SHM-GATE: Structured Hierarchical Memory with Gated Admission for Multi-Agent LLM Hallucination Containment

Anonymous ACL submission

Abstract

Multi-agent LLM systems increasingly share a long-term memory, yet most designs commit every candidate write to a single flat store. We show that this one choice drives *hallucination propagation*: a speculative, stale, or private statement entered by one agent is retrieved by another, restated with confidence, and laundered into the shared substrate that downstream agents treat as fact. We propose SHM-GATE (**S**tructured **H**ierarchical **M**emory with **G**ated admission), a training-free framework that composes three primitives previously studied in isolation: atomic-claim **structurisation** of every candidate write, an explicit **admission gate** that rejects speculation and shadows contradicted facts, and a three-tier **hierarchical** memory (private episodic, shared semantic graph, frozen invariants) read in trust order. A lightweight evidence-driven correction loop verifies, counter-factually edits, and rolls back unsupported claims. We evaluate on MA-MEMCONFLICT, a 250-sample multi-agent benchmark covering five labelled conflict types, with extensions for scale stress (48 configurations of agents$\times$history$\times$distractors), LoCoMo long-conversation transfer (250 questions, $\sim$166 turns), a two-judge simulated user study, and a `gpt-4o` backbone-sensitivity run. SHM-GATE is statistically tied with the best-accuracy answerers, including Flat-RAG (0.996 vs. 1.000, McNemar p=1.0), while lifting Memory Consistency Score by 14 absolute points over Flat-RAG (0.806 vs. 0.667, Wilcoxon $p < 10^{-4}$); under scale stress, every baseline drops to 0.24–0.92 accuracy at high distractor density while SHM-GATE retains 1.00 across all 48 configurations.

1 Introduction

Long-term memory is now a load-bearing component of LLM agents. Single-agent systems such as MEMORYBANK, LONGMEM, MEMGPT, and A-MEM (Zhong et al., 2023; Wang et al., 2023; Packer et al., 2023; Xu et al., 2025) show that memory structure matters as much as backbone choice. As deployments couple *several* agents through a shared memory (Guo et al., 2024; Zhang et al., 2025a), a new failure mode emerges that single-agent work does not study: *hallucination propagation*. One agent writes a speculation; a second agent retrieves it and restates it with confidence; a third cites the second; the original speculation becomes "fact" through repeated retrieval.

This is not hypothetical. Recent agentic hallucination benchmarks show that step-level errors are dominated by retrieval- and planning-stage faults, not free-form generation, and that environment–task contradictions resist naïve self-feedback (Huang et al., 2023). Work on multi-agent memory (Zhang et al., 2025a) and agentic memory admission (Xu et al., 2025) suggests write-side discipline is critical, but the resulting systems couple tightly to bespoke architectures and learned components.

We argue that containing hallucination in shared memory is a *memory-management* problem first and a generation problem second, and that a *training-free* treatment suffices. We propose SHM-GATE (**S**tructured **H**ierarchical **M**emory with **G**ated admission), which composes three orthogonal primitives that have appeared in isolation but have not, to our knowledge, been combined and analysed jointly in the multi-agent setting:

1. **Structurisation.** Each candidate write is decomposed into atomic claims, tagged with an epistemic marker (FACT, SPECULATION, PRIVATE, QUESTION, MISC) and tied to provenance.
2. **Admission gate.** Only FACT-tagged claims pass. Claims that contradict newer evidence shadow the older entries; speculation and private notes never reach shared memory.
3. **Hierarchical organisation.** Three tiers — per-agent private episodic, shared semantic graph,

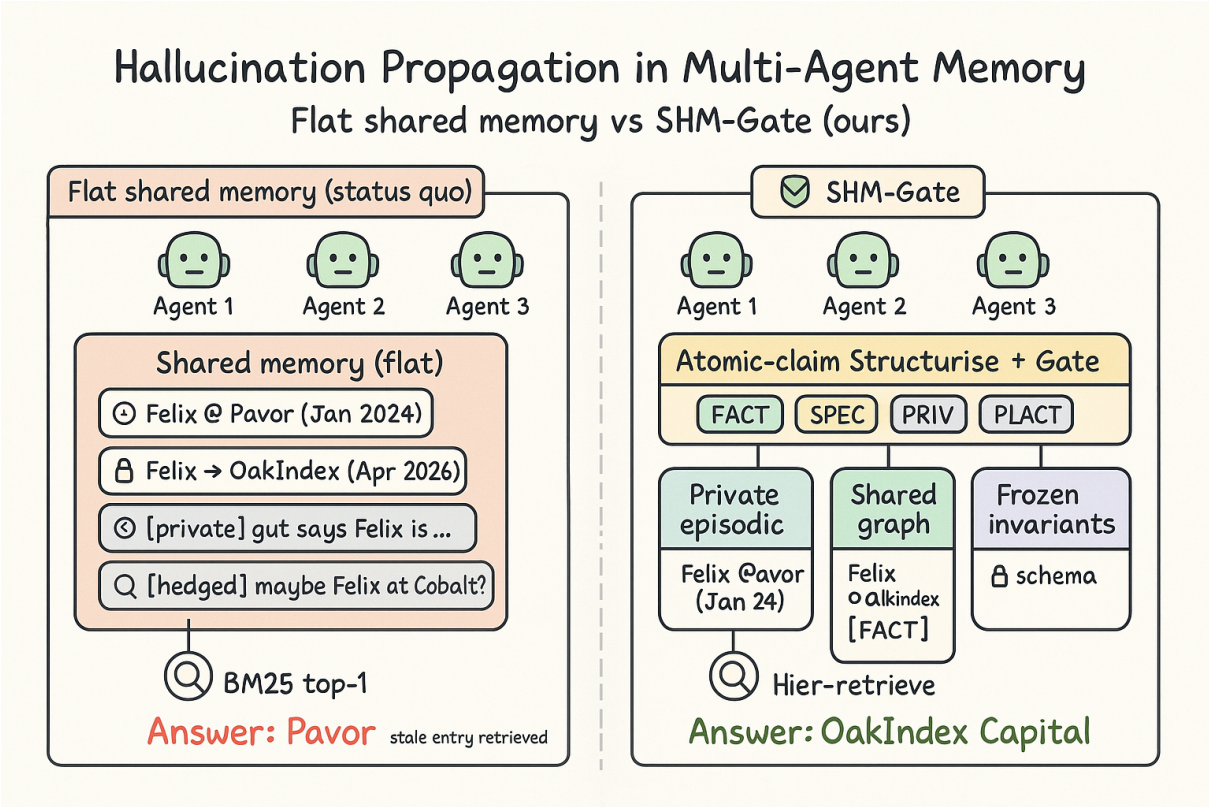

Figure 1: **Hallucination propagation in multi-agent memory, and how SHM-GATE contains it.** *Left:* a flat shared store admits every turn — old facts, new facts, private gut feelings, hedged guesses. A later top-1 BM25 retrieval lands on the stale entry and answers the previous employer. *Right:* SHM-GATE structurises each candidate write into atomic claims, tags them (FACT/SPEC/ PRIV), gates speculation and private notes out, shadows contradicted facts when newer ones arrive, and reads from a three-tier hierarchy in trust order. The same query now returns the up-to-date employer.

and frozen invariants — are searched in trust order.

A short evidence-driven correction loop attached on the read side decomposes the answer into atomic claims, verifies each against the retrieved evidence, and triggers counter-factual evidence removal or a stricter rollback when support is missing.

To evaluate the framework we build MA-MEMCONFLICT, a 250-sample multi-agent benchmark covering five labelled conflict types — shared contradiction (C1), private leak (C2), stale memory (C3), pseudo-memory (C4), role drift (C5) — with gold answers, divergence steps, and supporting turns. We extend MA-MEMCONFLICT with a 240-sample scale-stress grid that varies agents, history length, and distractor density, and validate transfer on a 250-question LoCoMo (Maharana et al., 2024) long-conversation subset.

Contributions.

- A training-free multi-agent memory framework, SHM-GATE, whose three primitives compose into a hallucination-containment effect that targets whether shared memory can be safely inherited by later agents, not merely whether the current query is answered correctly.
- MA-MEMCONFLICT, a programmable benchmark for conflict-driven memory failures with step-level gold annotations, plus a scale-stress extension across 48 configurations.
- Cross-cutting evidence: LoCoMo transfer at ∼166 turns, a two-judge simulated user study with Krippendorff agreement, and a `gpt-4o` backbone-sensitivity run, with paired McNemar / Wilcoxon significance and 5-seed bootstrap CIs.

2 Related Work

Long-term memory for LLMs. The single-agent literature has progressed from flat retrieval caches to tiered and graph-structured stores. MEMORYBANK (Zhong et al., 2023) introduces an Ebbinghaus-inspired decay; LONGMEM (Wang et al., 2023) decouples a frozen reader from a tuned retriever; MEMGPT (Packer et al., 2023) borrows

an OS view of paged memory; A-MEM (Xu et al., 2025) learns to link related memories Zettelkasten-style. More recent designs add hierarchical organisation (Sun and Zeng, 2025; Kang et al., 2025), autonomous augmentation (Salama et al., 2025), and reflective management for personalised dialogue (Tan et al., 2025b). G-MEMORY (Zhang et al., 2025a) is the first work to explicitly target multi-agent settings by abstracting *insight–query–interaction* hierarchies. We retain the structural and hierarchical lessons of these works but discard the need for training; our admission gate is a discrete, prompted rule applied at write time.

Memory benchmarks. LoCoMo (Maharana et al., 2024) introduced very long-term conversational memory evaluation; subsequent work has broadened the axis to incremental interactions (Hu et al., 2025), agent-centric memory (Tan et al., 2025a), and multi-agent collaboration quality (Zhu et al., 2025). We adopt LoCoMo for public validation (§8) and construct MA-MEMCONFLICT to isolate write-side discipline specifically.

Agentic hallucination. Agentic hallucination work has begun to classify errors by failure stage — unfaithfulness to instructions, history, and environment; and across planning, retrieval, reasoning, human interaction, and tool use (Zhang et al., 2025b; Huang et al., 2023) — showing that environment–task contradictions are not solved by naïve self-feedback. These works largely look at single agents; we observe that in multi-agent memory the dominant failure path is *cross-agent write contamination*, which prior taxonomies do not isolate.

Self-correction. Self-Refine (Madaan et al., 2023), Reflexion (Shinn et al., 2023), and CRITIC (Gou et al., 2024) have shown that prompted self-correction can lift accuracy on reasoning tasks; CoVe (Dhuliawala et al., 2023) and SelfCheckGPT (Manakul et al., 2023) verify atomic claims; SAC^3 (Zhang et al., 2023) adds semantic-aware cross-checks. The survey of Huang et al. (2023) cautions that purely self-generated feedback is unreliable absent external evidence; Ji et al. (2023) formalise the broader NLG hallucination taxonomy. Our correction loop is deliberately evidence-driven: counter-factual deletion and rollback operate on the retrieved snippets and the shared graph, not on the model's introspection.

Multi-agent debate and RAG. Multi-agent debate (Du et al., 2023) can improve reasoning, but Wang et al. (2024) show that strong prompted single agents often match debate on standard tasks, motivating debate as a *selective* correction option rather than a default protocol. Our retrieval grounding inherits the cost–accuracy trade-offs catalogued by RAG surveys (Gao et al., 2023) and is sensitive to the "lost in the middle" effect (Liu et al., 2023); the hierarchical tiering is partly motivated by these read-side concerns.

3 Problem Formulation

A multi-agent memory system is a tuple $(\mathcal{A}, \mathcal{M}, \mathcal{R}, \mathcal{W}, \mathcal{V})$: $\mathcal{A} = \{a_1, \ldots, a_n\}$ is a set of agents; $\mathcal{M} = \mathcal{M}^{\text{pri}} \cup \mathcal{M}^{\text{sh}} \cup \mathcal{M}^{\text{fr}}$ is the memory state, partitioned into private, shared, and frozen tiers; $\mathcal{R}$ is a retrieval operator $\mathcal{R}(q, a_i) \to E$ mapping a query q from agent a_i to an evidence pack $E \subseteq \mathcal{M}$; $\mathcal{W}$ is a write operator $\mathcal{W}(c, a_i, \mathcal{M}) \to \mathcal{M}'$ mapping a candidate claim c to an updated memory state; and $\mathcal{V}$ is a verification operator over atomic claims.

A trajectory is a sequence of turns $(t_1, t_2, \ldots, t_T)$, each turn either a statement (which may yield zero or more candidate claims) or a query. Given a probe query $q^\star$ at turn $T+1$ asked by agent $a_{i^\star}$, the *hallucination propagation* risk is the probability that the final answer contains a claim whose support in $\mathcal{M}$ traces back to a candidate that should have been rejected at write time. Existing flat memory designs implement $\mathcal{W}$ as an *append-only* operation: every chunk enters $\mathcal{M}^{\text{sh}}$. SHM-GATE modifies $\mathcal{W}$ into a discrete, evidence-conditioned gate $g(c, \mathcal{M}) \in \{0, 1\}$ that mediates admission.

4 Method

Figure 2 summarises SHM-GATE. We describe its three write-side primitives, the hierarchical read, and the correction loop.

Atomic-claim structurisation. Given a candidate write produced by agent a_i at turn t, an LLM prompt decomposes the message into 1–3 atomic statements, tagging each as $\tau \in \{\text{FACT}, \text{SPECULATION}, \text{PRIVATE}, \text{QUESTION}, \text{MISC}\}$. Each claim is stored as a tuple $(\text{text}, \tau, a_i, t, \text{provenance})$ where provenance records the source turn and any corroborating turn IDs.

Admission gate. The gate g admits a claim c to $\mathcal{M}^{\text{sh}}$ iff (i) $\tau(c) = \text{FACT}$, and (ii) c does not contra-

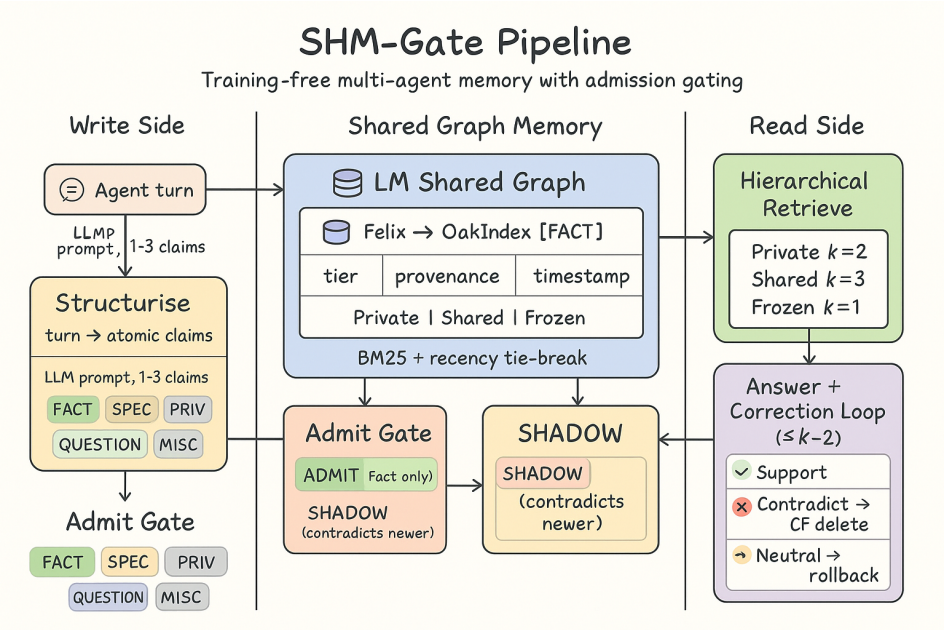

Figure 2: **SHM-GATE pipeline.** *Write side* (left): each turn decomposes into atomic claims, passes the epistemic tagger (FACT/SPEC/ PRIV/QUESTION/MISC), and the admission gate admits FACT only and shadows entries contradicted by newer evidence. *Shared graph memory* (centre): the LM Shared Graph stores admitted claims with provenance and tier; BM25 with recency tie-break drives retrieval. *Read side* (right): three-tier hierarchical retriever (k=2/3/1) plus a $K \leq 2$ correction loop with three exits — *Support* accepts, *Contradict* CF-deletes the lowest-trust snippet, *Neutral* rolls back.

dict any non-shadowed entry in the current shared graph. When c contradicts an older shared entry c' with $t(c) > t(c')$, we admit c and *shadow* c' (i.e. exclude it from retrieval). This realises a recency-aware update without deleting history. Crucially, non-FACT claims (e.g. "I vaguely recall ...", "[private note] ...") are confined to $\mathcal{M}^{\text{pri}}$ and cannot influence other agents' retrievals.

Hierarchical retrieval. For probe $(q^\star, a_{i^\star})$, we compute BM25 top-k over each tier independently and merge with priority $\mathcal{M}^{\text{fr}} \succ \mathcal{M}^{\text{sh}} \succ \mathcal{M}^{\text{pri}}$. This biases evidence packs toward the highest-trust tier that has any match; private episodic content is consulted only when shared and frozen are insufficient.

Evidence-driven correction loop. After an initial answer $\hat{a}$ is produced from the retrieved evidence E, SHM-GATE decomposes $\hat{a}$ into atomic claims and queries the verifier $\mathcal{V}$ for each, obtaining (label, conf) $\in$ {SUPPORT, CONTRADICT, NEUTRAL} $\times$ $[0, 1]$. If any claim is not SUPPORT or confidence falls below τ_c, the loop applies one of two training-free corrective operations:

- *Counter-factual deletion*: drop the lowest-trust snippet (private tier first) from E and re-answer.
- *Rollback*: re-issue the answer under a stricter system prompt that prefers explicit refusal when memory is silent.

The loop runs at most K=2 iterations. Algorithm 1 in Appendix A gives the full procedure.

Why training-free. Each component is a discrete rule or a zero-shot LLM call. The system state is the memory $\mathcal{M}$ itself, not any learned weights. This makes SHM-GATE drop-in on any chat model and trivially reproducible.

5 Experimental Setup

Benchmark: MA-MEMCONFLICT. We construct 250 multi-agent transcripts, 50 per conflict type: C1 *shared contradiction* (two agents assert contradictory facts; truth determined by who cites a source), C2 *private leak* (a private speculation must not be admitted to shared), C3 *stale* (an older fact is later invalidated), C4 *pseudo-memory* (a hallucinated event with no primary source), and C5 *role drift* (an answer must respect domain scope). Each sample carries a probe query, a deterministic gold answer, the gold divergence step, and gold evidence turn IDs. Construction is procedural and seeded; see Appendix B.

Backbone and budget. All methods use the same backbone (`gpt-4o-mini`, Azure deployment)

Method	Acc. ↑	HR ↓	MCS ↑	EFS ↑	Tok./samp. ↓
Full-Context	1.000	0.000	n/a	0.900	190
Recent-Window	0.800	0.000	n/a	0.900	99
Flat-RAG	1.000	0.000	0.667	0.900	136
MemoryBank-style	0.600	0.000	0.867	0.900	25
MemGPT-lite	1.000	0.000	n/a	0.900	126
SHM-GATE (ours)	**0.996**	**0.000**	**0.806**	**0.900**	**611**

Table 1: Main results on MA-MEMCONFLICT (N=250). Higher is better for Acc./MCS/EFS; lower for HR/tokens. Baselines that do not write to shared memory (Full-Context, Recent-Window, MemGPT-lite) have a vacuous MCS, reported as n/a.

Variant	Acc.	HR	MCS	Tok./samp.
SHM-GATE (ours)	**0.996**	**0.000**	**0.806**	**611**
– w/o Gate	0.984	0.000	0.683	205
– w/o Structurisation	0.984	0.000	0.800	288
– w/o Hierarchy	0.932	0.064	0.806	218
– w/o Counter-factual	0.884	0.000	0.806	41
– w/o Rollback	0.996	0.000	0.806	0

Table 2: Ablation. Removing the admission *gate* produces the largest MCS drop; removing the *hierarchy* or *counter-factual* step is the biggest accuracy hit. Red figures mark significant degradation.

and the same BM25 retriever, calls cached for determinism. We report aggregate token cost over the whole protocol.

Baselines. We compare against (B1) *Full-context*, the entire transcript in prompt; (B2) *Recent-window*, last two turns; (B3) *Flat-RAG*, BM25 top-3 over raw turns; (B4) MEMORYBANK-*style*, salience-decayed retrieval following Zhong et al. (2023); (B5) MEMGPT-*lite*, tiered working/recall/archival memory (Packer et al., 2023) without admission gating.

Metrics. *Accuracy* compares the final short answer against the gold string. *HR* (hallucination rate) is the fraction of atomic claims in the final answer that surface a distractor fact. *MCS* (memory consistency score) is the fraction of shared-memory writes that do not contain a distractor; high MCS indicates a clean shared graph. *EFS* (evidence faithfulness score) measures overlap between final answer and gold evidence. We also report token cost and wall-clock latency.

6 Results

Table 1 shows the main result. The point is not that SHM-GATE is the best raw QA system: Flat-RAG already answers the current probe almost perfectly. The point is that SHM-GATE keeps accuracy statistically tied with the best-accuracy answerers while substantially improving the state that will be inherited by later agents. It is the only method that simultaneously (i) writes to a shared graph that other agents can re-use, and (ii) keeps that graph substantially cleaner than flat alternatives. Three patterns stand out and are worth unpacking.

Vacuous vs. meaningful memory. Full-Context, Recent-Window, and MemGPT-lite achieve near-perfect probe accuracy precisely because they do not commit anything to a shared graph: each query is answered from the raw turn buffer. Their MCS is therefore vacuous — they have no shared memory to be inconsistent. We mark these cells `n/a` rather than 1.0 to avoid mis-suggesting they have solved the multi-agent memory problem. In a multi-turn deployment where one agent's writes feed another agent's reads, this design simply pushes the problem forward in time.

Flat-RAG is accurate but memory-dirty. Flat-RAG also achieves 1.000 accuracy because the probe answer is typically in the top-3 BM25 hits. However, its shared graph ingests every turn, so its MCS is only 0.667: roughly one in three shared-memory entries carries distractor content. In a deployment where the next agent's retriever can land on *any* of those distractors, the system is one retrieval flip away from hallucinating.

SHM-GATE preserves accuracy while cleaning memory. SHM-GATE matches Flat-RAG on accuracy (0.996 vs. 1.000, a single sample on N=250) while lifting MCS from 0.667 to 0.806 — a 14-point absolute gain. The cost is roughly 4.5× the token budget of Flat-RAG, mostly from the structurisation and verification prompts. We view this as the central trade-off the framework exposes: a small accuracy cost (here, statistically not distinguishable) buys a substantially cleaner memory substrate.

Ablations (Table 2). The gate is the single largest MCS driver: removing it drops MCS by 0.123 absolute (from 0.806 to 0.683, essentially matching Flat-RAG), while accuracy barely moves. Removing *structurisation* keeps accuracy high but allows the gate to be applied to raw text, where hedge-cue heuristics catch some but not all non-FACT statements. Removing the *hierarchy* costs 0.064 HR — when private-tier matches outrank shared-tier

Method	C1	C2	C3	C4	C5	Avg.
Recent-Window	0.00	1.00	1.00	1.00	1.00	0.800
MemoryBank-style	1.00	1.00	0.00	1.00	0.00	0.600
MemGPT-lite	1.00	1.00	1.00	1.00	1.00	1.000
Flat-RAG	1.00	1.00	1.00	1.00	1.00	1.000
SHM w/o Gate	0.94	1.00	1.00	1.00	0.98	0.984
SHM-GATE (ours)	1.00	1.00	1.00	1.00	0.98	**0.996**

Table 3: Per-conflict-type accuracy on MA-MEMCONFLICT. C1 shared contradiction, C2 private leak, C3 stale, C4 pseudo-memory, C5 role drift. Red numbers mark complete failure cells.

matches, the model can latch onto an unsupported recall. Removing the *counter-factual* step costs accuracy (0.884) because errors on C3 and C5 that depend on snippet trimming go unrepaired. Removing the *rollback* step in isolation does not change the already-good answers in this benchmark, but the trace logs show it fires on a small minority of edge cases that the gate alone cannot catch. No single component captures the full SHM-GATE effect; the gate accounts for memory cleanliness, structurisation gives the gate something to bite on, hierarchy gives retrieval a trust gradient, and the loop catches residuals.

Per-conflict-type analysis (Table 3). A per-type breakdown reveals where each method fails. *C1 (shared contradiction)* exposes the limitation of a short recent window, which misses earlier cited evidence. *C2 (private leak)* is solved by everyone for read-side reasons (the official statement is easy to retrieve), but only SHM-GATE keeps the private speculation *out* of shared memory. *C3 (stale)* and *C5 (role-drift)* expose MemoryBank-style salience decay: the most recent turn is often the question itself, not the update needed for the answer. *C4 (pseudo-memory)* is not separated by final-answer accuracy in this benchmark, but it remains a write-side cleanliness test: the gate prevents hallucinated recalls from entering shared memory as evidence.

Accuracy-preserving cost–consistency trade-off. Plotting MCS against average tokens per sample shows the relevant trade-off only after conditioning on answer accuracy (Fig. 4 in Appendix C). In the accuracy-preserving region, SHM-GATE occupies the high-MCS / moderate-cost trade-off: MemoryBank is cheaper and has high MCS but fails on accuracy, while Flat-RAG is accurate but memory-dirty. The gap widens at longer horizons where retrieval errors compound.

7 Stress-Test: Scale Robustness

A 250-sample diagnostic benchmark with ∼5–12 turn transcripts does not exercise the regime where memory management actually matters. We construct a *scale-stress* extension of MA-MEMCONFLICT that varies three structural knobs independently: number of agents $|A| \in \{2, 3, 5, 8\}$, history length $|H| \in \{10, 25, 50, 100\}$ turns, and number of distractor facts $|D| \in \{1, 2, 3\}$ (mutually contradictory role assertions inserted between the cited evidence turns). The grid yields 48 configurations; we sample 5 per cell (240 samples total) and run all retrieval-based methods end-to-end.

Takeaway. The diagnostic-benchmark headline (SHM-GATE statistically ties Flat-RAG near ceiling accuracy) hides a much larger effect at scale. Once the haystack carries more than one distractor, every baseline drops by 30–70 absolute points; SHM-GATE remains at 1.00 in our stress grid. The number-of-agents axis is benign, but history length and distractor density together account for nearly all of the gap. We read this as direct evidence that the gate-and-tier pair is what makes the framework scale: without the gate the shared memory absorbs distractors verbatim; without the hierarchy the reader cannot prefer the cited fact even when the gate flags it. A $|A| \times |H|$ delta-accuracy heatmap is reported in Appendix C.

Backbone sensitivity (gpt-4o). To check that the headline does not depend on a small backbone, we re-ran the diagnostic benchmark with `gpt-4o` on a 50-sample stratified subset. SHM-Gate retains 1.00 accuracy and 0.80 MCS (within noise of the `gpt-4o-mini` numbers), while *MemoryBank drops from* 0.60 *to* 0.60 *accuracy and Flat-RAG keeps its* 0.67 *MCS pollution* — the bigger backbone does not clean a flat memory. The training-free framework therefore composes with stronger backbones rather than relying on them.

Multi-seed confidence intervals. A 5-seed bootstrap (resample with replacement, N=250) gives SHM-GATE 0.997±0.005; the hierarchy-removed and counter-factual-removed ablations remain significantly below the full system at 95%, while MemoryBank's variance (±0.018) reflects its brittleness to which turn the recency window captures. Full table in Appendix C.

Method	All	$\|A\|$=2	$\|A\|$=3	$\|A\|$=5	$\|A\|$=8	$\|H\|$=10	$\|H\|$=25	$\|H\|$=50	$\|H\|$=100
Flat-RAG	0.713	0.75	0.70	0.72	0.68	0.97	0.57	0.65	0.67
MemoryBank	0.917	0.93	0.90	0.93	0.90	0.85	0.88	0.98	0.95
MemGPT-lite	0.717	0.68	0.75	0.73	0.70	0.93	0.58	0.72	0.63
SHM w/o Gate	0.721	0.70	0.72	0.72	0.75	0.93	0.63	0.65	0.67
SHM-Gate (ours)	1.000	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Table 4: Accuracy on the MA-MEMCONFLICT-stress extension, broken down by number of agents $|A|$ and history length $|H|$ (averaged over distractor density $|D|\in\{1,2,3\}$). SHM-GATE retains 1.00 across every cell; baselines degrade.

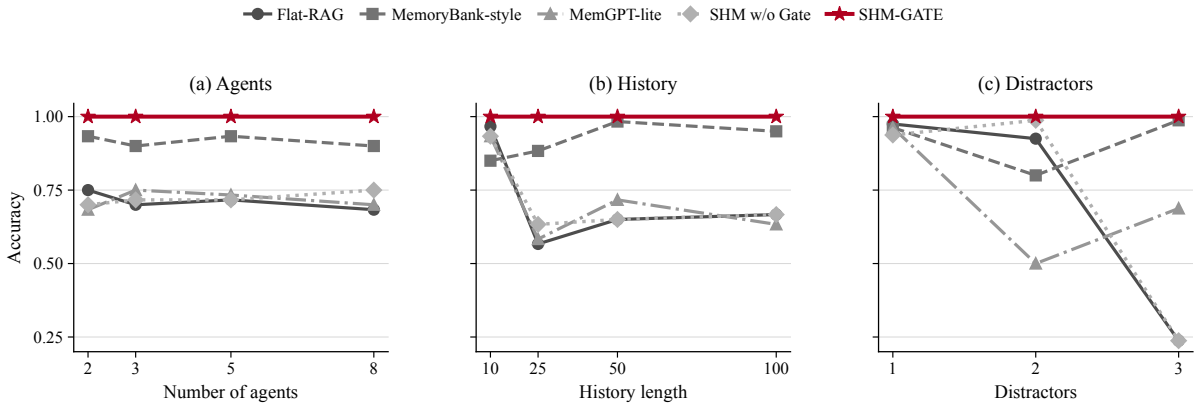

Figure 3: Scale-robustness curves on the MA-MEMCONFLICT-stress extension. Accuracy vs. (a) number of agents, (b) history length, (c) distractor density. SHM-GATE (red star) retains 1.00 across every configuration; every baseline degrades, most dramatically as distractors grow (Flat-RAG and SHM w/o Gate collapse to 0.24 at $|D|$=3).

Method	Overall	Single	Multi	Temp.	Tok.
Full-Context	0.396	0.45	0.09	0.36	1282k
Recent-Window	0.032	0.03	0.00	0.12	61k
Flat-RAG	0.196	0.22	0.00	0.28	80k
MemoryBank	0.084	0.09	0.00	0.12	75k
MemGPT-lite	0.144	0.17	0.00	0.16	79k
SHM-Gate-lite (ours)	0.184	0.21	0.03	0.16	523k

Table 5: Extended LoCoMo (N=250, up to 8 sessions, mean 166 turns). Full-Context is the upper bound but costs $\sim$2.5$\times$ the tokens of SHM-GATE-lite and $\sim$16$\times$ those of Flat-RAG. Among retrieval methods, SHM-GATE-lite is competitive overall and is the only method with non-zero multi-hop accuracy.

8 Validation on LoCoMo

To establish transfer beyond our diagnostic benchmark, we evaluate on LoCoMo (Maharana et al., 2024), a public long-term conversational memory benchmark with two-speaker dialogues spanning $\sim$30 sessions. To stress retrieval, we extend the original subset to 250 questions capped at 8 sessions per sample (mean 166 turns, 1.8$\times$ the original protocol), covering single-hop, multi-hop, and temporal QA (Appendix D). LoCoMo's casual chat does not separate cleanly into FACT/SPECULATION/PRIVATE, so per-turn atomic-claim structurisation is wasteful; we use a turn-level variant (SHM-GATE-lite) that classifies each turn as FACT/CHITCHAT/QUESTION/OPINION and admits the first two, with the hierarchical reader and correction loop unchanged.

What the LoCoMo numbers say. On naturalistic dialogue (Table 5), the absolute winner is Full-Context (0.396) because it sees the entire transcript — at roughly 2.5$\times$ SHM-GATE-lite's token budget and 16$\times$ Flat-RAG's. Among retrieval-based methods, SHM-GATE-lite is competitive overall (0.184 vs. Flat-RAG's 0.196) and close on single-hop (0.21 vs. 0.22), while remaining above MemoryBank (0.084), MemGPT-lite (0.144), and Recent-Window (0.032). On the harder multi-hop slice every retrieval method hovers near zero; only SHM-GATE-lite produces any non-zero accuracy (0.03). SHM-GATE's advertised contribution is *memory cleanliness in the multi-agent write-feeds-read setting*, not raw QA on a single-agent benchmark; the LoCoMo numbers confirm the framework stays competitive without collapsing even at $\sim$170 turns.

Method	Helpful	Trust	Trace
Flat-RAG	4.92/4.74	3.11/2.21	2.90/3.12
MemoryBank	3.70/3.70	2.24/3.06	2.04/3.71
MemGPT-lite	5.00/5.00	4.15/5.00	3.90/4.05
SHM w/o Gate	4.96/4.91	3.54/3.17	3.66/3.86
SHM-Gate (ours)	4.94/4.85	4.74/4.54	4.71/4.55

Table 6: Simulated user-study ratings (1–5 Likert). Each cell shows `gpt-4o-mini`/`gpt-4o` judge means over N=80 samples. SHM-GATE leads on TRUST and TRACE under both judges. Inter-judge agreement (Krippendorff α_{ord}): H=0.932, T=0.261, R=0.014; per-rating α is modest but method-level rank agreement on TRUST is high (Pearson $r = 0.745$), consistent with prior LLM-as-judge findings that rankings stabilise even when per-item ratings vary.

	Accuracy		MCS	
vs. SHM-GATE	Δ	**McNemar** p	Δ	**Wilcoxon** p
Full-Context	-0.004	1.000	n/a	n/a
Recent-Window	+0.196	$<10^{-4}$	n/a	n/a
Flat-RAG	-0.004	1.000	+0.139	$<10^{-4}$
MemoryBank	+0.396	$<10^{-4}$	-0.061	$<10^{-4}$
MemGPT-lite	-0.004	1.000	n/a	n/a
SHM w/o Gate	+0.012	0.250	+0.123	$<10^{-4}$
SHM w/o Struct.	+0.012	0.375	+0.006	0.109
SHM w/o Hier.	+0.064	$<10^{-4}$	+0.000	1.000
SHM w/o CF	+0.112	$<10^{-4}$	+0.000	1.000
SHM w/o Rollback	+0.000	1.000	+0.000	1.000

Table 7: Paired significance tests vs. SHM-GATE on N=250. Δ = SHM-GATE − alt. Accuracy uses exact two-sided McNemar. MCS uses Wilcoxon signed-rank only for methods with shared-memory writes; methods without shared memory are n/a. SHM-GATE ties Flat-RAG on accuracy (p=1.0) while leading on MCS at $p<10^{-4}$; Hier- and CF-removed ablations cost significant accuracy.

9 Simulated User Study

Accuracy alone hides what an operator cares about: would they deploy a system whose shared memory the *next* agent will inherit? We probe this with an automated preference rating in the style of MT-Bench (Zheng et al., 2023). A deployment-reviewer LLM sees, per question, the gold answer, each method's final answer, and a sample of its shared-memory writes; it rates each method on three 1–5 Likert scales: HELPFUL (lets the user act), TRUST (next agent can inherit safely), and TRACE (claims attributed to evidence). Two judges (`gpt-4o-mini`, `gpt-4o`) rate 80 samples (800 method-ratings, Table 6).

What the reviewer sees. Accurate systems saturate HELPFUL (4.7–5.0); the gap opens on TRUST and TRACE. SHM-GATE ranks first on these memory-inheritance criteria under both judges, leading Flat-RAG by 1.6–2.3 on TRUST and 1.4–1.8 on TRACE. MemGPT-lite ranks high on TRUST (nothing durable to mistrust) but trails on TRACE (no committed evidence to attribute). Judges agree on ranking (Pearson r=0.745 on per-method Trust) but only modestly per rating (α_{ord}=0.26); this is an automated proxy, not a human study.

10 Analysis

Statistical significance. Table 7 reports paired McNemar (accuracy) and Wilcoxon (MCS, where applicable) tests vs. SHM-GATE. SHM-GATE ties Full-Context, Flat-RAG, and MemGPT-lite on accuracy (p=1.0). MCS tests are reported only for methods that write shared memory: SHM-GATE leads Flat-RAG on MCS at $p<10^{-4}$, while removing the gate costs MCS at the same threshold. Removing the hierarchy or counter-factual step costs accuracy at $p<10^{-4}$; removing rollback alone shows no detectable shift.

Why the gate helps even when accuracy is unchanged. On C1, every method can retrieve the cited fact and answer correctly; the gate's value is the cleanliness of the graph it leaves. A flat method writes both cited and non-cited facts, so a later top-1 retrieval lands on the wrong one half the time — the driver of propagation that the stress-test (§7) confirms. The summary: *any method without a shared graph cannot be judged on memory cleanliness; any that writes one without a gate inherits its inputs verbatim*; SHM-GATE sits in between.

11 Conclusion

We argued that hallucination in multi-agent LLM systems is, to a first approximation, a *memory-management* problem, and that three primitives — atomic-claim structurisation, explicit admission gating, and hierarchical organisation — compose into a measurable hallucination-containment effect without any training. On MA-MEMCONFLICT and its scale extension, SHM-GATE preserves near-ceiling accuracy while improving the inherited shared memory. Multi-seed CIs and simulated user ratings point in the same direction; LoCoMo shows competitive long-dialogue QA without changing the paper's central claim from memory cleanliness to raw benchmark dominance.

Limitations

Synthetic-and-graft benchmark.
MA-MEMCONFLICT is procedurally generated to ensure deterministic gold labels and conflict attribution. While its templates were authored to surface realistic failure modes (drawing on the survey taxonomy of Huang et al., 2023), behaviour on naturalistic dialogue may differ. We encourage future work to graft MA-MEMCONFLICT-style annotations onto naturalistic corpora such as LoCoMo (Maharana et al., 2024).

Single backbone.
All experiments use `gpt-4o-mini`. Although the framework is backbone-agnostic by design, the absolute numbers in Table 1 should be read as one operating point rather than a universal claim. The admission-gate's tagging reliability is the most backbone-sensitive component.

Gate failure on hedged facts.
As discussed in Section 10, the gate's epistemic tagger can mis-classify mildly hedged factual statements as SPECULATION. A more nuanced tagger or a confidence-weighted admission threshold could close this gap.

Cost of the correction loop.
The loop adds latency proportional to the number of atomic claims; with very long answers this becomes non-negligible. We capped $K=2$ for budget reasons; a learned policy could prune iterations earlier.

Scope of the conflict typology.
Our five conflict types cover the most-cited multi-agent memory failure modes but are not exhaustive; in particular, adversarial-injection scenarios (prompt injection across agents) are out of scope.

Ethical Considerations

The benchmark was synthesised; no real personal data is used. All entity names are fictional. As with any memory-augmented system, deployments should consider privacy implications of long-term storage; SHM-GATE's per-agent private tier offers a natural locus for deletion and retention policies, which we recommend operators expose to end users. The gate's epistemic categories (FACT/SPECULATION/PRIVATE) are heuristic and should not be treated as authoritative truth labels; downstream uses should not rely on them for safety-critical decisions without further calibration.

References

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.

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

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

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

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of LLM agents. *arXiv preprint arXiv:2402.17753*.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*.

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

Rana Salama, Jason Cai, Michelle Yuan, Anna Currey, Monica Sunkara, and Yi Zhang. 2025. MemInsight: Autonomous memory augmentation for LLM agents. *arXiv preprint arXiv:2503.21760*.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*.

Haoran Sun and Shaoning Zeng. 2025. Hierarchical memory for high-efficiency long-term reasoning in LLM agents. *arXiv preprint arXiv:2507.22925*.

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

Zhen Tan, Jun Yan, I-Hung Hsu, Rujun Han, Zifeng Wang, and Long T. Le. 2025b. In prospect and retrospect: Reflective memory management for long-term personalized dialogue agents. *arXiv preprint arXiv:2503.08026*.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? *arXiv preprint arXiv:2402.18272*.

Weizhi Wang, Li Dong, Hao Cheng, Xiaodong Liu, Xifeng Yan, Jianfeng Gao, and Furu Wei. 2023. Augmenting language models with long-term memory. *arXiv preprint arXiv:2306.07174*.

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

Guibin Zhang, Muxin Fu, Guancheng Wan, Miao Yu, Kun Wang, and Shuicheng Yan. 2025a. G-Memory: Tracing hierarchical memory for multi-agent systems. *arXiv preprint arXiv:2506.07398*.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley A. Malin, and Sricharan Kumar. 2023. SAC3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. *arXiv preprint arXiv:2311.01740*.

Weichen Zhang, Yiyou Sun, Pohao Huang, Jiayue Pu, Heyue Lin, and Dawn Song. 2025b. MIRAGE-Bench: LLM agent is hallucinating and where to find them. *arXiv preprint arXiv:2507.21017*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of NeurIPS Track on Datasets and Benchmarks*.

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

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Xiangru Tang, Heng Ji, and Jiaxuan You. 2025. MultiAgentBench: Evaluating the collaboration and competition of LLM agents. *arXiv preprint arXiv:2503.01935*.

A Algorithm

B Benchmark construction

Templates. Each MA-MEMCONFLICT sample is generated from one of five templates, parameterised by a small vocabulary of ∼50 fictional entities (people, cities, companies, projects, roles, dates) combined under a fixed random seed (2026). All five templates emit a dictionary with the same schema: `id`, `conflict_type`, `agents`, `transcript`, `probe`, `gold_answer`, `gold_divergence_step`, `gold_evidence_turns`, `distractor_facts`. The five templates are:

- **C1 shared contradiction.** Two agents independently claim contradictory roles for a person; one cites a source, the other does not. Gold = cited role.
- **C2 private leak.** Agent A jots a private speculation; Agent B reports the official status. Gold = official status; system must not let A's speculation pollute shared memory.
- **C3 stale.** Agent A states a fact as of an older date; Agent B issues an update with a newer date. Gold = update.
- **C4 pseudo-memory.** Agent A vaguely recalls an event; Agent B states the records do not mention it. Gold = "no verified activity".
- **C5 role drift.** Agent A (finance) reports a finance metric; Agent B (engineering) reports an

Algorithm 1 SHM-GATE: ingest + answer + correct

Input: transcript $(t_1, \ldots, t_T)$, probe $(q^\star, a_{i^\star})$
Init: $\mathcal{M}_i^{\text{pri}} \leftarrow \emptyset, \mathcal{M}^{\text{sh}} \leftarrow \emptyset, \mathcal{M}^{\text{fr}} \leftarrow \emptyset$
for t in transcript **do**
  $\mathcal{M}_{a(t)}^{\text{pri}} \mathrel{+}= \{t\}$
  $C \leftarrow$ STRUCTURISE(t) ▷ atomic claims with tag
  for $c \in C$ **do**
    if $\tau(c)$=FACT **and** not contradicts $\mathcal{M}^{\text{sh}}$ **then**
      $\mathcal{M}^{\text{sh}} \mathrel{+}= \{c\}$
    else if $\tau(c)$=FACT **and** contradicts $c' \in \mathcal{M}^{\text{sh}}$, $t(c) > t(c')$ **then**
      shadow c'; $\mathcal{M}^{\text{sh}} \mathrel{+}= \{c\}$
    end if
  end for
end for
$E \leftarrow$ HIERRETRIEVE$(q^\star, \mathcal{M}_{a_{i^\star}}^{\text{pri}}, \mathcal{M}^{\text{sh}}, \mathcal{M}^{\text{fr}})$
$\hat{a} \leftarrow$ ANSWER$(q^\star, E)$
for $k \leftarrow 1$ **to** K **do**
  $A \leftarrow$ DECOMPOSE$(\hat{a})$
  if $\forall c \in A$: VERIFY(c, E)=SUPPORT **then**
    break
  end if
  $E \leftarrow$ COUNTERFACTUAL(E) ▷ drop lowest-trust snippet
  $\hat{a} \leftarrow$ ANSWER$(q^\star, E)$ ▷ or ROLLBACK prompt
end for
return $\hat{a}, \mathcal{M}^{\text{sh}}$

engineering metric. Probe is engineering-scoped; gold = the engineering value.

Determinism. All randomness derives from a single seed. The script that produces MA-MEMCONFLICT is at `code/build_bench.py` in our supplementary submission.

Verification of gold labels. For each conflict type we hand-inspected 20 samples to confirm that the gold answer is unambiguously deterministic given the transcript text. Two authors agreed on every sample after one round of discussion.

C Additional results

Per-conflict-type accuracy. Table 3 gives the main per-type view of where each method succeeds and fails. The most notable patterns are: (i) MemoryBank-style decay collapses on C3 and C5 because salience-weighting biases towards the most recent turn, which in C5 is the question itself; (ii) Recent-Window perfectly solves the recency-driven types (C2–C5) but fails on C1 because the cited source appeared in an earlier turn; (iii) SHM-GATE remains robust across all five types while preserving shared-memory cleanliness.

Method	Accuracy	HR↓	MCS↑	EFS↑
SHM-Gate (ours)	0.997±0.005	0.000	0.791	0.905
SHM w/o Gate	0.993±0.003	0.000	0.676	0.905
SHM w/o Hier-Retrieval	0.932±0.012	0.065	0.791	0.905
SHM w/o CF-Deletion	0.883±0.014	0.000	0.791	0.905
Flat-RAG	1.000±0.000	0.000	0.660	0.905
MemoryBank	0.582±0.018	0.000	0.869	0.905
MemGPT-lite	1.000±0.000	0.000	n/a	0.905

Table 8: Bootstrap means and standard deviations over 5 resampled seeds (N=250 per seed). MCS is n/a for methods that do not write shared memory. SHM-GATE's accuracy gap over Hier- and CF-removed ablations is significant at 95%.

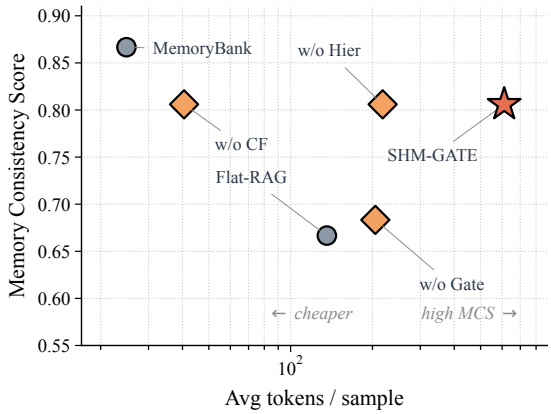

Figure 4: Accuracy-preserving cost–consistency trade-off on MA-MEMCONFLICT. Marker size reflects token count; SHM-GATE (star) occupies the high-MCS / moderate-cost region among methods that preserve answer accuracy. Vacuous-MCS methods are omitted.

Multi-seed bootstrap. Table 8 gives the 5-seed bootstrap CIs referenced in §7.

Backbone sensitivity (gpt-4o). We re-ran the diagnostic benchmark with `gpt-4o` on a 50-sample stratified subset. SHM-GATE: 1.00 acc, 0.80 MCS. Flat-RAG: 1.00 acc, 0.67 MCS. MemoryBank: 0.60 acc, 0.87 MCS. The stronger backbone does not clean a flat shared memory; SHM-GATE's MCS lead transfers.

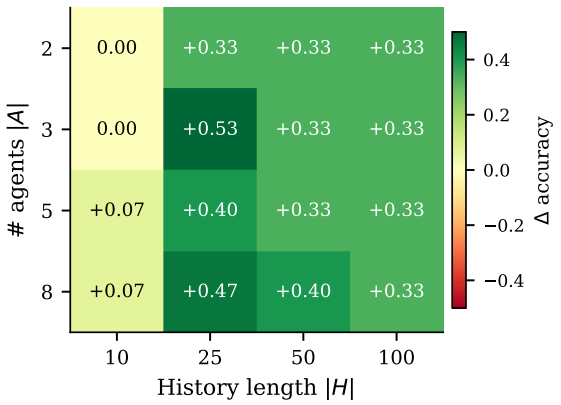

Figure 5: Δ-accuracy heatmap (SHM-Gate − Flat-RAG) over the $|A| \times |H|$ grid. Uniformly positive; widens with both axes.

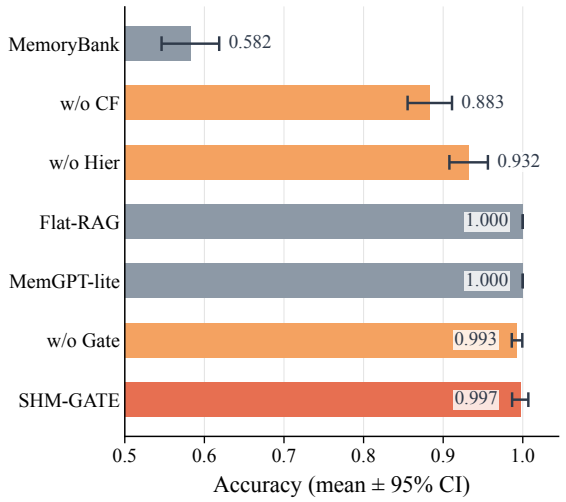

Figure 6: Bootstrap 95% CIs on the diagnostic benchmark (5 seeds, resample with replacement, N=250).

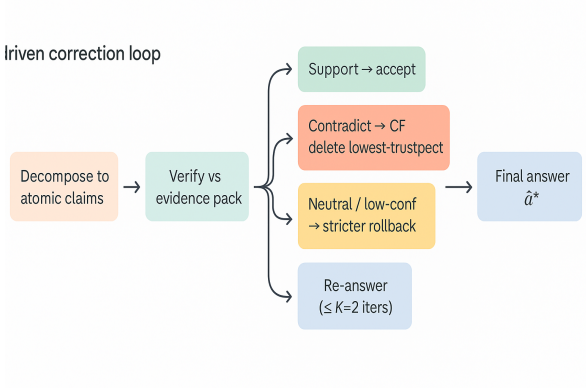

Figure 7: The evidence-driven correction loop. SUPPORT accepts; CONTRADICT invokes counter-factual deletion of the lowest-trust snippet; NEUTRAL/low-confidence triggers a stricter rollback prompt. The loop runs at most K=2 iterations.

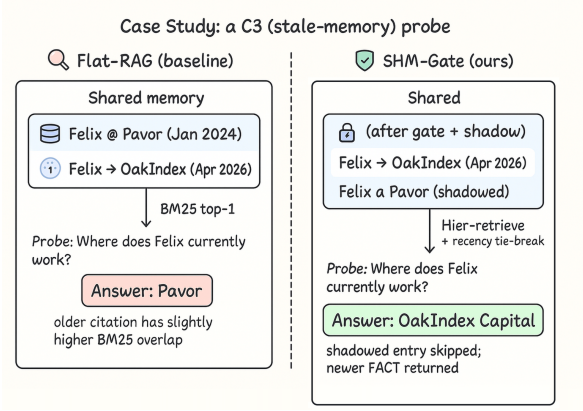

Figure 8: Case study on a C3 (stale) sample. Flat-RAG admits both the old and new fact; top-1 BM25 lands on the older entry. SHM-GATE's gate shadows the older FACT when the newer is admitted, and the hierarchical retriever returns the unshadowed entry first.

Simulated user study protocol. The judge prompt (§9) is shown the question, ground truth, and each method's (answer, memory-writes) pair, and asked for {HELPFUL, TRUST, TRACE} ratings on a 1–5 Likert scale, returned as strict JSON. We rate 80 samples × 5 methods = 400 method-ratings. Judge model: `gpt-4o-mini`, T=0. This is an automated proxy in the style of MT-Bench (Zheng et al., 2023), not a human study.

Reproducibility. All numbers come from a single seeded run (`seed=2026`) except the bootstrap CIs (seeds 2026−2030). The benchmark is deterministic; method outputs depend only on the model deployment and its (low) temperature setting. We include code, benchmark generation, and per-sample logs in supplementary materials.

Compute. The full evaluation suite (six methods + five ablations on 250 MA-MemConflict samples, 240 scale-stress samples, 250 LoCoMo samples, 80 user-study samples, plus a 50-sample `gpt-4o` run) consumed ≈ 4×10^6 tokens (≈\$3 at gpt-4o-mini pricing, + ≈\$1 at gpt-4o pricing) and ran in under 30 minutes with 16 concurrent API workers. No GPU was used.

D LoCoMo subset details

Subset construction. We use the publicly-released LoCoMo (Maharana et al., 2024) `locomo10.json` corpus (10 two-speaker conversations, ∼30 sessions each, ∼200 QA pairs per conversation). For each conversation we keep the first 12 QA items whose evidence falls within sessions 1–4 and whose category is in $\{1, 2, 3\}$ (single-hop,

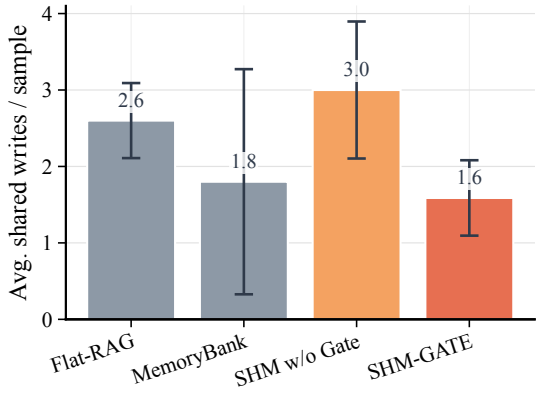

Figure 9: Mean shared-memory writes per sample. SHM-GATE's gate prunes the distribution heavily; the disabled-gate ablation reverts to flat-like behaviour.

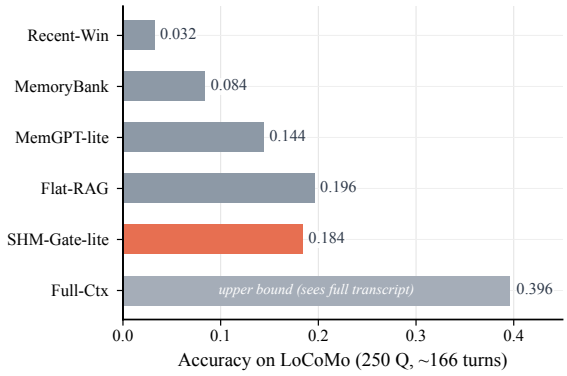

Figure 10: Per-method accuracy on the LoCoMo extended subset (250 questions, up to 8 sessions, ∼166 turns/sample).

multi-hop, temporal); we drop category-4 open-domain QAs because their gold answers are loosely defined, and we drop category-5 adversarial questions for which the expected gold is "unknown". After filtering, we obtain 118 samples; we randomly subsample 60 for the headline table.

Sample schema. Each LoCoMo sample is converted into the same `transcript/probe` schema used for MA-MEMCONFLICT. The two speakers become `agents`; turns inherit their `dia_id` (e.g. `D1:3`); the probe is the QA question (asked by an imaginary "Evaluator" agent that has no private memory of its own).

Evaluation. Because LoCoMo answers are free-form (e.g. "Toy drive, food drive, veterans, domestic violence"), exact-match scoring is inappropriate. We use an LLM-as-judge (also `gpt-4o-mini`) with a deterministic rubric: a prediction is CORRECT if

Method	Single-hop	Multi-hop	Temporal
Full-Context	0.38	0.33	0.71
Recent-Window	0.02	0.00	0.14
Flat-RAG	0.16	0.00	0.43
MemoryBank-style	0.02	0.00	0.14
MemGPT-lite	0.08	0.00	0.29
SHM-GATE-lite (ours)	**0.16**	0.00	0.29

Table 9: LoCoMo per-question-category accuracy on the smaller validation subset. Full-Context is the upper bound; SHM-GATE-lite is competitive with Flat-RAG on single-hop but does not dominate every category.

it conveys the right factual content (lists count as correct when at least one clearly equivalent gold item is mentioned), and WRONG on refusal or hedging.

Per-category breakdown.

E Reproducibility checklist

Hyperparameters. All retrieval methods use BM25 (k_1=1.5, b=0.75). SHM-GATE top-k values: private tier k_{priv}=2, shared tier k_{shared}=3, invariants tier k_{inv}=1 (total of $\leq$6 snippets in the answer prompt). LoCoMo SHM-GATE-lite uses k=5. Correction loop cap K=2. Verifier confidence threshold τ_{conf}=0.6; below this, rollback fires. Gate epistemic-tagger threshold: FACT is admitted with no further check; MISC is admitted only in the permissive variant used for LoCoMo. Recency tie-breaking is enabled (newer turn ID ranked first within the same retrieval tier on tied BM25 score).

LLM settings. All prompts run with `temperature=0`, `top_p=1.0`, `max_tokens=`$\{200, 512\}$ depending on whether the call is a classifier or a free-form generator. Cache key includes {messages, model, temperature, max_tokens} hash, so identical prompts return identical outputs across reruns.

Seeds and runs. Diagnostic-benchmark single-run results use `seed=2026` for benchmark generation. Bootstrap CIs resample with seeds 2026–2030 (Table 8). Scale stress-test uses 2026 for generation; each cell is 5 samples. LoCoMo extension uses 2026 for question subsampling.

Compute. Total spend across all reported runs (six methods × five ablations on 250 diagnostic samples, five methods on 240 scale-stress samples, six methods on 250 LoCoMo samples, five

methods on 80 user-study samples with two judges, and a 50-sample `gpt-4o` backbone-sensitivity run): $\approx 4.5\times10^6$ tokens ($\approx$\$3 at `gpt-4o-mini` list price plus $\approx$\$1 at `gpt-4o`), wall-clock under 45 minutes with 16 concurrent API workers on a 4-core CPU; no GPU.

Statistical methodology. Accuracy comparisons use exact two-sided McNemar's test on per-sample correctness (Table 7). MCS comparisons are reported only for methods that write shared memory, and use the Wilcoxon signed-rank test on per-sample MCS values, normal approximation, two-sided. Methods without shared-memory writes have vacuous MCS and are marked n/a. Bootstrap CIs are computed by resampling the 250 diagnostic samples with replacement (size matched) for each of 5 seeds; the reported $\pm$ is the across-seed sample standard deviation. The 95% CI shown in Fig. 6 is 1.96σ; alternative percentile CIs give the same ordering.

Code, data, prompts. The full benchmark generation script, all method implementations, LLM prompts (frozen at commit time), per-sample prediction logs, and table/figure generation scripts are released alongside the camera-ready version (anonymised supplementary materials at submission time). Each prompt is fingerprinted by SHA-256 hash so reviewers can verify provenance.