalone capability claims.

C.1 Full Memory-Age and Evidence-Condition Results

These tables expand the compact main-text results without changing the aggregation convention. Table 9 reports the full Fresh/Saturated/ΔForget profile, Table 10 adds confidence intervals for conflict and boundary probes, and Table 11 shows the rank stability behind the question-type comparison.

C.2 Additional Supporting Views

The following figures and tables preserve supporting results that are too large for the main text. Figure 6 shows the Fresh-to-Saturated endpoints by question type, Figure 7 shows additional memory-window diagnostics, **??** reports abstention precision, and **??** gives representative qualitative probes.

C.3 Failure-Origin Details

The failure-origin analysis has two parts. The oracle block asks whether the answer generator can recover when the relevant memory evidence is supplied directly. The replay block then checks, with a transparent Text-emb-3-small proxy, whether the same evidence was reachable in the original production setting.

Oracle subset	n	Base	Oracle	Lift
Trajectory overall	80	16.9%	82.5%	+65.6
All-13-fail traj.	40	0.0%	85.0%	+85.0
Dynamic-changed traj.	40	33.8%	80.0%	+46.2
Hard cur./hist.	40	8.5%	80.0%	+71.5
Reach/use replay with Text-emb-3-small				
Evidence reach	300	–	93.0%	–
Reach gap ($R = 0$)	300	–	7.0%	[3.8, 10.6]
Use gap ($R = 1, U = 0$)	300	–	73.3%	[67.1, 79.2]
Success ($R = 1, U = 1$)	300	–	19.7%	[14.8, 25.3]
Use-gap range	–	–	69–88%	–

Table 12: Numerical companion to the main oracle/recovery decomposition figure. *Base* is the mean production-system solve rate; *Oracle* is Gist accuracy when relevant memory evidence is supplied directly. The replay rows use the transparent Text-emb-3-small reach proxy: R marks whether the proxy reaches the benchmark source session for the gold answer, and U marks whether the original production answer solves the probe. Lift and range values are percentage points.

D Judge Reliability

MemTrace uses GPT-4o as the primary scoring judge. As a reliability check, we rescore a stratified 200-probe sample with a second LLM judge using the same rubric and JSON schema, then compute Cohen's κ with bootstrap confidence intervals.

E Sensitivity Checks

The main paper reports controlled MemTrace adapter configurations. The checks below vary prompts, paper-native settings, answer backbones, or diagnostic setup to quantify sensitivity around those main rows. They are not counted as additional benchmark systems.

System	Paradigm	Question type	Fresh	Saturated	ΔForget
HippoRAG-v2	RAG	Current	69.7	45.4	24.4
		Historical	67.6	50.9	16.7
		Trajectory	33.0	13.4	19.7
Mem-T	Agentic Memory	Current	58.9	40.4	18.5
		Historical	59.8	47.3	12.5
		Trajectory	47.2	19.8	27.4
SimpleMem	External Memory	Current	60.4	43.8	16.6
		Historical	58.0	44.8	13.2
		Trajectory	40.1	17.3	22.8
text-emb-3-small	RAG	Current	65.7	43.4	22.3
		Historical	59.3	47.5	11.8
		Trajectory	19.7	9.4	10.3
REMem	External Memory	Current	62.1	43.1	18.9
		Historical	53.7	43.1	10.6
		Trajectory	24.1	12.6	11.5
Qwen3-Emb	RAG	Current	61.2	41.1	20.1
		Historical	59.2	48.2	11.0
		Trajectory	15.7	8.6	7.1
Gemini-3-Flash	Long Context	Current	66.0	39.7	26.2
		Historical	59.6	47.1	12.5
		Trajectory	31.5	11.0	20.5
BM25	RAG	Current	61.2	39.9	21.3
		Historical	56.2	46.9	9.3
		Trajectory	18.3	9.2	9.1
AMem	External Memory	Current	55.3	42.0	13.3
		Historical	47.7	39.3	8.5
		Trajectory	15.6	10.9	4.7
Qwen3.5-35B	Long Context	Current	68.8	32.7	36.1
		Historical	66.2	44.2	21.9
		Trajectory	49.0	6.7	42.3
GPT-5-nano	Long Context	Current	67.8	34.8	33.0
		Historical	62.4	41.9	20.5
		Trajectory	38.4	6.5	31.9
Mem0	External Memory	Current	50.1	31.5	18.7
		Historical	37.5	29.7	7.7
		Trajectory	9.6	8.2	1.5
MIRIX	Agentic Memory	Current	56.4	26.4	30.0
		Historical	40.5	23.8	16.7
		Trajectory	12.4	4.8	7.6

Table 9: Full main-adapter Fresh, Saturated, and ΔForget table by question type. Fresh/Saturated Gist is shown as percentages and ΔForget as percentage points. Systems are ordered by mean Saturated Gist in the main leaderboard.

E.1 Configuration and Prompt Sensitivity

This section keeps the main benchmark rows fixed and reports variants that stress prompts, paper-native settings, and matched results. The goal is to make clear which conclusions are stable and which numbers should be read only as configuration checks rather than main-system replacements.

E.2 Backbone and Oracle Sensitivity

We also check whether the conflict and oracle conclusions depend on the answer backbone. These

System	Conflict (W7–W8)			Boundary		
	Gist	Abst	Hallu	Gist	Abst	Hallu
Qwen3.5-35B	88.6 [84.9,92.2]	8.4 [6.4,10.6]	3.9 [1.9,5.9]	89.6 [79.6,97.8]	24.8 [18.9,30.8]	10.4 [2.3,21.3]
HippoRAG-v2	80.2 [73.9,85.9]	11.7 [8.1,15.7]	11.1 [6.6,15.9]	94.4 [88.9,98.4]	57.5 [50.1,64.0]	5.2 [1.1,10.0]
Gemini-3-Flash	79.8 [75.2,84.0]	15.2 [12.4,18.6]	6.4 [3.2,9.9]	96.0 [92.6,98.7]	90.3 [85.9,94.0]	3.8 [1.1,7.1]
GPT-5-nano	79.0 [73.9,83.3]	16.9 [13.6,19.9]	5.7 [3.5,8.1]	95.1 [91.2,98.1]	68.8 [64.2,73.1]	4.9 [1.9,9.1]
BM25	76.7 [72.4,80.6]	20.3 [15.9,24.8]	3.6 [1.9,5.5]	96.6 [92.2,99.5]	96.4 [92.4,99.4]	3.4 [0.6,7.5]
SimpleMem	74.0 [67.9,80.3]	13.9 [9.0,19.2]	12.6 [8.6,17.1]	91.6 [86.4,95.6]	28.7 [21.2,35.9]	7.6 [3.1,12.6]
Mem-T	73.0 [67.9,78.1]	18.3 [15.6,21.2]	11.4 [8.0,15.1]	91.1 [87.5,94.1]	69.7 [65.8,73.8]	8.6 [5.4,12.8]
Qwen3-Emb	69.8 [64.5,74.6]	23.1 [18.2,29.1]	8.0 [5.1,11.1]	95.8 [91.7,99.4]	94.5 [89.7,98.1]	4.2 [0.8,9.2]
Text-emb-3-small	68.3 [62.6,73.8]	25.9 [20.8,31.5]	6.9 [3.6,10.6]	96.4 [92.1,99.6]	95.6 [91.1,98.9]	3.6 [0.4,8.0]
MIRIX	57.1 [51.7,62.4]	27.5 [22.4,32.6]	18.1 [13.5,23.1]	97.1 [94.4,99.2]	96.9 [94.4,99.1]	2.9 [0.8,5.6]
REMem	35.1 [28.5,42.4]	61.8 [55.0,68.4]	4.9 [2.1,8.1]	94.5 [88.0,98.9]	94.0 [90.4,97.2]	3.6 [0.9,6.5]
AMem	20.1 [12.6,27.4]	78.3 [70.6,85.7]	1.1 [0.2,2.1]	97.4 [93.8,99.9]	97.4 [93.4,99.9]	2.6 [0.1,6.3]
Mem0	14.6 [10.1,19.1]	82.5 [77.8,87.9]	2.7 [0.8,4.8]	99.3 [97.9,100.0]	99.3 [97.9,100.0]	0.7 [0.0,2.1]

Table 10: Main-adapter 13-system conflict and boundary table with bootstrap 95% confidence intervals. Cells report percentage [lower, upper]. This is the confidence-interval companion to Table 4.

System	Paradigm	Current Sat. [95% CI]	Historical Sat. [95% CI]	Trajectory Sat. [95% CI]
HippoRAG-v2	RAG	45.4 [40.5,50.2]	50.9 [46.4,55.3]	13.4 [10.4,16.3]
Mem-T	Agentic Memory	40.4 [35.3,45.5]	47.3 [43.8,50.7]	19.8 [16.3,23.3]
SimpleMem	External Memory	43.8 [37.9,49.7]	44.8 [37.9,51.7]	17.3 [13.9,20.8]
text-emb-3-small	RAG	43.4 [39.6,47.2]	47.5 [42.3,52.6]	9.4 [7.0,11.8]
REMem	External Memory	43.1 [39.2,47.1]	43.1 [40.0,46.2]	12.6 [9.7,15.6]
Qwen3-Emb	RAG	41.1 [35.9,46.2]	48.2 [43.9,52.5]	8.6 [6.5,10.7]
Gemini-3-Flash	Long Context	39.7 [36.4,43.0]	47.1 [43.7,50.6]	11.0 [8.7,13.3]
BM25	RAG	39.9 [36.0,43.9]	46.9 [41.7,52.0]	9.2 [6.8,11.6]
AMem	External Memory	42.0 [36.6,47.4]	39.3 [35.0,43.6]	10.9 [7.9,13.8]
Qwen3.5-35B	Long Context	32.7 [28.4,36.9]	44.2 [40.0,48.5]	6.7 [4.7,8.7]
GPT-5-nano	Long Context	34.8 [30.9,38.7]	41.9 [37.0,46.8]	6.5 [4.5,8.5]
Mem0	External Memory	31.5 [27.8,35.1]	29.7 [25.4,34.1]	8.2 [6.3,10.0]
MIRIX	Agentic Memory	26.4 [23.5,29.4]	23.8 [20.3,27.4]	4.8 [3.2,6.4]
Rank-correlation across question types (1,000-fold user bootstrap; from followup_robustness_report.md §11.4)				
Comparison		Median ρ	95% CI	% boots < 0.6
Cur+Hist mean vs Traj		0.473	[0.247, 0.747]	82.1%
Current vs Trajectory		0.516	[0.291, 0.736]	77.8%
Historical vs Trajectory		0.385	[0.192, 0.692]	91.1%

Table 11: Per-system bootstrap 95% confidence intervals on main-adapter Saturated Gist by question type, shown as percentages, and rank-correlation distribution between current/historical and trajectory ranks.

checks swap only the answer-time generator while keeping the rest of the diagnostic setup fixed.

E.3 Evidence-Interface and Actionability Checks

The following checks are diagnostic interventions, not additional benchmark systems. They ask whether changing the evidence interface or using oracle benchmark labels can change the observed failure pattern.

Dimension	n	Cohen's κ	95% bootstrap CI
Gist accuracy	200	0.772	[0.680, 0.858]
Response type	200	0.703	[0.623, 0.785]

Table 13: Second-judge reliability on a stratified 200-probe sample. Agreement is computed between the primary GPT-4o judge and a Gemini-3-Flash cross-judge using the same scoring rubric.

Question type	KP type	n	Gist disag.	Any disag.
Trajectory	Dynamic	24	6	45.8%
Current	Static	23	5	43.5%
Historical	Static	21	3	33.3%
Current	Preference	22	3	27.3%
Historical	Preference	22	4	27.3%
Historical	Dynamic	24	1	25.0%
Current	Dynamic	22	0	18.2%
Trajectory	Preference	19	0	5.3%
Trajectory	Static	23	0	4.3%

Table 14: Second-judge disagreement breakdown on the same 200-probe sample. "Any disagreement" counts probes where the judges disagree on either binary Gist accuracy or response type.

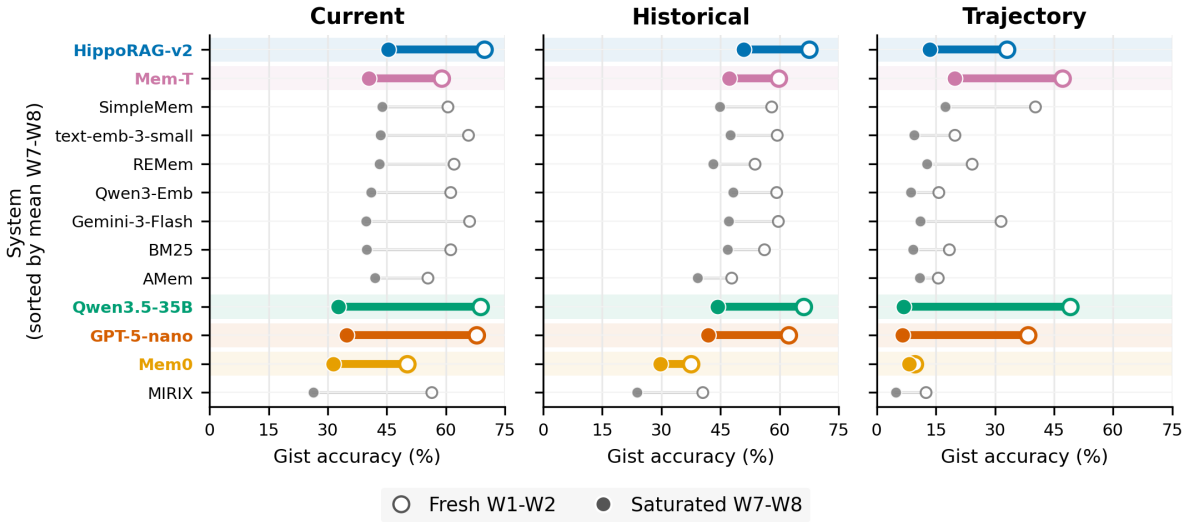


Figure 6: Per-system Fresh-to-Saturated endpoints by question type for the main MemTrace adapter configurations. Filled markers denote Fresh W1–W2 and hollow markers denote Saturated W7–W8. This is the endpoint-level companion to the main memory-age trace in Figure 3.

System	Main adapter	Variant	Coverage	Reading
HippoRAG-v2	main 42.4%	unified 38.3%	20 users	Unified prompt lowers overall Gist.
REMem	main 36.3%; paper config 36.7%	unified 29.9%	20 users	Paper config ties main; unified is more conservative.
SimpleMem	main 39.0%	unified 26.7%; gpt-4.1-mini 37.1%	20 users	Prompt/backbone sensitive.
MIRIX	main 22.5%	gpt-4.1-mini 37.4%	20 users	Paper-native backbone is stronger.
Mem-T	main 40.4%; matched subset 41.0%	unified 42.1%	20 users	Paper config

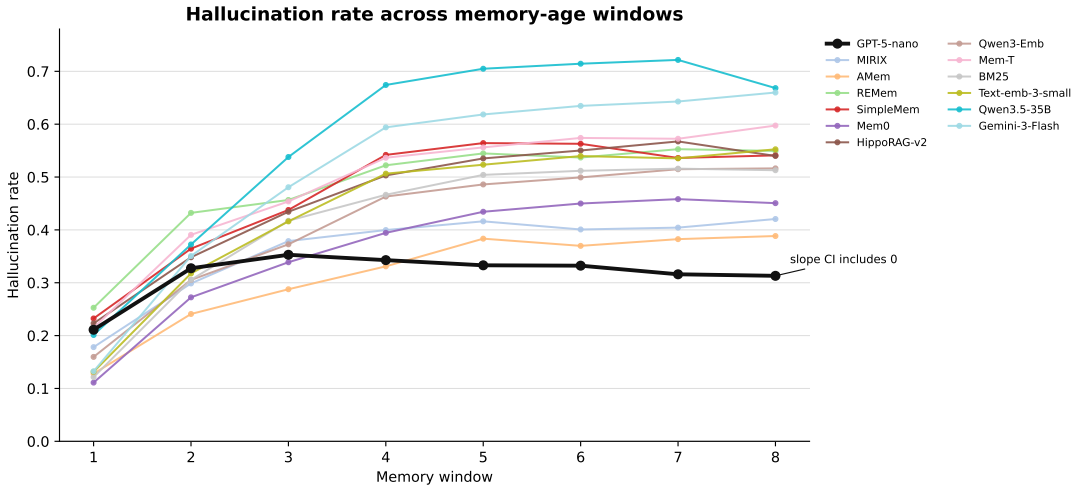
Table 15: Configuration sensitivity checks. Values are overall Gist under the core-KP user-macro mean-over-question-types convention from the ablation inventory, shown as percentages. These variants are not mixed into the main leaderboard.

System	Conflict Gist	Conflict Abst.	Boundary Abst.	Coverage	Reading
HippoRAG-v2	80.2% → 33.8%	11.7% → 62.7%	57.5% → 96.5%	20 users	Native is more conflict-committing.
REMem	35.1% → 13.6%	61.8% → 83.2%	94.0% → 99.9%	20 users	Unified is more abstention-heavy.
SimpleMem	74.0% → 63.4%	13.9% → 33.7%	28.7% → 95.9%	20 users	Unified trades conflict commitment for refusal.
Mem-T	72.0% → 64.7%	20.0% → 32.0%	67.0% → 92.7%	20 users	safer but less committing.

Table 16: Distractor prompt-control checks. Arrows report main adapter → unified where available. The table shows configuration sensitivity for the conflict/boundary results, with rates shown as percentages.

Scope	Variant	n	Gist [95% CI]	Lift [95% CI]	Secondary change
Saturated trajectory	baseline	426	20.0 [14.2, 26.2]	–	Hallu. 71.3; Abst. 8.9
Saturated trajectory	tr+cr+ac	426	31.2 [22.4, 39.9]	+11.1 [+4.4, +18.3]	Hallu. -5.8; Abst. -6.1
Conflict	baseline	800	68.1 [62.0, 73.9]	–	Hallu. 7.0; Abst. 25.9
Conflict	cr	800	73.6 [68.6, 78.2]	+5.5 [+1.4, +9.9]	Hallu. -2.2; Abst. +0.7
Boundary	baseline	800	96.6 [92.7, 99.6]	–	Hallu. 3.4; Abst. 95.8
Boundary	tr	800	97.3 [93.4, 100.0]	+0.6 [+0.0, +1.5]	Hallu. -0.6; Abst. +1.1

Table 17: Diagnostic actionability check using benchmark labels as oracle routing signals on top of the transparent dense-RAG baseline. Variants are not deployable memory methods and are not counted as benchmarked systems; they only test whether the diagnostic axes can guide targeted interventions.



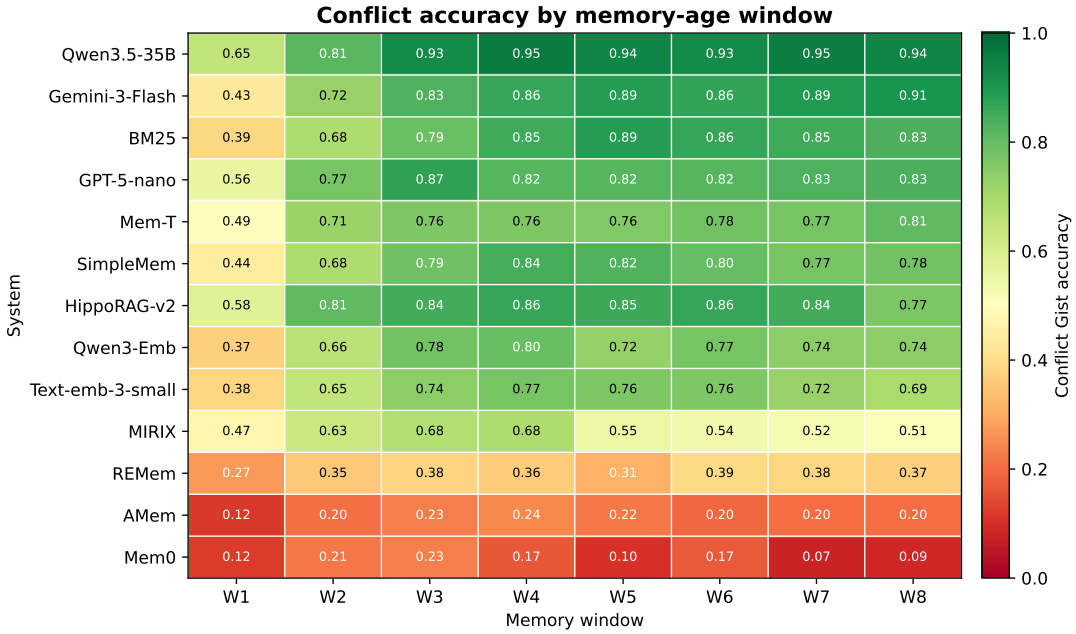


Figure 7: Additional memory-window diagnostics. Top: main-adapter per-system hallucination rate across W1–W8. Bottom: main-adapter Conflict Gist accuracy by system and memory window, expanding the conflict/boundary profile and its confidence-interval companion table.

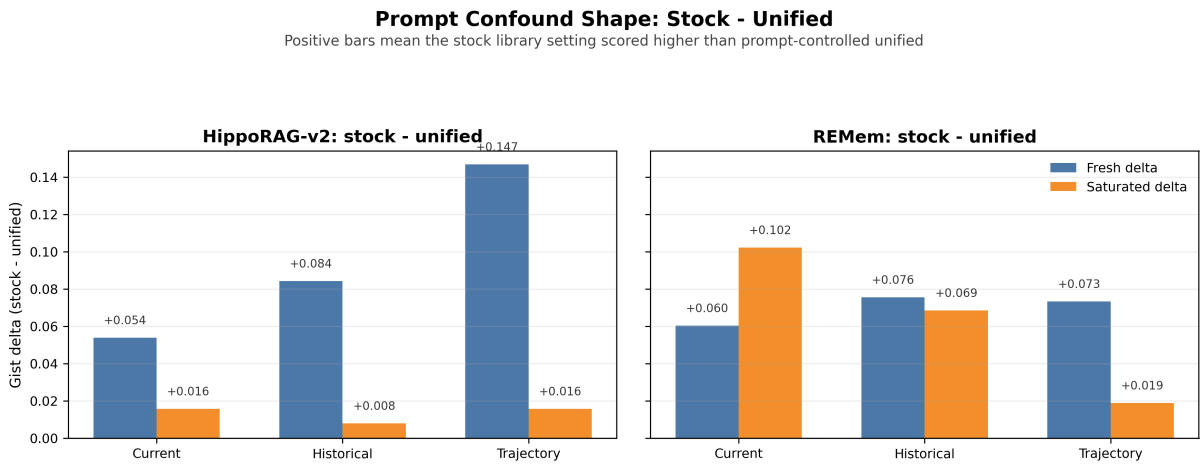


Figure 8: Prompt-control ablation by horizon and question type. This sensitivity check asks whether horizon and question-form changes alone explain the retention patterns; it does not replace the main adapter Fresh-to-Saturated trace.

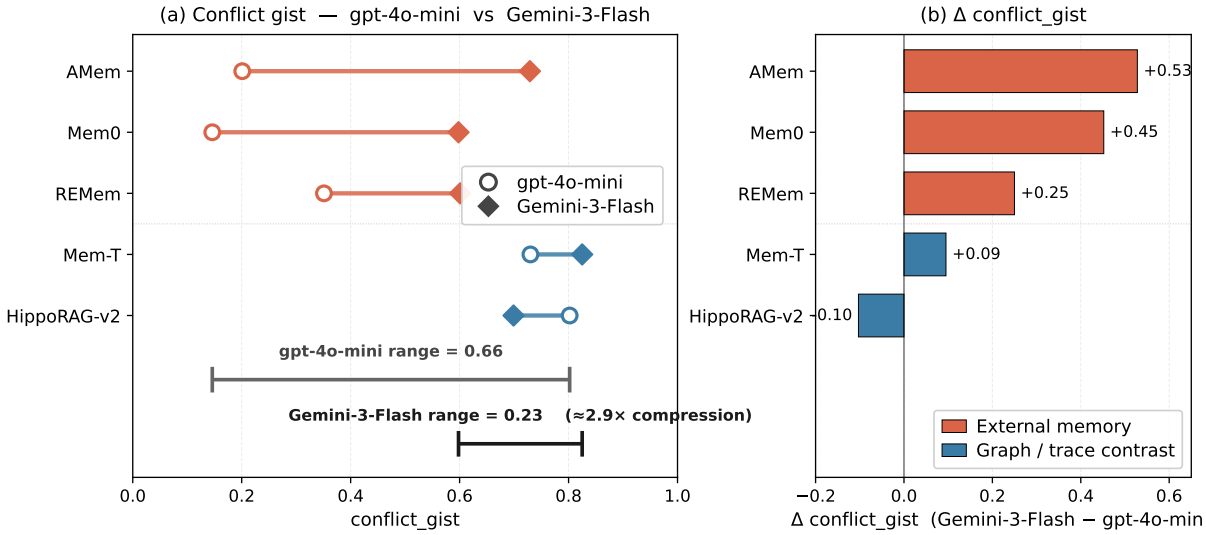


Figure 9: Backbone-sensitivity panel on conflict probes. The main-adapter gpt-4o-mini condition is the audit baseline, while the answer-time generator is swapped to Gemini-3-Flash.

System	Group	4o-mini	Gemini	Δ
HippoRAG-v2	Graph RAG	80.2%	69.9%	-10.3
Mem-T	Trace	73.0%	82.5%	+9.5
REMem	External	35.1%	60.1%	+25.0
AMem	External	20.1%	72.9%	+52.8
Mem0	External	14.6%	59.8%	+45.2

Generator	Subset	Base	Oracle	Lift
4.1-mini	All-fail traj.	0.0%	70.0%	+70.0
4.1-mini	Dynamic traj.	33.8%	72.5%	+38.7
4.1-mini	Hard cur./hist.	8.5%	75.0%	+66.5
4o-mini	All-fail traj.	0.0%	85.0%	+85.0
4o-mini	Dynamic traj.	33.8%	80.0%	+46.2
4o-mini	Hard cur./hist.	8.5%	80.0%	+71.5

Table 18: Backbone-sensitivity numeric checks. Left: Conflict Gist shifts when the answer generator is swapped from gpt-4o-mini to Gemini-3-Flash; AMem and Mem-T are exploratory 3-user subsets, while the other rows are full 20-user runs. Right: oracle-evidence recoverability on the same 120-probe balanced sample, showing large lifts under both oracle backbones. Deltas and lifts are percentage points.

Audit signal	Mem0	HippoRAG-v2
Evidence form	Atomic facts	Raw passages
Returned items	10.0 avg. facts	Variable passages
High-relevance fragment in top 3	32/40	n/a
Answer begins with refutation	n/a	32/40

Table 19: Retrieval-interface audit on 40 user-0 conflict probes under the audited system pipelines. The audit illustrates differences in retrieved evidence form and should be read as an interface check, not as a substitute for the main rows.

Mem0 setting	gpt-4o-mini	Gemini-3-Flash	Gap
Vanilla Mem0	13.3%	80.4%	+67.1 pp
Mem0 + score visibility	26.3%	72.1%	+45.8 pp
Gap reduction	21.3 pp, about 32%		

Table 20: Mem0 score-visibility intervention on paired users {0,1,2}. The vanilla gpt-4o-mini value is a paired-subset baseline, not the full 20-user Mem0 main-row value. Surfacing retrieval similarity scores reduces Mem0's backbone-coupling on conflict Gist, with side costs in the same audit: Gemini-3-Flash conflict hallucination rises from 7.1% to 12.1% and boundary abstention falls from 88.7% to 68.8%.