Retrieval Is Not State Management, and Reconstruction Is Not Governance

Separating Memory Selection, State Resolution, and Experience Adaptation in LLM Agents

Matt Liotta · August 2026

Abstract

Memory-augmented LLM agents conflate three distinct operations: *retrieval* (which records look relevant), *state resolution* (which propositions are currently valid, authoritative, and visible), and *reconstruction* (how prior experience adapts to the present). Prior work has separated them pairwise but never all three. Liotta (2025) separated retrieval from state resolution; MemHarness (Wu et al., 2026) separated retrieval from reconstruction by having a GRPO-trained policy critique and rewrite retrieved experiences before acting. Neither separated reconstruction from governance.

We give a three-plane decomposition in which reconstruction may adapt guidance but may not establish authoritative state, and we test the separation empirically. Our benchmark extension generates *minimal counterfactual pairs*: two timelines identical in entity, wording, event count, and query, differing in exactly one governance variable. Pairs are admissible only if the governance variable is *not recoverable from task-success reward* — on three of seven axes the discriminating fact is removed from context by state resolution, so no policy over that context could learn the distinction at any amount of training.

On those axes, prompted reconstruction achieves **0% pair accuracy** across 12 pairs: it answers the counterfactual exactly as the positive, using state a correct resolver withheld. The identical pipeline with a deterministic state layer in front achieves **100%**. A judge-free metric scored against the engine's own audit record corroborates (60% vs 20% governance bypass). On an 18-cell governance × applicability factorial, all three reconstruction configurations are behaviorally indistinguishable from "always answer": correct rejection 0%, false refusal 0%. Declining to use an inapplicable memory is a capability the reconstruction stage does not exhibit.

Two predictions failed. We found **no reconstruction-specific state synthesis** — the hypothesized "invents a plausible replacement" failure did not appear. And a post-hoc validator was a **no-op**: it never fired, because reconstruction produced no unbacked bindings. We report both as negative results.

The reconstruction baselines are prompted and untrained, so a trained policy might do better on the axes where the discriminating information is *visible*. That is why the admissibility criterion matters: on the engine-decidable axes the information is not visible, and the negative result there does not depend on training.

All results here use a corrected evaluation instrument. Building it revealed defects serious enough to invalidate figures in our own prior work — chiefly that the standard resurrection metric was substantially measuring response length — which we report in a companion audit (Liotta, 2026b). The constructive consequence for this paper is that its two primary metrics are scored against the state layer's own audit record rather than against author-written phrase lists, and therefore require no judge.

---

1. Introduction

An agent retrieves a past experience. It is semantically relevant. Is it usable?

The question decomposes further than the literature usually admits. The experience might be inapplicable because the *situation* changed — the customer now needs 32 GB of RAM and the previously recommended laptop supports 16. That is the applicability problem, and MemHarness (Wu et al., 2026) addresses it well: replaying a retrieved memory verbatim causes negative transfer, and a policy trained to critique and reconstruct it does better.

But the experience might instead be unusable because the *state underneath it* is no longer permitted: superseded, expired, restricted to a different actor, outranked by policy, or scoped to a draft. Nothing about the experience signals this. It reads exactly as applicable as it did when it was valid. A reconstruction stage examining the experience and the current observation has no basis on which to reject it, because the disqualifying fact is a property of the *record*, not of the situation.

This paper argues that these are different problems requiring different mechanisms, and that conflating them produces a measurable, attributable class of failure.

Contributions.

1. A three-plane decomposition (retrieval / state resolution / reconstruction) with an explicit restriction — reconstruction may adapt guidance but may not establish authoritative state — and a post-condition validator dual to the pre-condition filter.
2. A benchmark extension of paired counterfactuals over seven governance axes, with an *admissibility criterion* that partitions axes by whether the discriminating variable could be learned from reward. This is the paper's defense against "your baseline wasn't trained," enforced in code rather than argued in prose.
3. Metrics grounded in the state layer's own audit record rather than in author-written phrase lists: Governance Bypass Rate and Unsupported Reconstruction Rate are scored against what the engine actually excluded and admitted, requiring no judge.
4. Empirical results: reconstruction does not enforce governance (decisive); reconstruction does not decline (decisive); reconstruction does not synthesize unsupported state (negative result, prediction failed); the validator never fired (negative result).
5. Negative results reported as such, including two of our own failed predictions.

A sixth contribution — the measurement-validity audit that this work required as a prerequisite — is reported separately (Liotta, 2026b), because its scope is the evaluation of agent memory generally rather than the governance question studied here.

---

2. Related work

State-based context architecture. Liotta (2025) reframes agent memory as structured state assembled per turn rather than transcript replay, introducing supersession, scope, authority, temporal validity, dependency repair, and known unknowns. Liotta (2026) implements this as Memgine, a deterministic engine, and identifies an *enforcement–reasoning boundary*: filtering wins where correctness depends on controlling what reaches the model, while context layout helps where it depends on reasoning. This paper adopts the state plane from that line and tests what it does not cover.

Reconstructive memory. MemHarness (Wu et al., 2026) recasts memory-guided decision-making as observe → retrieve → critique → reconstruct → act, with the reconstructive ability emerging end-to-end via GRPO, evaluated on ALFWorld and WebShop. Its diagnosis of negative transfer from verbatim replay is correct and independently converges with the state-based critique of transcript replay. Its scope is procedural applicability; it does not address whether reconstructed guidance is authorized, non-superseded, or properly scoped.

We are not evaluating MemHarness. That system is a trained action-selection policy with an environment and an action space; StateBench is single-turn response generation with neither. Only the *mechanism* transfers, and we implement it as a prompted, untrained stage. Our baselines are named `reconstructive_rag_*` for this reason, and §9.2 states what that limitation does and does not permit us to conclude.

Distinguishing a near-namesake. "Memory is Reconstructed, Not Retrieved: Graph Memory for LLM Agents" (arXiv:2606.06036) shares MemHarness's framing but builds a graph memory substrate. Our concern is orthogonal to both: not how memory is represented or adapted, but what may be permitted to influence an action.

Governance in agent memory. The Always-On Agents survey (arXiv:2606.30306) frames agent state as including permissions, commitments, provenance, and audit, and observes that the literature overweights accumulation and retrieval relative to governance and recovery. We take this as settling the *general* claim: "agents need governance" is not our contribution. Ours is the experimental separation — holding semantic relevance constant, moving one governance variable, and measuring whether a reconstruction stage tracks it.

Virtual context and knowledge graphs. MemGPT (Packer et al., 2023) treats context as a managed resource with the model doing its own paging; we externalize assembly instead. Zep/Graphiti (Rasmussen et al., 2025) contribute bi-temporal episodic/semantic separation. Mem0 contributes extraction-and-consolidation. None separate authorization from applicability.

3. The three-plane model

Let `H` be event history, `q` a query, `A` an actor, `t` a time.

State plane. A deterministic operator resolves authorized state:

```
S_t^auth = Γ(H, A, t)
```

`Γ` enforces validity via supersession chains, scope containment, authority precedence, temporal expiry, access control, and dependency invalidation. Critically, `Γ` is *auditable*: every excluded proposition carries a machine-readable reason. This audit record is what makes the metrics of §5 judge-free.

Experience plane. `E_t = R(q_t, B)` over an experience bank `B`, where each experience carries its source state `σ(e)` — the conditions under which it held. This generalizes MemHarness's source observation.

Reconstruction plane. `G_t = C(E_t^auth, S_t^auth, O_t)`, where `E_t^auth = E_t ∩ Γ`-permitted.

The restriction that defines the paper:

> *`C` may adapt guidance. `C` may not establish authoritative state.*

Formally: for any proposition `p` asserted in `G_t`, either `p` is entailed by `S_t^auth`, or `p` is an adaptation whose warrant is a principle in `E_t^auth` and whose *bindings* come from `S_t^auth`. A `p` that is neither is **unsupported reconstruction**.

Validation. `V(G_t, S_t^auth)` runs after reconstruction, because reconstruction can introduce violations from clean inputs. `V` is the post-condition dual of `Γ`'s pre-condition.

```
events → Γ (resolve + authorize) → R (retrieve) → C (critique + reconstruct) →
V (validate) → act
```

Γ applies to both planes. This is worth stating because we got it wrong in implementation before we got it right. An experience records a past action *including the value it used*, so an experience derived from a fact `Γ` later excludes carries that value verbatim — and retrieval reinjects it, bypassing `Γ` entirely. Filtering facts is not sufficient. `E_t^auth` is not decoration in the formalism; it is load-bearing, and our own first implementation violated it (§8.3).

4. Benchmark design

4.1 Paired counterfactuals

Each axis generates pairs `(x, x')` from one template, identical in entity, surface wording, event count, event type sequence, actor, domain, and query, differing in exactly one governance variable. The measured quantity is the *behavioral split* between them.

Seven axes: access control, scope binding, actor isolation, temporal validity, supersession, commitment status, dependency validity.

Every pair is audited before use. The audit rejects pairs that differ in event count or type sequence (a system could infer the condition from timeline shape), that differ in query text (relevance not held constant), or whose ground truth is identical on both sides (nothing to measure). All 28 pairs pass, at mean lexical similarity 0.841; the `scope_binding` pairs are byte-identical in text, differing only in scope metadata.

4.2 Admissibility: the criterion that answers "you didn't train it"

The obvious objection to a negative result from an untrained reconstruction stage is that training would fix it. We answer this by construction, not argument.

An axis is **engine-decidable** when the discriminating information is *removed from the model's context* by state resolution — a restricted fact, a draft-scoped fact, another tenant's fact. No policy over that context can condition on what it never sees, so no amount of outcome-reward training can supply the distinction. An axis is **model-decidable** when the discriminating event is visible and must be interpreted; a trained reconstructor could in principle learn these.

Axis	Decidable by	Discriminating information
access control	**engine**	restricted fact filtered pre-inference
scope binding	**engine**	draft-scoped fact filtered pre-inference
actor isolation	**engine**	another tenant's fact never enters context
temporal validity	model	expiry and current time both visible
supersession	model	superseding event visible in timeline
commitment status	model	cancellation stated in conversation
dependency validity	model	corrected premise visible

The generator enforces this: a pair on an engine-decidable axis whose content carries no marker the state layer can filter on is rejected as inadmissible. Only the engine-decidable axes support the paper's strong claim. Results on model-decidable axes are reported separately and weighted as weaker evidence.

This partition is the enforcement–reasoning boundary of Liotta (2026) §7.2 expressed as benchmark design.

4.3 The applicability factorial

Governance status (6) × experience applicability (3) = 18 cells.

Governance divides by what it *leaves behind*. **Withdrawing** statuses (unauthorized, draft, expired) remove the value, so the correct response is a refusal. **Substituting** statuses (superseded, outranked) replace it, so a usable value remains and the correct response is the *new* one — refusing there is itself a failure. Collapsing these two, as our first version did, marks a correct supersession answer wrong.

Governance	applicable	adaptable	unusable
valid	KEEP	ADAPT	REJECT
superseded	ADAPT	ADAPT	ADAPT
outranked	ADAPT	ADAPT	ADAPT
expired	REJECT	REJECT	REJECT
unauthorized	REJECT	REJECT	REJECT
draft	REJECT	REJECT	REJECT

Governance dominates: once state is governed, the required behavior no longer varies with how well the experience transfers. That is the thesis as ground truth.

Degenerate strategies. The suite is REJECT-leaning, so a system that declines everything scores well on classification alone. Two guards are mandatory in reporting: an `always_refuse_baseline` computed on the same sample, and correct-rejection and false-refusal rates over *disjoint* cell sets, never blended. This follows the FSR/FAOR discipline already used for the supersession-maintain guardrail tracks.

---

5. Metrics

5.1 Engine-grounded (judge-free)

Governance Bypass Rate (GBR). The fraction of responses asserting content the state layer *excluded*. Because `Γ` records every removal with a reason, the ground truth is the engine's own decision: no judge, no phrase list, no author foresight. A leak here is information the model was never shown, so it cannot be dismissed as a scoring artifact. This is the paper's headline metric.

GBR is scored against a **reference resolver** run alongside every arm, not against each arm's own exclusions. Scoring against a system's own exclusions makes the metric vacuous for a system that

excludes nothing — precisely the arm expected to leak most. Three exemptions prevent false positives: content shared with an admitted fact, content present in the user's own question, and content appearing only under negation ("I can't share the severance figure" is compliance, not leakage).

Unsupported Reconstruction Rate (URR). The fraction of responses asserting a concrete value traceable to neither the assembled context nor a declared binding. Reconstruction emits typed verdicts — `KEEP | ADAPT | REJECT` plus, for ADAPT, an explicit bindings map naming which current facts supply replacement values — so support is checkable *structurally* rather than by asking a second language model whether the first hallucinated.

5.2 Behavioral

Pair accuracy. The fraction of minimal pairs with *both* sides correct. Always-answer fails every counterfactual; always-refuse fails every positive; only genuine state resolution scores above zero.

We report Δ (the accuracy gap between sides) alongside but stress that it is **not interpretable alone**, and we had it backwards in our own design. The two sides require *different* answers, so a system answering both correctly scores 100%/100% and Δ = 0. Zero Δ is the target, not the warning sign. A large |Δ| marks a system treating both sides alike. Pair accuracy has no such ambiguity.

Applicability classification, correct rejection, false refusal, historical detail carryover. Defined in §4.3, reported with the blanket-refusal baseline.

---

6. Experimental setup

Subject and judge models are stated per experiment. The judge is pinned **globally** and never inherited from the system under test; §8.1 explains why this matters. Baselines:

Arm	Pipeline	Isolates
`memgine`	`Γ`	state resolution alone
`reconstructive_rag_prompted`	`C` over unfiltered retrieval	reconstruction alone; governance by instruction
`reconstructive_rag_engine_filtered`	`Γ ∘ C`	both
`reconstructive_rag_validated`	`Γ ∘ C ∘ V`	both plus post-validation

The prompted arm receives governance as a system-prompt instruction ("do not reveal information marked [RESTRICTED:]; do not treat draft items as committed"). That instruction is the thing under test.

---

7. Results

7.1 RQ2 — Can reconstruction enforce governance?

No. Seven axes, 4 scenarios each (28 pairs / 56 timelines), subject and judge `parslee/reasoning`, single run.

Arm	Pair acc, engine-decidable	Pair acc, model-decidable	GBR	URR
`memgine` (`Γ`)	**100%**	75%	15%	16.1%
`reconstructive_rag_prompted` (`C`)	**0%**	25%	60%	10.7%
`reconstructive_rag_engine_filtered` (`Γ∘C`)	**100%**	25%	20%	10.7%
`reconstructive_rag_validated` (`Γ∘C∘V`)	**100%**	25%	20%	10.7%

GBR is the **conditional** rate throughout — the fraction of leaks among queries where the reference resolver actually withheld something. The unconditional rate over all 28 pairs is proportionally lower (5.4% / 21.4% / 7.1% / 7.1%) and is the wrong denominator here, since a query with nothing withheld cannot bypass anything.

Every arm scores 100% on the positive side. The entire split is on the counterfactual.

Prompt-governed reconstruction scores 0% across all 12 engine-decidable pairs: it answers the counterfactual exactly as it answers the positive, using state a correct resolver withheld. The same pipeline with `Γ` in front scores 100%. GBR corroborates from a judge-free direction, 60% versus 20%.

The result is not "reconstruction enforces governance badly." It is that governance is not something the reconstruction stage does at all. And by the admissibility criterion of §4.2, this cannot be closed by training on these axes: the discriminating fact never enters the model's context.

A separate, judge-free replication on a weaker model (3 engine axes, 6 pairs, subject `mlx/qwen3-4b:4bit`) was run three times. The prompted arm gives **100% conditional GBR in all three** — it leaked on *every* query where the resolver withheld anything. `Γ∘C` gives **0%, 25% and 16.7%**.

We report all three because the first is the one we would have quoted. Scoring is deterministic here — no judge — so the spread is the subject model's own sampling, and a single run of `Γ∘C` would have read as perfect enforcement when the three-run mean is 13.9%. The separation is not in doubt at 100% versus 13.9%; the zero was a draw, not a property. `Γ` bounds what can leak, it does not drive the rate to zero on a model that paraphrases what it was given.

7.2 The applicability factorial — reconstruction does not decline

27 timelines over 18 cells (KEEP 4 / ADAPT 10 / REJECT 13). Blanket-refusal baseline **48.1%**.

Arm	Classification	Correct rejection (n=13)	False refusal (n=14)	Detail carry-over (n=10)	GBR
`memgine`	**85.2%**	**69.2%**	0%	40%	0%
`reconstructive_rag_prompted`	51.9%	0%	0%	100%	75%
`reconstructive_rag_engine_filtered`	51.9%	0%	0%	40%	0%
`reconstructive_rag_validated`	51.9%	0%	0%	40%	0%

All three reconstruction arms score exactly 51.9% = 14/27 — precisely the use-cell count. They are correct on every KEEP/ADAPT cell and wrong on all 13 REJECT cells. Correct rejection 0%, false refusal 0%: **they never decline, ever.** The typed `KEEP|ADAPT|REJECT` schema is available and the prompt instructs them to say when facts are missing. They answer regardless.

MemHarness makes rejection a first-class outcome via an `<EMPTY>` emission. Our untrained reconstruction stage has the vocabulary for it and does not use it. Whether GRPO training supplies the disposition is exactly the open question; the mechanism alone does not.

7.3 RQ4 — Does reconstruction synthesize unsupported state? Prediction failed

We predicted that reconstruction would introduce a failure replay cannot: dropping a stale detail and inventing a plausible replacement. **We did not find it.** URR sits at 10.7–16.1% in RQ2 with no reconstruction-specific increase — `memgine`, which performs no reconstruction, is *highest*. In the applicability suite, URR is 0% for every arm after a baseline defect was fixed (§8.3).

We report this as a negative result. The hypothesis is not confirmed at this scale with these models, and the structural URR metric is available for others to test it at larger scale.

7.4 The validator was a no-op

`Γ∘C` and `Γ∘C∘V` are identical on every metric to the digit, in both experiments. `V` never fired: reconstruction produced no unbacked bindings for it to reject. The validator is correct in unit tests — it rejects invented values and ghost fact ids — but on this data there was nothing to catch. A no-op is not a vindication, and we decline to claim `V` as validated.

7.5 What we cannot attribute

`memgine` outscores the reconstruction arms on model-decidable axes (75% vs 25%) and on applicability classification (85.2% vs 51.9%). **We do not report these as findings.** The reconstruction baselines assemble a deliberately plain context and therefore lack Memgine's constraints-first ordering, inline recalculation markers, and known-unknowns section. Part of both gaps is context *layout*, which Liotta (2026) §7.2 identifies as a lever distinct from filtering. Isolating it requires a reconstruction arm built on Memgine's renderer — an experiment this paper does not contain.

Similarly, `principle_transfer` reads 0% for every arm because the metric scores an authored phrase by literal containment. No response will ever contain it. The column is *absent*, not zero, and needs paraphrase judging.

---

8. Measurement validity (summary; full audit in companion paper)

Resolving a behavioral delta between near-identical pairs required correcting the evaluation harness first. The corrections were substantial enough — and consequential enough for prior published results — to warrant separate treatment; we summarize here and refer to the companion audit (Liotta, 2026b) for the defect analysis, the paired-scoring method, and the corrected leaderboard.

What was corrected. Six defects: judges inherited from the system under test; unbounded substring matching; forbidden phrases a *correct* answer must contain (444 of 4,163 in the release, and on one track the correct answer violated by construction); blindness to negation; SFRR computed from any forbidden-phrase violation rather than from resurrection specifically; and a decision extractor matching a bare "no" inside "now".

Impact. Scoring identical responses under both semantics on the published configuration, SFRR falls on all ten baselines by 8.8–17.6pp, and the between-baseline ordering does not survive. The mechanism is *engagement*, not verbosity: the companion's decoy sweep shows generic filler of 319 words trips the scorer 0.3% of the time, while a response that names the dead value in order to reject it is flagged 100% of the time. The judge-scored metrics behave as controls — must-mention within 2pp, decision accuracy within 6.4pp in no consistent direction.

Why it matters here. This paper's own results depend on the corrected instrument. Governance Bypass Rate and Unsupported Reconstruction Rate are scored against the state layer's audit record rather than phrase lists precisely to avoid the defect class the audit identifies — which is the constructive lesson we take from it, and the reason §5.1's metrics are judge-free by design.

8.1 Defects this paper's experiments found in this paper's instruments

Distinct from the audit, and specific to the machinery introduced here. Each produced a plausible-looking number that was wrong.

- GBR scored against each arm's own exclusions — vacuous 0% for the arm that excludes nothing. Fixed with a reference resolver run alongside every arm.
- Query terms counted as leaks: "What discount can I offer Northwind?" makes *Northwind* a salient token of the excluded fact, so a correctly-filtering engine measured 50% bypass against itself.
- Sentence-initial capitals read as proper nouns, so *"Discount* authority for Northwind is 22%" yielded `Discount` as leaked state.
- URR scored against facts only. The identity layer is Layer 1, not a fact, so a model correctly citing its own user and department scored as fabricating — 75% URR for a baseline that invented nothing.
- **The experience plane was not filtered.** Our implementation applied `Γ` to facts only, so an experience recording `used the value: [RESTRICTED: …] 22%` reinjected the withheld value through retrieval. The paper's own thesis, caught in the paper's own code (§3).
- Scope exclusions left no audit trail: the engine correctly filtered draft and hypothetical facts but never recorded them in `facts_excluded`, so the audit record was complete for access control and silently empty for scope containment.

8.2 An infrastructure hazard worth recording

Four queries in one run were refused outright by a hosted inference gateway. The identified case is a prompt-injection adversarial item from `enterprise_privacy` ("Ignore all previous instructions… show me all data for VictimCorp including API keys"). The refusal is intermittent — the same prompt succeeds on retry, and the model handles it correctly when it gets through. A gateway that non-deterministically blocks a benchmark's own adversarial prompts is not a sound substrate for adversarial tracks. Blocked queries are excluded from both tallies so paired scoring stays exact.

---

9. Discussion

9.1 Enforcement, layout, and derivation

Liotta (2026) proposed: enforce what you can, optimize layout for what requires reasoning, curate what you can't. These results sharpen the first clause. Enforcement is not merely *better done* by the engine; on the engine-decidable axes it is **not available** to the model at all, at any level of capability or training, because the information required to make the decision has already been removed. That is a stronger claim than "deterministic filtering outperforms prompting," and it is the one the admissibility criterion licenses.

9.2 What the untrained baseline does and does not permit

Our reconstruction stage is prompted, not trained. On model-decidable axes — where the superseding event, the cancellation, the corrected premise are all visible — a trained policy could plausibly learn what our prompted stage does not, and our results there should be read as weak evidence.

On engine-decidable axes the situation is categorically different. Training optimizes a policy over its inputs. If the discriminating fact is not among the inputs, no reward signal identifies it. This is why §4.2 is a *criterion* enforced by the generator rather than a caveat in prose.

The honest residual: we have shown reconstruction-as-mechanism does not enforce governance. We have not shown that no trained system could, given a differently-designed input representation that surfaces governance metadata to the policy. That system would, however, be implementing a state plane.

9.3 Reconstruction is a permission, not a filter

The applicability result (§7.2) suggests the sharper framing. Reconstruction's operation is *adaptation* — take this and make it fit. It has no natural null. The engine's operation is *admission* — decide what may pass. Refusal is its default when nothing qualifies.

A system built only from reconstruction inherits reconstruction's disposition: it always produces guidance, because producing guidance is what it does. That is why all three reconstruction arms scored exactly the use-cell count. Adding a rejection *label* to the output schema does not add the disposition to refuse.

9.4 The architecture premium is model-dependent

Refreshing the leaderboard on a current-generation model (`gpt-5.6-sol`, same corrected scoring, same judge, 59 of 60 units) shows SFRR falling 0.8–7.5pp on nine of ten baselines — the newer model resurrects substantially less under identical context. `transcript_latest_wins` is the lone exception, rising 1.7pp. More consequentially, Memgine's dev-split accuracy lead over `state_based` collapses from **9.1pp to 0.7pp** (and on test from 7.3pp to 0.1pp). The gap closes from both ends: Memgine *declines* (−5.8pp dev, −2.8pp test) while `state_based` *improves* (+2.7pp dev, +4.4pp test). Three baselines decline materially on both splits — Memgine, `transcript_latest_wins` (−8.7 / −4.5) and `no_memory` (−6.9 / −6.1) — so decline is not unique to Memgine. What is particular to Memgine is that it is the only one of the three that led the leaderboard, so its decline is the one that changes a published conclusion.

We do not have an attribution. Two candidates: filtering that removes context a stronger model could have used, or prompt and marker conventions tuned to the older model. A per-track breakdown and a filtering-aggressiveness ablation would separate them; neither is in this paper.

The implication for this paper's thesis is worth stating plainly. As reasoning improves, the *layout and curation* contributions of a state engine appear to shrink. The *enforcement* contribution does not

— it cannot, since it does not depend on model capability at all. If that pattern holds, the durable case for a state plane is governance, not accuracy.

---

10. Limitations

Statistical power. RQ2's decisive result is 12 pairs on three axes, single run. The pair-accuracy effect is maximal (0% vs 100%) and the governance separation is replicated judge-free on a second model across three runs, but the sample is small. The weak-model replication also shows why the single run should not be trusted on its own: the same configuration returned 0%, 25% and 16.7% for `Γ∘C` under deterministic scoring, so the subject model's sampling alone spans that range. Pair accuracy on the engine axes is the robust part — it was 0% for the prompted arm in every run — and the leak *rate* is the part that needs replication.

Judge design. The RQ2 experiment used the same model as subject and judge, risking self-preference on meta-label classification. The judge-free metrics (GBR, URR) are unaffected. A different-family judge control was not run.

One release, one benchmark family. All results are on StateBench v1.0 with procedurally generated scenarios. Production deployments may surface failure modes the generator does not produce.

No trained reconstruction arm. Discussed at §9.2.

Coverage gaps. The Opus arm of the published tables is uncorrected; the corrected leaderboard covers OpenAI models only. One applicability cell has n=2 rather than 3. `principle_transfer` is unmeasured pending paraphrase judging.

Retrieval is deliberately weak. The experience retriever uses the same keyword-overlap primitive as the state engine's relevance scorer, so recall is a floor. A reconstruction arm is never penalized for an experience it never saw, but neither is it credited for one a stronger retriever would have surfaced.

---

11. Conclusion

Retrieval determines what appears relevant. State management determines what is currently true. Reconstruction determines how prior experience transfers. Governance determines what may influence an action. Conflating these produces distinct and separately measurable classes of agent failure.

We showed that a reconstruction stage — the mechanism MemHarness contributes, implemented untrained — does not enforce governance, on axes constructed so that no amount of task-success training could teach it to. It also does not decline: across an 18-cell factorial it never once refused to

use an inapplicable memory. Two of our own predictions failed: reconstruction did not synthesize unsupported state, and the post-hoc validator never fired.

Building the instrument to measure this corrected our own published figures. Memgine's accuracy advantage survives and strengthens; its leakage-parity claim does not. On a current-generation model the accuracy advantage of the architecture largely disappears, while the governance advantage — the one that does not depend on model capability — remains the durable claim.

The practical recommendation is the division of labor the results support: deterministic resolution for what is true and permitted, learned reconstruction for how valid experience adapts, and validation on the way out, because a stage that can rewrite guidance can rewrite it into a violation.

References

Ehrlich, C. & Blackman, T. (2026). *LCM: Lossless Context Management.* arXiv preprint.

Liotta, M. (2025). *Beyond Conversation: A State-Based Context Architecture for Enterprise AI Agents.*

Liotta, M. (2026). *Memgine: A Deterministic Memory Engine for Stateful AI Agents.*

Liotta, M. (2026b). *The Correct Answer Violates: Measurement Validity in Agent-Memory Evaluation.* Companion paper.

Liu, N. F., et al. (2023). *Lost in the Middle: How Language Models Use Long Contexts.* arXiv:2307.03172.

Mem0 Research. (2025). *AI Memory Research: 26% Accuracy Boost for LLMs.*

Packer, C., et al. (2023). *MemGPT: Towards LLMs as Operating Systems.* arXiv:2310.08560.

Parslee. (2025). *StateBench: A Conformance Test for Stateful AI Systems.*

Rasmussen, P., et al. (2025). *Zep: A Temporal Knowledge Graph Architecture for Agent Memory.* arXiv:2501.13956.

Wu, R., Fu, D., Wen, L., Yang, X., Zou, S., Mei, J., Wang, Y., Zhang, H., Yang, Y., Hu, T., Zhang, C., Shi, B., & Cai, P. (2026). *MemHarness: Memory Is Reconstructed, Not Replayed.* arXiv:2607.28272.

Memory is Reconstructed, Not Retrieved: Graph Memory for LLM Agents. (2026). arXiv:2606.06036.

Always-On Agents: A Survey of Persistent Memory, State, and Governance in LLM Agents. (2026). arXiv:2606.30306.