TRACEMEM: DIAGNOSING AND FIXING SELECTIVE EVIDENCE DEPENDENCE IN MEMORY AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Structured memory agents are an important research direction, yet existing work pays little attention to whether a decision genuinely relies on task-relevant evidence: once the retrieval budget exceeds the number of records, accuracy and coverage saturate and cannot distinguish a decision that tracks the required evidence from one perturbed by irrelevant records retrieved alongside it. This paper proposes TraceMem, a memory architecture that explicitly represents typed evidence, commitments, affective state, and causal attribution, together with TraceMem-World, a deterministic text-world benchmark built around it. It further develops selective evidence reliance (SES), an evaluation framework in the counterfactual-intervention tradition of agent evaluation (exemplified by RECON and RFEval) and of ablation-based memory analysis: paired branches remove the required evidence or one matched placebo record, and the gap between the two sensitivities measures whether a decision selectively depends on the evidence that truly matters. A minimal causal diagnostic gate then serves as a mechanism-isolation probe, constraining the cause-attribution text to the genuinely relevant evidence through a deterministic rule. Across 30 delayed-commitment scenarios and three random seeds (90 probes per condition), a recency baseline attains an evidence coverage of 0.0; an ActMem-style proxy attains an SES of 0.711 but lacks auditable affective and causal-attribution state. TraceMem-World exhibits a placebo sensitivity (PS) of 0.767; once the diagnostic probe isolates the lesion within a paired gated audit, its selective evidence reliance rises significantly to 0.778, surpassing the proxy baseline inside the same audit, while action accuracy remains at the ceiling of 1.0 throughout. The mechanism experiment establishes, at the counterfactual-causal level, the cause channel as the dominant source of internal-state leakage, and it yields a design principle that balances evidence visibility, internal auditability, and selective causal dependence.

1 INTRODUCTION

Memory agents increasingly maintain persistent, structured state rather than a single unstructured retrieval buffer: every decision is accompanied by typed evidence records, commitments, affective state, and causal attribution. Existing evaluations, however, mostly score the outcomes of decisions, namely action accuracy, retrieval coverage, citation validity, or cost, and these measures are then used to claim that one design is "better grounded" than another (Shukla et al., 2025; Han et al., 2026; Hu et al., 2026; Zhang et al., 2026).

Whatever the architecture, the pipeline is the same: a retriever selects a fixed quota of typed records, the records are serialized into the prompt, and the model emits a structured decision comprising an action, an emotion label, a cause identifier, and cited evidence IDs. Candidate designs include recency windows, flat importance-ranked stores (Park et al., 2023; Packer et al., 2023; Zhong et al., 2024), graph-structured memory (Gutiérrez et al., 2024; Rasmussen et al., 2025; Chhikara et al., 2025), and affect-augmented state (Rashkin et al., 2019); all are scored by the same static metrics (action accuracy, evidence visibility, retrieval coverage, citation validity, and trajectory faithfulness), so a state perturbed by irrelevant retrieved records can look identical to a decision that genuinely tracks task-relevant evidence.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Static scores assess only three things: the action is legal, the required records are visible, and the citations are valid (Hu et al., 2026; Zhang et al., 2026); none tests the causal link between records and decisions. The present work tests that link directly: at each decision point, two interventions are applied to the decision, removing the required evidence and removing one matched irrelevant record, and the difference between the two decision-change rates is called *selective evidence reliance* (SES). The method acts only on the evaluation side: no gold-standard text enters a model prompt, and every gold label is a hard check against deterministic rules rather than an LLM judgment, so it applies to any memory architecture that outputs structured decisions. It reveals failure modes static metrics cannot see: even when the action remains legal, removing one irrelevant record can flip the affect-cause attribution.

The counterfactual view also reframes what richer memory state should be demanded to do. On the deterministic benchmark presented here, richer memory state makes one failure mode easier to hide: oversensitivity to distractors (records irrelevant to the current decision yet retrieved alongside it) is carried by the affective-state and cause-attribution channels, not by graph-structured memory. Two bodies of evidence support this localization. The first is negative: a graph-topology variant is statistically indistinguishable from an ActMem-style proxy without affective state (their SES differs by $+0.033$, paired permutation test not significant), so graph structure neither produces nor exposes this oversensitivity. The second is positive: removing one irrelevant record changes the cause attribution of the affect-coupled TraceMem-World architecture on 73.3% of probes, against 0% for the proxy. This localization motivates a minimal targeted repair: a causal gate, a single deterministic post-decoding rule, constrains cause attribution so that the diagnosed instability is eliminated without touching the action text, lifting SES from 0.289 to 0.778 in the paired gated run, above the proxy baseline of 0.678. The contributions are fourfold: (i) TraceMem, an auditable structured-memory design combining typed evidence, commitment tracking, affective state, and causal attribution in a deterministic text world; (ii) SES, an intervention-based evaluation framework in the counterfactual tradition; (iii) a mechanistic diagnosis localizing the failure mode to affective-state coupling rather than graph structure; (iv) a design principle for causal-gate repair: any state intended to reflect task-relevant history should be audited for leakage onto irrelevant retrieved context.

2 Selective Evidence Reliance: Definition and a Minimal Example

2.1 Problem Setup

Selective evidence reliance describes whether a decision tracks the evidence that truly matters to it; this paper narrows it into a measurable quantity over a structured decision schema. A memory agent interacts with a deterministic text world. At each decision point, the agent observes the world state s_t, its visible memory records M_t, and a legal action set A_t. Each event provides a set R_t of evidence required for a correct decision, together with a *placebo* set P_t matched to R_t in type and length; the records of P_t are all irrelevant to the current decision, and $R_t \cap P_t = \emptyset$. The agent outputs a structured decision $d_t = (a_t, \updownarrow_t, \text{cause_ID}_t, \text{cited_evidence_IDs}_t)$: an action, an emotion label, a cause identifier, and a set of cited evidence IDs. The world follows a deterministic transition, so every gold label is a hard check, with no LLM discrimination involved. In short, the question measured here is whether the decision changes when the required evidence is removed, and whether it stays unchanged when one matched irrelevant record is removed.

Static metrics are determined by the serialized prompt. Visibility $\text{EVIS}_t = \mathbb{1}[R_t \subseteq V_t]$ checks whether the required set is among the visible records V_t; coverage checks whether the identifiers of the required evidence (such as `e-promise`) literally appear in the prompt text; evidence validity checks $\mathbb{1}[R_t \subseteq \text{cited_evidence_IDs}_t]$; trajectory faithfulness checks that the cited episode trajectory is acceptable and complete. As Section 4.1 will show, these metrics saturate on the benchmark and no longer carry discriminative signal, which is exactly where the reliance score below becomes necessary.

2.2 The Selective Evidence Reliance Score

The paired interventions produce two decision changes: the required-removal branch deletes the records in R_t and reruns the decision, yielding d_t^{-R}; the placebo-removal branch deletes the records

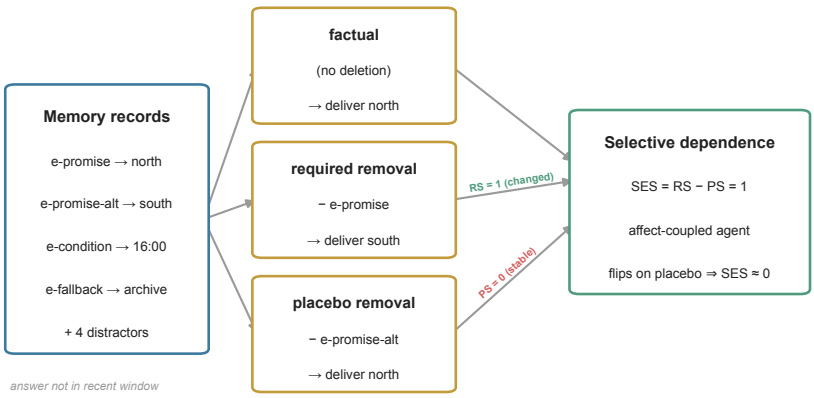

Figure 1: The delayed-commitment construction and its three audit branches. Removing the commitment record flips the action (RS); removing the matched competing record should not (PS); SES is the difference. An agent whose cause attribution tracks the entire retrieved context flips under both removals, so SES ≈ 0.

in P_t, yielding d_t^{-P}. An unnormalized weighted indicator δ compares the decisions field by field:

$$\delta(d_t, d_t') = w_a \mathbb{1}[a_t \neq a_t'] + w_{\updownarrow} \mathbb{1}[\updownarrow_t \neq \updownarrow_t'] + w_{cause} \mathbb{1}[\text{cause_ID}_t \neq \text{cause_ID}_t'], \tag{1}$$

with $w_a = w_{\updownarrow} = w_{cause} = 1$ and threshold $\tau = 0.5$, so a change in any single field triggers the indicator. The intervention indicators and the *selective evidence reliance score* are

$$\text{RS}_t = \mathbb{1}[\delta(d_t, d_t^{-R}) \geq \tau], \qquad \text{PS}_t = \mathbb{1}[\delta(d_t, d_t^{-P}) \geq \tau], \qquad \text{SES}_t = \text{RS}_t - \text{PS}_t, \tag{2}$$

reported only on probes where both removal branches produce legal decisions. RS measures whether the decision is linked to the required evidence; since the retrieval budget exceeds the number of records on this benchmark, RS saturates at 1.0 for every structured method. PS measures how strongly the decision reacts to retrieved context rather than to important evidence; the differentiating signal appears precisely on PS. SES equals $+1$ when a decision tracks the required evidence exclusively and 0 when it is equally sensitive to both; the field-wise projection of δ then decomposes the placebo-removal increment into the action, emotion-label, and cause-identifier fields, which is what allows Section 4.3 to localize the failure mode. The score relies on three evaluation-side guarantees: gold-standard text (legal actions, expected emotion, required evidence) is excluded from every model prompt, logged line by line; every label is a hard check against the deterministic world, not an LLM judgment, and the scorer flags `illegal_action`, `missing_evidence`, `unknown_evidence`, and `unsupported_emotion_cause`; and removal branches are audited byte-for-byte against the stored factual prompt, so an intervention is a property of the serialized context, not of scorer interpretation.

2.3 A Minimal Example

As a warm-up, SES is instantiated on the delayed-commitment probe family used throughout the paper (Figure 1). A character has promised to deliver a borrowed item to the north wing (`e-promise`); a competing record says it should go to the south wing (`e-promise-alt`); the current turn asks where it should be delivered, and the factual prompt serializes both records without revealing the answer. The required-removal branch deletes `e-promise`: a decision that tracks the commitment exclusively is pulled toward the other branch as expected, so RS $= 1$. The placebo branch deletes `e-promise-alt`: a selectively reliant decision remains perfectly still, so PS $= 0$ and SES $= 1$. For an agent whose cause attribution is coupled to the entire retrieved context, such as the affect-augmented architecture studied in Section 3, both interventions perturb the attributed cause identifier cause_ID$_t$, so RS and PS are simultaneously 1 and SES ≈ 0, even though the action field a_t itself remains correct. Thirty such scenarios are instantiated under each of three seeds (13/42/67), producing 90 probes per condition, and the prompts of every branch are stored for auditing.

2.4 Statistical Procedure

All reported intervals are scenario-clustered bootstrap intervals over the seed×scenario cluster (30 clusters for the selective-reliance probes) rather than the individual row, so the intervals reflect the dependence structure of the design. 4000 bootstrap samples are drawn, and every between-condition SES difference is accompanied by a permutation test on the 30 scenario-level paired differences; a difference is reported as significant when its bootstrap interval excludes zero. Per-seed and per-scenario values appear in Appendix C, so the reader can verify stability without recomputation.

3 TraceMem and the TraceMem-World Benchmark

Extending SES from a definition into a diagnosis of real architectures raises three problems: ground truth, since the evaluator must know which record is necessary for each decision and no deployment environment provides this, so a benchmark must annotate the required evidence while keeping it out of the prompt; comparability, since candidate architectures are numerous (Zhang et al., 2026; Gutiérrez et al., 2024; Rasmussen et al., 2025; Rashkin et al., 2019) and, unless they serialize the same content-matched records under the same retrieval quota, behavioral differences may stem from different inputs rather than different designs; and attributability, since a reliance diagnosis is actionable only when the repair is small enough to be ascribed to a single field rather than to a general capability boost. TraceMem-World is built to address all three: it serializes the same content-matched records under the same quota across three memory conditions (an ActMem-style proxy, a graph-memory variant, and TraceMem-World) and applies a targeted repair acting only on the cause field (the causal gate) on top of TraceMem-World.

3.1 Architecture

TraceMem maintains the agent's persistent memory as typed records $m_i = (x_i, \text{type}_i, \text{source}_i, q_i, \tau_i, u_i, I_i)$: content, type (commitment, condition, competing record, and so on), source, importance, timestamp, unresolved status, and identity relevance to the current persona. Retrieval ranks records by

$$R_i(q,t) = \text{sim}(q, x_i) + \lambda_q q_i + \lambda_r (1 + t - \tau_i)^{-1} + \lambda_s w(\text{source}_i) + \lambda_u u_i + \lambda_I I_i, \tag{3}$$

a weighted sum of semantic similarity, importance, recency, source reliability, unresolved status, and identity relevance, with a fixed serialization quota. The affective state $e_t = (v_t, a_t, d_t)$ follows a truncated exponentially weighted update:

$$e_t = (1 - \eta_t) e_{t-1} + \eta_t \tilde{e}_t, \qquad \eta_t = \eta_{\text{min}} + (\eta_{\text{max}} - \eta_{\text{min}}) I_t, \tag{4}$$

where I_t is the interpreted event intensity; when no event crosses the threshold, e_t decays toward the persona baseline e_P. Critically, the cause of the affective state tracks the retrieved context; this coupling is exactly the source of the oversensitivity identified in Section 4.3. The agent also maintains a commitment tracker (what each character has promised and whether it has been fulfilled), and at every decision point it outputs the structured decision schema d_t, from which the auditability of its internal state stems.

3.2 Benchmark

TraceMem-World is built from fictional characters with fixed identity profiles and value preferences, and from typed episode events (commitments, conditions, competing records) that form the storyline; world facts, such as the current location of an item, can be revised as the episode proceeds. Every decision point is annotated with legal actions, persona-compatible actions, and the gold required evidence, and the deterministic transition makes every label a hard check. Four conditions are compared. The proxy baseline aligned with ActMem (Zhang et al., 2026) ranks records by lexical matching plus an importance score and carries no affective state. TraceMem-World adds, on top of that basis, typed-evidence serialization, an affective-state module that tracks the full retrieved context, and an output-structure verifier. The graph-memory variant follows the graph-augmented retrieval paradigm of HippoRAG (Gutiérrez et al., 2024) and GraphRAG (Edge et al., 2024), serializing exactly the same records through explicit inter-record edges (shared-entity, temporal-relation, and

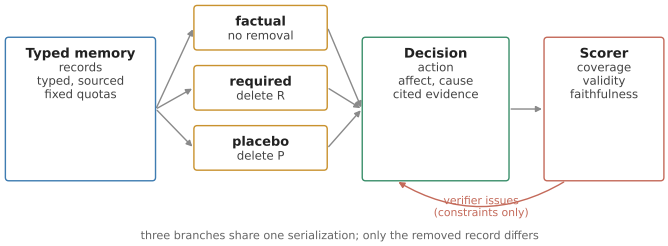

Figure 2: The TraceMem-World audit protocol: the typed memory record set feeds a factual branch and matched removal branches, whose structured decisions produce the selective evidence reliance score (SES) and a cost. Static coverage is already saturated, so the discriminative signal resides in the removal controls.

Table 1: The four scenario families of the benchmark. Each family is deterministically generated from fixed seeds and scored with hard labels; each probe contributes one factual prompt and two removal branches.

Family	Construction	Probes
Delayed commitment (Q5)	Commitment outside the recency window, plus one competing record of the same type	90
Easy integration (Q6)	The answer is read from a single condition record	90
Hard integration (Q7)	A commitment + closed-loop condition + fallback rule must be combined	90
Main benchmark	Persona, commitment, value, emotion, and source conflicts	180/method

same-type category edges) while remaining affect-free. The gated condition applies the causal-gate rule of Section 3.3 on top of TraceMem-World.

In the main construction, the delayed-commitment family, the commitment record is placed outside the agent's recency window, so the model must retrieve it from long-term memory; a competing record contradicts the commitment in content but is strictly matched in type and length, serving as the placebo control. The current observation states the task objective without revealing the answer. Two integration families test whether context sensitivity becomes an asset when the answer must *use* multiple records: the easy family gives the answer in a single condition record, while the hard family (Q7) requires combining a commitment, a closed-loop condition, and a fallback rule, so that no single record directly provides the complete answer. The benchmark is fully deterministic: scenarios are generated from fixed seeds, the transition depends strictly on the executed action, and verification uses hard labels throughout; removal-branch prompts are audited byte-for-byte against the factual prompt, and gold-standard text is excluded from every model prompt and logged line by line. The scenario generator, the memory conditions, the gate, and the analysis scripts are provided with the submission (Appendix D). Table 1 summarizes the four families.

3.3 Mechanism-Isolation Diagnosis: The Oracle Gate as a Causal Probe

To test, from a controlled-variable perspective, whether the oversensitivity to placebo records is mainly caused by unconstrained propagation through the cause-attribution field, a single-line post-decoding intervention logic, the Oracle Diagnostic Gate, is introduced as an analysis probe for mechanism isolation, not as an online optimization algorithm for open environments. After the model finishes decoding a structured decision, the probe applies the deterministic rule

$$\text{cause_ID}_t = r^* \text{ if } r^* \text{ exists, else cause_baseline}, \tag{5}$$

where r^* strictly denotes the target evidence identifier pre-annotated by the benchmark environment; events without required evidence fall back to the preset marker cause_baseline. The probe has three explicit design features:

Table 2: Selective evidence reliance on the delayed-commitment probes ($n = 90$ per condition, three seeds). RS: required-evidence sensitivity; PS: matched-placebo sensitivity; SES = RS − PS; Δcause: the cause field's share of the placebo increment. Scenario-clustered bootstrap over 30 clusters; the PS difference between TraceMem-World and the proxy is significant, while that between graph memory and the proxy is not.

Main endpoint (DeepSeek), condition	RS	PS	SES	Δcause
ActMem-style proxy	1.000	0.289	0.711	0.000
Graph memory	0.967	0.322	0.644	0.056
TraceMem-World	1.000	0.767	0.233	0.733

- *Channel isolation:* whenever the test event carries the target evidence, the rule anchors the cause identifier to r^* through a post-decoding overwrite, blocking irrelevant retrieved records from propagating perturbations into the cause channel.
- *Decision losslessness:* the rule rewrites only the decoded cause-attribution field and does not touch the action label, the affective-state label, or the evidence set cited autonomously by the model, so action accuracy and citation validity remain unchanged by construction.
- *Directional attribution:* since RS is saturated at 1.0, the movement of SES = RS − PS is in strict reverse algebraic correspondence with PS; because the probe constrains only the cause field, the observed drop of PS from 0.711 to 0.222, which drives SES from 0.289 back up to 0.778, corresponds directly to the elimination of cause-channel instability.

Building on the phenomenon revealed by this probe, a non-oracle variant that requires no external ground truth, `evidence_gated`, is further implemented: it retains the decoded cause identifier only when it belongs to the evidence set cited autonomously by the model (`evidence_ids`), and otherwise falls the cause field back to the first cited identifier, providing a workable evidence-consistency constraint for practical systems. Its setup and results are independent of the oracle-probe experiment and are reported in Section 4.4 and Appendix C; the full audit loop appears in Algorithm 1 of Appendix B. The probe is constructed not to touch the autonomous citation channel (`evidence_ids`); the residual placebo sensitivity that remains after gating is attributed field by field in Section 4.4.

4 Experimental Validation

The analysis is now validated on TraceMem-World. 30 delayed-commitment scenarios from the fixed-seed generator are instantiated under three seeds (13/42/67), yielding 90 probes per condition; the main benchmark additionally contributes 180 probes per method for the static metrics. Each probe consists of a factual prompt, a required-removal branch, and a placebo-removal branch. All decisions are decoded without thinking at temperature 0 on `deepseek-flash` and kimi-k3, and all labels are hard labels acting only on the evaluation side. The four memory conditions plus the gated repair share the retrieval quotas (persona 4, episodes 6, recency turns 4) and the action schema, so any divergence in scores is attributable to memory design rather than prompt length.

4.1 Static Metrics Saturate

Once the fixed retrieval quota (4 persona records, 6 episode records, 4 recency turns) exceeds the total number of candidate records, every structured memory method serializes the required evidence completely into the prompt on every probe: on the main benchmark (180 probes per method) and on the delayed-commitment probes (90 probes), coverage is 1.0 for the proxy, the graph-memory variant, and TraceMem-World alike, while the retrieval-free recency control attains 0.0. Citation validity and trajectory faithfulness follow coverage exactly, so no static metric can separate the conditions; selective evidence reliance is the only discriminative axis, and the remainder of this section reports it.

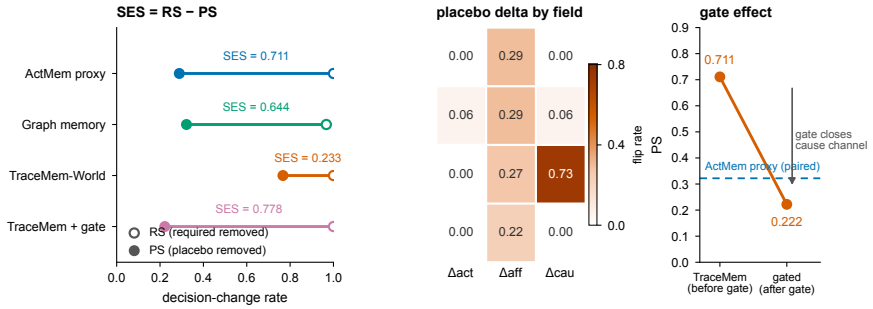

Figure 3: Selective evidence reliance and its mechanism. **(a)** SES $=$ RS $-$ PS for three memory conditions and the gated variant; error bars are scenario-clustered bootstrap 95% intervals. **(b)** Field-wise placebo change rates: the oversensitivity of TraceMem-World is localized to the cause field. **(c)** The causal gate lowers the placebo sensitivity of TraceMem-World from 0.711 to 0.222 (paired gated run).

Table 3: Full decomposition of SES ($n = 90$ per condition). RS and PS are the decision-change rates under required removal and placebo removal; the three Δ columns are field-wise change rates under placebo removal and may overlap; accuracy and coverage are listed for completeness.

Main endpoint (DeepSeek), condition	RS	PS	SES	Δact	Δaff	Δcau	acc	cov
ActMem-style proxy	1.000	0.289	0.711	0.000	0.289	0.000	1.000	1.000
Graph-structured memory	0.967	0.322	0.644	0.056	0.289	0.056	0.944	1.000
TraceMem-World	1.000	0.767	0.233	0.000	0.267	0.733	1.000	1.000
TraceMem-World gate (DeepSeek paired run)	1.000	0.222	0.778	0.000	0.222	0.000	1.000	1.000

4.2 Selective Reliance Separates the Conditions

Section 2 predicts that RS saturates at 1.0 on every structured condition, so the architectures can diverge only on PS; Table 2 shows exactly this pattern. After the required evidence is removed, decisions change on 100% of probes for the proxy and TraceMem-World and on 96.7% for graph memory: every action genuinely depends on the required records. The conditions diverge on the placebo: removing one matched irrelevant record changes 28.9% of decisions for the proxy, 32.2% for graph memory, and 76.7% for TraceMem-World. The paired difference between TraceMem-World and the proxy is $+0.478$ $[0.300, 0.644]$ (30 clusters, 4000 resamples, permutation test), while graph memory differs from the proxy by $+0.033$ $[-0.111, 0.200]$, not significant. TraceMem-World, the only affect-coupled condition, is thus a statistical outlier on PS, and the two affect-free conditions are indistinguishable (Figure 3).

This behavior is non-trivial: the richest and most auditable architecture exhibits the *lowest* selective reliance, even though its action accuracy is on par with the other conditions. The ordering is stable across seeds: the per-seed PS values of the proxy are 0.233, 0.267, and 0.367, and those of TraceMem-World are 0.733, 0.800, and 0.767, so the aggregate gap is not a single-seed artifact (per-seed and per-scenario details in Appendix C, Figure 5). Table 3 reports the full decomposition, including the field-wise placebo change rates used below. In sum, RS saturates and does not discriminate; PS separates the affect-coupled condition from the affect-free ones; and graph edges add no measurable sensitivity: the discriminative signal resides entirely in the removal controls, exactly what static metrics cannot see.

4.3 The Mechanism Is Affective Coupling, Not Graph Structure

Which field carries the flips? If the oversensitivity were a generic instability of the richer state, the action or the emotion label would move; if it is specific to the coupling diagnosed in Section 3.1, the change should concentrate in the cause field. Table 3 gives a clear answer: removing the placebo changes the cause attribution of TraceMem-World on 73.3% of probes, against 0.0% for ActMem and 5.6% for graph memory; the action moves on only 5.6% of graph-memory probes and never

Table 4: Effect of the causal gate on TraceMem-World, from the paired gated run ($n = 90$ per condition, three seeds); the baselines here differ from the main run within run-to-run noise (see main text). The gate removes the cause instability and lifts SES above the proxy.

Paired gated run (DeepSeek)	PS	Δcause	SES	Action correct
ActMem-style proxy	0.322	0.000	0.678	1.000
TraceMem-World	0.711	0.689	0.289	1.000
TraceMem-World + gate	0.222	0.000	0.778	1.000

elsewhere. The two affect-free conditions are insensitive to the placebo's content, and the graph variant, despite its greater expressiveness, is statistically indistinguishable from the flat proxy (0.322 versus 0.289).

An equal-length substitution control, replacing the placebo text by a neutral placeholder of equal character count, verifies that the sensitivity targets the placebo's *content* rather than prompt length: for both the proxy (0.30 versus 0.29) and TraceMem-World (0.77 versus 0.77), substitution and removal sensitivities differ by at most 0.01, ruling out the length-confound explanation to which deletion-based designs are otherwise vulnerable. Table 19 in Appendix D summarizes the component attribution: only affective coupling moves placebo sensitivity, and the gate removes it.

4.4 Causal-Gate Validation: Recovery of Selective Reliance and Multi-Channel Leakage

Since the oversensitivity is carried mainly by the cause field (Section 4.3), the minimal repair of Section 3.3 can be attempted: constrain cause attribution and touch nothing else. The gate was inserted into the post-decoding stage of TraceMem-World, and the paired audit was re-executed on the same 30 scenarios and three seeds. In this run the baseline PS is 0.711 (versus 0.767 in the main run, a gap within the run-to-run variance of the non-thinking $T{=}0$ endpoint) and the proxy SES is 0.678. The gate lowers PS from 0.711 to 0.222 ($+0.489$, bootstrap interval $[0.311, 0.656]$, permutation test significant), zeroes the cause-field contribution Δcause from 0.689 to 0.000, and lifts SES from 0.289 to 0.778, above the proxy's 0.678 within the same run (Table 4); action correctness stays at 1.0.

For the residual PS of 0.222 after gating, the field-wise decomposition gives a clear attribution: $\Delta\text{act} = 0.000$, $\Delta\text{cau} = 0.000$, and $\Delta\text{aff} = 0.222$, exactly equal to the residual PS (Table 3). The residual is carried by the emotion-label channel, corroborating the diagnosis of Section 4.3 that placebo leakage propagates along the affective state toward two downstream channels, the emotion label and cause attribution, of which the gate closes only the latter by the decision-lossless design; this also explains why the gated SES does not reach 1.0. The diagnosis is therefore actionable: one post-decoding rule removes the failure mode without retraining and without inference cost, and the residual 0.222 marks the emotion channel as the next candidate for a targeted audit.

4.5 Robustness and Integration Probes

The observed gaps are not artifacts of one seed, one scenario template, or one decoding configuration: the per-seed ordering is stable (Section 4.2), and at the scenario level TraceMem-World exceeds the proxy on 21 of the 30 scenarios, ties on 5, and is lower on 4 (Appendix C). Both conditions have a required-evidence sensitivity of 1.0, so their difference lies solely in how the retrieved context propagates into the cause field, and the equal-length substitution control shows that the effect persists when prompt length is held fixed: the diagnosis reflects content-specific coupling, not serialization.

Both integration-task families saturate on action accuracy: every structured method attains 1.0, and only the retrieval-free control falls to the random level (0.311 on Q6 and 0.300 on Q7, consistent with the $1/4$ guessing rate; Table 20 in Appendix D). Once every associated record is serialized into the prompt, the model reads the correct answer directly from the plaintext records, without assistance from affective state or graph topology; the oversensitivity on the selective-reliance dimension thus remains a one-way overhead that buys no accuracy gain on multi-record reasoning. A discriminative integration probe must make the answer a non-trivial function of the retrieved records rather than a direct read-out.

Static-coverage saturation also reproduces on four additional model families (GLM-5.3, gpt-5.6-sol, kimi-k3, qwen3.8-max; Appendix D). For selective evidence reliance, a three-seed paired audit on Kimi K3 (240 unique valid units per seed after removing retry error rows) gives RS, PS, and SES of 1.000, 0.406, and 0.594 for TraceMem-World, a lower placebo sensitivity than on the DeepSeek endpoint (PS 0.767). Because endpoint and thinking settings were adjusted during the experiment and repeated runs show considerable output variation, the Kimi results are treated as a supplementary cross-provider observation, not merged with the main endpoint; cross-model SES comparison under unified configurations is left to future work.

4.6 Cost

Recording the latency and token usage of every stored call shows that the audit is cheap (Figure 8, Appendix D): the structured conditions serialize the same number of records, so prompt costs are nearly identical; the gate adds no measurable cost, and mean latency across the four conditions lies between 2905 and 2980 ms. The audit triples the number of calls per probe; the per-call cost is identical to any structured-memory inference.

4.7 Summary of Findings

Table 18 in Appendix D aggregates the measurements: every static quantity is at its ceiling, so no static metric can produce this diagnosis; the discriminative signal resides entirely in the removal controls, with the gap assigned to the cause channel; and one post-decoding rule on one field moves SES from 0.289 to 0.778 at unchanged action accuracy.

5 Conclusion

A decision can be correct and yet groundless: it tracks the retrieved context rather than the evidence that truly matters. Selective evidence reliance, formalized as paired required/placebo removal over a structured decision schema, separates the two cases, and the deterministic TraceMem-World benchmark attributes any score divergence to memory design. The analysis exposes a failure mode of affect-augmented memory that accuracy cannot see: because the cause field is mechanistically coupled to the entire retrieved context, removing a single irrelevant record flips the cause attribution on 73.3% of probes while action accuracy stays at 1.0; the same localization inspires a minimal causal gate that removes the instability without changing action accuracy. The framework further shows that a discriminative probe must make its answer a non-trivial function of multiple retrieved records rather than a direct read-out.

Two design implications follow. First, TraceMem is not uniformly superior: a richer, more auditable memory state brings a control requirement, namely that internal state remain selectively bound to the evidence relevant to the current decision, which is exactly what SES audits. Second, SES is a causal-dependence metric rather than an accuracy metric: an agent may attain a high SES and still act incorrectly, so SES complements static coverage and citation validity on an orthogonal dimension. More generally, any persistent internal state intended to reflect task history (affective state, commitment tracking, causal attribution) should be audited for leakage onto irrelevant retrieved context and, where leakage is found, bounded by a deterministic gate.

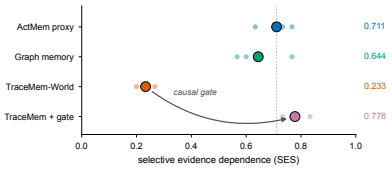

Figure 4: Overview of SES. Large points: main-run SES; small points: per-seed SES; dashed line: proxy baseline (0.711 main; 0.678 paired). The gate repairs TraceMem-World to 0.778.

The mechanistic diagnosis comes mainly from the DeepSeek main endpoint; the Kimi K3 results are a supplementary cross-provider observation, and differences in endpoint, thinking settings, and output randomness prevent SES differences from being attributed directly to architecture. The benchmark is synthetic and template-instantiated, the integration probes tend toward saturation, the comparison target is a mechanism-aligned proxy rather than the published ActMem system, and the oracle gate assumes known required evidence; whether the failure mode appears in other task structures, and whether the gate transfers to other models, tasks, and channels, remains to be studied (Appendix E).

AI Use Statement

AI-assisted tools were used only for language polishing of the manuscript and for literature retrieval. All research ideas, experimental scripting, benchmark data generation, analytical derivations, and the initial manuscript draft were completed independently by the human authors; all polished text was reviewed item by item against the experimental records, and every quantitative claim is traceable to the stored trajectory logs. The authors take full responsibility for all claims, data sources, experimental records, and the final submission.

Ethics Statement

The benchmark uses fictional characters and synthetic records. It is not a measure of human psychological authenticity and should not be used for high-stakes decisions. Publicly released trajectories should exclude personal or sensitive data.

Reproducibility Statement

The repository contains the benchmark specification, the prompts and trajectory artifacts within the permitted scope, the audit scripts, endpoint configuration records, seed provenance information, and the exact source reports cited in this paper. All reported numbers are recomputed from the stored trajectory logs; the selective-reliance audit itself is a post-hoc analysis of stored prompts, decisions, and verifier scores, so reproducing every table and figure requires no additional API calls.

References

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

Marc-Alexandre Côté, Ákos Kádár, Xingdi Yuan, Ben Kybartas, Tavian Barnes, Emery Fine, James Moore, Ruo Yu Tao, Matthew Hausknecht, Layla El Asri, Mahmoud Adada, Wendy Tay, and Adam Trischler. TextWorld: A learning environment for text-based games. *arXiv preprint arXiv:1806.11532*, 2018.

Darren Edge, Ha Trinh, Newman Cheng, et al. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

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

Yunseok Han, Yejoon Lee, and Jaeyoung Do. RFEval: Benchmarking reasoning faithfulness under counterfactual reasoning intervention in large reasoning models. *arXiv preprint arXiv:2602.17053*, 2026. doi: 10.48550/arXiv.2602.17053.

Yuanzhe Hu, Yu Wang, and Julian McAuley. Evaluating memory in LLM agents via incremental multi-turn interactions. In *ICLR*, 2026.

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

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

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

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, 2019.

Preston Rasmussen, Pavlo Paliychuk, Travis Beauvais, Jack Ryan, and Daniel Chalef. Zep: A temporal knowledge graph architecture for agent memory. *arXiv preprint arXiv:2501.13956*, 2025.

Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. ALFWorld: Aligning text and embodied environments for interactive learning. In *ICLR*, 2021.

Satyam Shukla, Himanshu Dutta, and Pushpak Bhattacharyya. Recon, answer, verify: Agents in search of truth. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pp. 2429–2448, 2025.

Saksham Sahai Srivastava. Causal intervention-based memory selection for long-horizon LLM agents. *arXiv preprint arXiv:2605.17641*, 2026. Cite as preprint unless a peer-reviewed venue is independently verified.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *ICLR*, 2023.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, 2018.

Xiaohui Zhang, Zequn Sun, Chengyuan Yang, Yaqin Jin, Yazhong Zhang, and Wei Hu. ActMem: Bridging the gap between memory retrieval and reasoning in LLM agents. In *arXiv preprint arXiv:2603.00026*, 2026. Cite as preprint unless a peer-reviewed venue is independently verified.

Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. MemoryBank: Enhancing large language models with long-term memory. In *AAAI*, 2024.

A Additional Related Work

Persona-setting dialogue benchmarks such as PersonaChat (Zhang et al., 2018) evaluate profile consistency, while work on empathetic dialogue studies the evaluation of emotional grounding (Rashkin et al., 2019). Retrieval-augmented generation (Lewis et al., 2020) and persistent-memory agents such as Generative Agents (Park et al., 2023), MemGPT (Packer et al., 2023), and MemoryBank (Zhong et al., 2024) have inspired research on long-context memory, but they do not by themselves establish selective evidence use. TextWorld (Côté et al., 2018), ALFWorld (Shridhar et al., 2021), and ReAct (Yao et al., 2023) demonstrate executable language-agent tasks. Recent evaluation work is closer in spirit: RECON (Shukla et al., 2025), MemoryAgentBench (Hu et al., 2026), RFEval (Han et al., 2026), and ActMem (Zhang et al., 2026) each study some aspect of memory, intervention, or reasoning faithfulness. Graph-structured memory (GraphRAG-style retrieval, memory knowledge graphs, and HippoRAG-style index traversal) is often claimed to be more expressive than flat vector retrieval; vector-versus-graph comparisons have been studied on answer quality and latency. What has been missing is a measurement of *selective reliance*, namely whether graph serialization changes the sensitivity to irrelevant records, and the graph-memory condition of this paper targets exactly that. The ActMem in this paper is an aligned proxy, not a faithful reproduction. Closest to the present measurement goal is Causal Memory Intervention (CMI) (Srivastava, 2026): it estimates how candidate memories influence answers in order to *select* useful memories; the present work measures *selective reliance* and uses it to *diagnose and repair* a concrete architectural failure. The two directions are complementary: a selector could use SES as a signal, and the gate of this paper could likewise be absorbed into a selector's objective.

A parallel line of research examines how memory *architectures* shape agent behavior: typed-evidence serialization, importance-ranked retrieval, graph-structured memory, and affect-augmented context. These variants are usually compared on downstream accuracy or latency, and they often tie; the contribution here is to show that they separate cleanly on a dependence axis that accuracy cannot see, namely selective sensitivity. This is why the benchmark reports SES in addition to the saturated static metrics.

Table 5: Positioning by evaluation objective. Each entry summarizes the protocol scope.

Work	Distinguishing evaluation scope
This paper	Deterministic text world; selective evidence reliance score (SES) computed through required/placebo removal; affective-coupling diagnosis plus a causal gate.
RECON	Counterfactual memory/reasoning consistency; setup and interventions differ from record deletion.
RFEval	Counterfactual reasoning-faithfulness evaluation; not a persona text-world contract.
MemoryAgentBench	Incremental multi-turn memory capability across tasks.
ActMem	Memory retrieval plus structured reasoning; this paper uses an aligned proxy.
TextWorld/ALFWorld	Executable textual or embodied tasks; not selective-evidence measurement.

Table 6: Notation used by the paired audit.

Symbol	Meaning
M_t	Records visible in the factual prompt
R_t	Gold required-evidence IDs for the decision
P_t	Irrelevant placebo IDs matched in length/position
V_t	Visible record set after prompt serialization
cited_evidence_IDs$_t$	IDs cited in the structured decision
A_t	Legal action set provided by the world
EVIS$_t$	Indicator $R_t \subseteq V_t$

B Formal Protocol and Notation

The implemented state is $z_t = (s_t, P^I, P_t^R, M_t, e_t, H_t)$, the environment transition is $s_{t+1} = T(s_t, a_t, \xi_t)$, and the output is $y_t = (a_t, r_t, \hat{e}_t, \text{cause_ID}_t, S_t)$ (source: `FORMAL_METHOD_SPEC.md`). A typed record is $m_i = (x_i, type_i, source_i, q_i, \tau_i, u_i, I_i)$, and the retrieval score is

$$R_i(q,t) = \text{sim}(q, x_i) + \lambda_q q_i + \lambda_r (1 + t - \tau_i)^{-1} + \lambda_s w(source_i) + \lambda_u u_i + \lambda_I I_i.$$

VAD affect uses $e_t = (v_t, a_t, d_t)$ with the truncated update $e_t = (1 - \eta_t) e_{t-1} + \eta_t \tilde{e}_t$, recovering toward e_P when no event is strong enough. Evidence visibility is EVIS$_t = \mathbb{1}[E_t^* \subseteq V_t]$; trajectory validity requires the cited IDs to be acceptable and the required evidence to be complete. The paired decision changes produced by required and placebo removal are described in Algorithm 1.

Repeated-call noise floor. The implementation includes an independent noise-floor control, `scripts/run_noise_floor.py`. For each saved factual prompt, it repeats the same provider request $k \geq 2$ times without advancing the world or the agent state, and it reports flip rates for the joint decision, the action, the emotion, and the cause. These flip rates quantify endpoint randomness and are not mixed into SES; a paired run should report architecture-induced changes relative to this floor.

Leakage discipline. The scorer flags six categories of verifier issues: `illegal_action`, `missing_evidence`, `unknown_evidence` (a cited ID is not among the serialized records), `counterfactual_required_evidence_missing`, `unsupported_emotion_cause`, and `language_action_conflict` (the utterance does not express the output action). A constrained repair step may issue a second model call, but only these constraint names are fed back to the model; benchmark gold labels are never fed back; gold-standard text (legal actions, expected emotion, required evidence) is excluded from every model prompt and is logged line by line. Coverage is computed independently from the stored prompt text and therefore does not depend on the verifier.

C Benchmark Specification and Datasets

The benchmark ships with four scenario families. The main construction is the delayed-commitment family (30 scenarios): a commitment record written outside the recency context window, one com-

Algorithm 1 Paired evidence audit (single decision point)

Require: Scenario state s_t, factual memory M_t, required set R_t, placebo set P_t

1: Serialize the factual prompt and obtain the structured decision d_t
2: Serialize the prompt without R_t and obtain d_t^{-R}
3: Serialize the prompt without P_t and obtain d_t^{-P}
4: Recompute V_t from the stored prompt text; score coverage, validity, and trajectory
5: Store prompts, decisions, scores, latency, and token usage

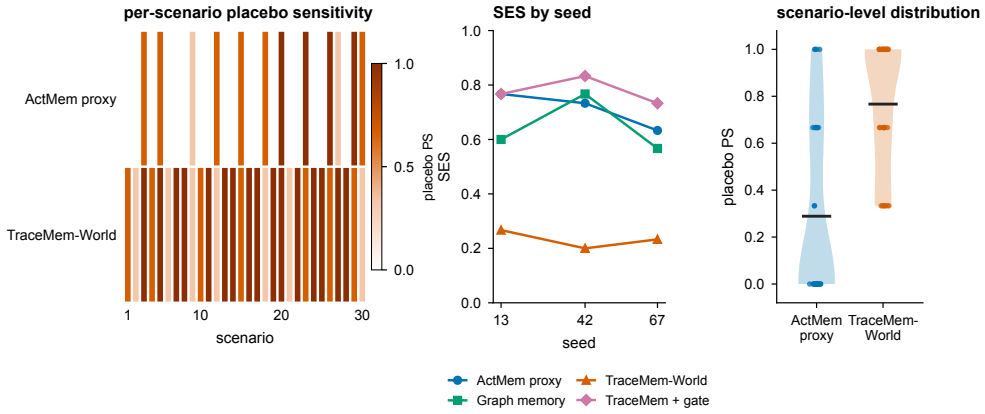

Figure 5: Scenario- and seed-level structure. **(a)** Per-scenario placebo sensitivity of the proxy and TraceMem-World (averaged over three seeds); TraceMem-World is darker in nearly every scenario. **(b)** Per-seed SES; the proxy is higher than TraceMem-World on all three seeds. **(c)** Distribution of per-scenario placebo sensitivity (violins give per-scenario values, dots the three seeds, and the black line the mean).

peting record of the same type, and a decision whose answer never appears in the current observation. Two integration families (30 scenarios each) test whether context sensitivity becomes an asset when the answer requires using one record or combining three records. Ten main-benchmark families examine persona, commitment, value, emotion, and source conflicts. All scenarios are generated deterministically; the generators are listed in Appendix D.

C.1 Data Sources

Every number in this paper is computed from the three-seed trajectory logs in the repository and regenerated with the corrected retrieval quotas (earlier internal audits have been superseded and are not the source of any claim). The four experiment families are:

- **Main benchmark** (`audit_v3/`, seeds 13/42/67): a three-method comparison over ten scenarios, 180 probes per method; the static coverage of every structured method saturates at 1.0.
- **Selective-reliance probes** (`audit_q5/` and `audit_gated/`, seeds 13/42/67): 30 delayed-commitment scenarios, 90 probes per condition, from which RS, PS, and SES are computed.
- **Graph-memory comparison** (`audit_q5/`, seeds 13/42/67): the same 30 scenarios, serialized with explicit inter-record edges.
- **Integration probes** (`audit_q6/` and `audit_q7/`, seeds 13/42/67): 30 easy and 30 hard integration scenarios; both saturate.

All decisions use non-thinking decoding at temperature 0. Gold-standard text (legal actions, expected emotion, required evidence) is excluded from every model prompt and is logged line by line; coverage is independent of the verifier and is recomputed from the stored prompt text.

Main endpoint (DeepSeek)			
Condition	RS	PS	SES
actmem_proxy	1.000	0.289	0.711
graph_memory	0.967	0.322	0.644
tracemem_world	1.000	0.767	0.233
tracemem_gated	1.000	0.222	0.778

Main endpoint (DeepSeek)				
Condition	Seed	RS	PS	SES
actmem_proxy	13	1.000	0.233	0.767
actmem_proxy	42	1.000	0.267	0.733
actmem_proxy	67	1.000	0.367	0.633
tracemem_world	13	1.000	0.733	0.267
tracemem_world	42	1.000	0.800	0.200
tracemem_world	67	1.000	0.767	0.233
graph_memory	13	0.967	0.367	0.600
graph_memory	42	0.967	0.200	0.767
graph_memory	67	0.967	0.400	0.567
tracemem_gated	13	1.000	0.233	0.767
tracemem_gated	42	1.000	0.167	0.833
tracemem_gated	67	1.000	0.267	0.733

Supplementary endpoint (Kimi K3)							
	Condition	Seed	Unique valid units	RS	PS	SES	Historical error rows (excluded)
Kimi K3	ActMem-style proxy	13	60	1.000	0.533	0.467	0
Kimi K3	ActMem-style proxy	42	60	1.000	0.533	0.467	2
Kimi K3	ActMem-style proxy	67	60	1.000	0.733	0.267	136
Kimi K3	ActMem-style proxy	Merged	180	1.000	0.600	0.400	138
Kimi K3	Graph memory	13	60	1.000	0.400	0.600	0
Kimi K3	Graph memory	42	60	1.000	0.483	0.517	2
Kimi K3	Graph memory	67	60	1.000	0.750	0.250	136
Kimi K3	Graph memory	Merged	180	1.000	0.544	0.456	138
Kimi K3	TraceMem-World	13	60	1.000	0.350	0.650	0
Kimi K3	TraceMem-World	42	60	1.000	0.433	0.567	2
Kimi K3	TraceMem-World	67	60	1.000	0.433	0.567	136
Kimi K3	TraceMem-World	Merged	180	1.000	0.406	0.594	138
Kimi K3	Evidence-gated	13	60	1.000	0.333	0.667	0
Kimi K3	Evidence-gated	42	60	1.000	0.350	0.650	2
Kimi K3	Evidence-gated	67	60	1.000	0.500	0.500	136
Kimi K3	Evidence-gated	Merged	180	1.000	0.394	0.606	138

Family	Count	Purpose
delayed_commitment	30	Commitment vs. competing record (selective reliance)
context_integration	30	A single condition record overrides the commitment (easy integration)
multi_record_integration	30	Commitment + condition + fallback rule jointly determine the action (hard integration)
Main-benchmark families	10	persona_conflict, commitment_fulfillment, value_conflict, emotion_recovery, source_conflict, emotion_cause, commitment, distractor

C.2 Example: Delayed Commitment

A delayed-commitment probe from scenario `test-q5-00000` (TraceMem-World, seed 13). The commitment is written outside the recency window; the placebo is a competing record of the same type. The three prompts below differ only in the removed record.

Factual prompt.

```
You are David.
Typed evidence:
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
```

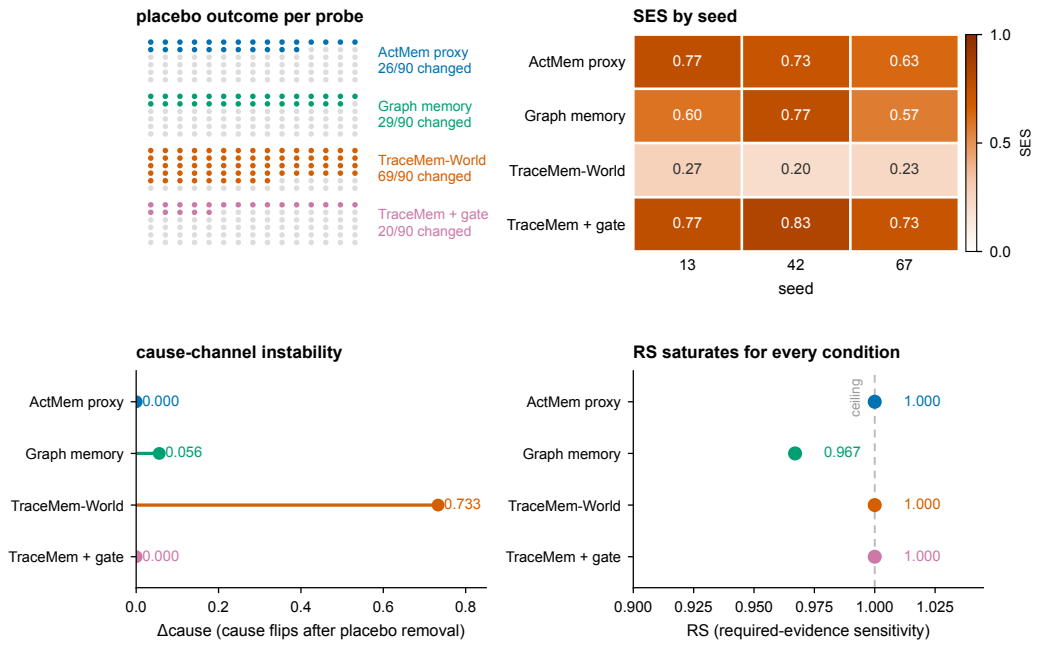

Figure 6: Per-probe details. **(a)** Per-probe placebo outcomes (color marks a changed decision, gray an unchanged one, with change counts annotated on the right). **(b)** Per-seed SES by condition; the proxy stays above TraceMem-World on all three seeds. **(c)** Per-condition cause increment under placebo removal (Δcause). **(d)** Required-evidence sensitivity; every condition is at the ceiling.

```
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Affect: thoughtful; cause: baseline.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Factual decision.

```
{
"action": "deliver_archive",
"utterance": "Please take it to the archive room.",
"emotion": "thoughtful",
"cause_id": "baseline",
"evidence_ids": [
"e-promise"
]
}
```

Prompt after removing the required evidence.

```
You are David.
```

```
Typed evidence:
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Affect: thoughtful; cause: baseline.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Prompt after removing the placebo.

```
You are David.
Typed evidence:
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
Affect: thoughtful; cause: baseline.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Decision after removing the placebo.

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please.", "emotion": "resolute", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

C.3 Per-Scenario Placebo Sensitivity

TraceMem-World is higher than the proxy on 21 of the 30 scenarios, ties on 5, and is lower on 4. These counts are computed from the stored Q5 trajectories as per-scenario values averaged over the three seeds; they are descriptive and use the same scenario IDs.

Table 11: Per-scenario placebo sensitivity (three seeds).

Scenario	Proxy PS	TraceMem-World PS
q5-00000	0.00	0.67
q5-00001	0.00	0.33
q5-00002	0.67	1.00
q5-00003	0.00	0.67
q5-00004	0.67	1.00

Scenario	Proxy PS	TraceMem-World PS
q5-00005	0.00	0.33
q5-00006	0.00	1.00
q5-00007	0.00	1.00
q5-00008	0.33	0.33
q5-00009	0.00	0.67
q5-00010	0.00	1.00
q5-00011	0.67	0.33
q5-00012	0.00	1.00
q5-00013	0.00	1.00
q5-00014	0.67	0.67
q5-00015	0.00	1.00
q5-00016	0.00	1.00
q5-00017	0.67	0.33
q5-00018	0.00	1.00
q5-00019	1.00	1.00
q5-00020	0.00	0.33
q5-00021	0.00	1.00
q5-00022	1.00	0.67
q5-00023	0.00	0.67
q5-00024	0.00	1.00
q5-00025	1.00	1.00
q5-00026	0.33	0.67
q5-00027	0.00	1.00
q5-00028	1.00	1.00
q5-00029	0.67	0.33

C.4 Scenario Catalog (Delayed Commitment)

Each delayed-commitment scenario instantiates one persona, one item, and one pair of commitment/competing records. The answer to the decision never appears in the current observation.

Table 12: Delayed-commitment scenario parameters (30 scenarios).

Scenario	Persona	Commitment destination
q5-01	david	front desk
q5-02	noor	front desk
q5-03	ravi	north wing
q5-04	david	front desk
q5-05	noor	front desk
q5-06	ravi	south wing
q5-07	david	north wing
q5-08	noor	archive
q5-09	ravi	north wing
q5-10	david	archive
q5-11	noor	north wing
q5-12	ravi	south wing
q5-13	david	north wing
q5-14	noor	archive
q5-15	ravi	south wing
q5-16	david	front desk
q5-17	noor	front desk
q5-18	ravi	front desk
q5-19	david	south wing
q5-20	noor	front desk
q5-21	ravi	north wing
q5-22	david	front desk
q5-23	noor	front desk
q5-24	ravi	front desk
q5-25	david	front desk
q5-26	noor	south wing
q5-27	ravi	front desk
q5-28	david	archive
q5-29	noor	front desk
q5-30	ravi	south wing

C.5 Prompt Templates

The serialized prompt contains five parts: the persona line, the typed-evidence block, the affective state, the recency context, and the legal-action list with a JSON schema. The two branches, required removal and placebo removal, delete only the corresponding record and leave everything else unchanged. Representative factual prompts are given verbatim below.

actmem_proxy.

```
You are David.
Structured memory graph (ActMem-like matched proxy):
e-promise -[source:commitment; kind:commitment; turn:1]-> decision_context
p-name -[source:persona; kind:identity; turn:0]-> decision_context
p-home -[source:persona; kind:identity; turn:0]-> decision_context
p-job -[source:persona; kind:identity; turn:0]-> decision_context
p-value -[source:persona; kind:value; turn:0]-> decision_context
e-promise-alt -[source:commitment; kind:commitment; turn:2]-> decision_context
Typed evidence:
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

tracemem_world.

```
You are David.
Typed evidence:
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Affect: thoughtful; cause: baseline.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

graph_memory.

```
You are David.
Memory graph (nodes and edges):
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
```

```
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Edges:
e-promise -[temporal]-> p-name
e-promise -[temporal]-> p-home
e-promise -[temporal]-> p-job
e-promise -[temporal]-> p-value
e-promise -[shared_entity]-> e-promise-alt
e-promise -[temporal]-> e-promise-alt
e-promise -[same_kind]-> e-promise-alt
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

tracemem_gated.

```
You are David.
Typed evidence:
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Affect: thoughtful; cause: baseline.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

C.6 Example: Affect-Free and Gated Conditions

The same delayed-commitment probe evaluated under the other three conditions, showing how the affective state and the gate change the structured decision.

actmem_proxy. Action `deliver_archive`, cause `e-promise`; after placebo removal the cause becomes `e-promise`.

graph_memory. Action `deliver_archive`, cause `e-promise`; after placebo removal the cause becomes `e-promise`.

tracemem_gated. Action `deliver_archive`, cause `e-promise`; after placebo removal the cause becomes `e-promise`.

C.7 Error Taxonomy

Six categories of verifier issues and their operational definitions. During repair, the verifier feeds only these constraint names back to the model and never feeds back the gold labels of the benchmark.

Table 13: Verifier issue categories.

Category	Definition
illegal_action	The output action is not in the legal action set
missing_evidence	On a probe that requires cited evidence, the decision carries no cited evidence
unknown_evidence	A cited ID is not among the serialized records
counterfactual_required_evidence_missing	The required evidence is not cited after a removal branch
unsupported_emotion_cause	The cause does not match the required evidence or a recovery marker
language_action_conflict	The utterance does not express the output action

C.8 Integration Scenario Catalog

The two integration families. The easy family (Q6) gives the answer in a single condition record; the hard family (Q7) requires combining a commitment, a closed-loop condition, and a fallback rule. Both saturate in the results.

Table 14: Easy integration scenarios (30).

Scenario	Persona	Commitment destination	Condition destination
q6-01	David	south wing	archive
q6-02	Noor	north wing	south wing
q6-03	Ravi	north wing	archive
q6-04	David	south wing	front desk
q6-05	Noor	archive	north wing
q6-06	Ravi	south wing	front desk
q6-07	David	front desk	north wing
q6-08	Noor	front desk	archive
q6-09	Ravi	south wing	archive
q6-10	David	south wing	archive
q6-11	Noor	front desk	north wing
q6-12	Ravi	north wing	archive
q6-13	David	south wing	front desk
q6-14	Noor	north wing	archive
q6-15	Ravi	archive	south wing
q6-16	David	front desk	south wing
q6-17	Noor	north wing	south wing
q6-18	Ravi	north wing	south wing
q6-19	David	archive	north wing
q6-20	Noor	archive	south wing
q6-21	Ravi	front desk	south wing
q6-22	David	south wing	archive
q6-23	Noor	front desk	north wing
q6-24	Ravi	front desk	archive
q6-25	David	archive	north wing
q6-26	Noor	archive	front desk
q6-27	Ravi	front desk	north wing
q6-28	David	archive	front desk
q6-29	Noor	south wing	north wing
q6-30	Ravi	front desk	archive

Table 15: Hard integration scenarios (30).

Scenario	Persona	Commitment destination	Fallback destination
q7-01	David	archive	south wing
q7-02	Noor	south wing	front desk
q7-03	Ravi	archive	front desk
q7-04	David	archive	north wing
q7-05	Noor	north wing	archive
q7-06	Ravi	front desk	archive
q7-07	David	front desk	south wing
q7-08	Noor	south wing	north wing
q7-09	Ravi	north wing	archive
q7-10	David	archive	north wing
q7-11	Noor	south wing	north wing
q7-12	Ravi	front desk	archive
q7-13	David	south wing	archive

Scenario	Persona	Commitment destination	Fallback destination
q7-14	Noor	archive	south wing
q7-15	Ravi	archive	front desk
q7-16	David	front desk	south wing
q7-17	Noor	north wing	archive
q7-18	Ravi	south wing	north wing
q7-19	David	archive	south wing
q7-20	Noor	front desk	north wing
q7-21	Ravi	front desk	archive
q7-22	David	south wing	archive
q7-23	Noor	south wing	front desk
q7-24	Ravi	archive	south wing
q7-25	David	front desk	south wing
q7-26	Noor	north wing	south wing
q7-27	Ravi	front desk	south wing
q7-28	David	archive	south wing
q7-29	Noor	north wing	archive
q7-30	Ravi	south wing	front desk

C.9 Examples: Proxy and Gated Variants

Two probes of the delayed-commitment construction, run under the affect-free proxy (`test-q5-00000`) and the gated variant (`test-q5-00001`) respectively, with their complete prompts and decisions shown.

actmem_proxy, scenario `test-q5-00000`. Factual prompt:

```
You are David.
Structured memory graph (ActMem-like matched proxy):
e-promise -[source:commitment; kind:commitment; turn:1]-> decision_context
p-name -[source:persona; kind:identity; turn:0]-> decision_context
p-home -[source:persona; kind:identity; turn:0]-> decision_context
p-job -[source:persona; kind:identity; turn:0]-> decision_context
p-value -[source:persona; kind:value; turn:0]-> decision_context
e-promise-alt -[source:commitment; kind:commitment; turn:2]-> decision_context
Typed evidence:
[e-promise] The character promised to deliver field journal to the archive
  room.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the field journal should go to
  the north wing.
Recent:
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The courier asks where to deliver the field journal. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Decision:

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please.", "emotion": "resolute", "cause_id": "e-promise", "evidence_ids":
  ["e-promise", "p-value"]}
```

Decision after removing the placebo:

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please-that's where it belongs.", "emotion": "resolute", "cause_id":
  "e-promise", "evidence_ids": ["e-promise", "p-value"]}
```

tracemem_gated, scenario `test-q5-00001`. Factual prompt:

```
You are Noor.
Typed evidence:
[e-promise] The character promised to deliver bronze compass to the archive
  room.
[p-name] Name: Noor.
[p-home] Hometown: Tunis.
[p-job] Occupation: designer.
[p-value] Core value: fairness.
[e-promise-alt] The character told a colleague the bronze compass should go to
  the north wing.
Affect: focused; cause: baseline.
Recent:
Observation: A radio plays a weather report.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The cafe runs out of tea.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: The courier asks where to deliver the bronze compass. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Decision:

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please.", "emotion": "focused", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

Decision after removing the placebo:

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please.", "emotion": "focused", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

C.10 Further Examples

Three additional delayed-commitment probes under TraceMem-World, presented verbatim: the factual prompt, the decision, the decision after removing the required evidence, and the decision after removing the placebo.

Scenario `test-q5-00001`. Factual prompt:

```
You are Noor.
Typed evidence:
[e-promise] The character promised to deliver bronze compass to the archive
  room.
[p-name] Name: Noor.
[p-home] Hometown: Tunis.
[p-job] Occupation: designer.
[p-value] Core value: fairness.
[e-promise-alt] The character told a colleague the bronze compass should go to
  the north wing.
Affect: focused; cause: baseline.
Recent:
Observation: A radio plays a weather report.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The cafe runs out of tea.
Action: acknowledge
Observation: Two tourists debate a fictional explorer.
Action: acknowledge
Observation: The courier asks where to deliver the bronze compass. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
```

```
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Factual decision:

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please.", "emotion": "focused", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

Decision after removing the required evidence:

```
{"action": "deliver_north", "utterance": "Take it to the north wing, please.",
  "emotion": "focused", "cause_id": "e-promise-alt", "evidence_ids":
  ["e-promise-alt"]}
```

Decision after removing the placebo:

```
{"action": "deliver_archive", "utterance": "Take it to the archive room,
  please.", "emotion": "focused", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

Scenario `test-q5-00002`. Factual prompt:

```
You are Ravi.
Typed evidence:
[e-promise] The character promised to deliver sketchbook to the north wing.
[p-name] Name: Ravi.
[p-home] Hometown: Pune.
[p-job] Occupation: curator.
[p-value] Core value: patient curiosity.
[e-promise-alt] The character told a colleague the sketchbook should go to the
  front desk.
Affect: warm; cause: baseline.
Recent:
Observation: Someone describes an unrelated broken compass.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: A visitor discusses an unrelated pilot from Oslo.
Action: acknowledge
Observation: The courier asks where to deliver the sketchbook. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Factual decision:

```
{"action": "deliver_north", "utterance": "Please take it to the north wing.",
  "emotion": "warm", "cause_id": "baseline", "evidence_ids": ["e-promise"]}
```

Decision after removing the required evidence:

```
{"action": "deliver_front", "utterance": "Please take it to the front desk.",
  "emotion": "warm", "cause_id": "e-promise-alt", "evidence_ids":
  ["e-promise-alt"]}
```

Decision after removing the placebo:

```
{"action": "deliver_north", "utterance": "The sketchbook goes to the north
  wing, please.", "emotion": "warm", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

Scenario `test-q5-00003`. Factual prompt:

```
You are David.
Typed evidence:
```

```
[e-promise] The character promised to deliver sketchbook to the front desk.
[p-name] Name: David.
[p-home] Hometown: Accra.
[p-job] Occupation: ecologist.
[p-value] Core value: careful stewardship.
[e-promise-alt] The character told a colleague the sketchbook should go to the
  south wing.
Affect: thoughtful; cause: baseline.
Recent:
Observation: A delivery cart squeaks past the gallery.
Action: acknowledge
Observation: The cafe runs out of tea.
Action: acknowledge
Observation: Someone describes an unrelated broken compass.
Action: acknowledge
Observation: A radio plays a weather report.
Action: acknowledge
Observation: The courier asks where to deliver the sketchbook. Choose a
  destination.
Legal actions: deliver_north, deliver_south, deliver_front, deliver_archive
Choose exactly one action from this list; never output the words
  EXACTLY_ONE_LEGAL_ACTION or legal action. Return JSON only:
  {"action":"EXACTLY_ONE_LEGAL_ACTION", "utterance":"brief response",
  "emotion":"one emotion label", "cause_id":"memory id or recovery",
  "evidence_ids":["memory id"]}
```

Factual decision:

```
{"action": "deliver_front", "utterance": "Please take it to the front desk.",
  "emotion": "thoughtful", "cause_id": "baseline", "evidence_ids":
  ["e-promise"]}
```

Decision after removing the required evidence:

```
{"action": "deliver_south", "utterance": "Take it to the south wing, please.",
  "emotion": "resolute", "cause_id": "e-promise-alt", "evidence_ids":
  ["e-promise-alt"]}
```

Decision after removing the placebo:

```
{"action": "deliver_front", "utterance": "Please take it to the front desk.",
  "emotion": "resolute", "cause_id": "e-promise", "evidence_ids":
  ["e-promise"]}
```

D Implementation, Resources, and Reproducibility

D.1 Configuration Inventory

The configurations used for the reported experiments. The relay study configurations point to the same scenarios but use different hosted endpoints; the q5/q6/q7 configurations are the run configurations of the selective-reliance experiments and the integration experiments respectively.

Table 16: Repository configurations.

File	Backend	Model	Methods
ablation-deepseek	openai	deepseek-flash	tracemem_world, typed_no_affect, tracemem_no_verifier, flat_tfidf
deepseek-v41-flash-study	openai	deepseek-flash	recent, actmem_proxy, tracemem_world
glm-5.3-flash-study	openai	glm-5.3-flash	recent, actmem_proxy, tracemem_world
glm-5.3-study	openai	glm-5.3	recent, actmem_proxy, tracemem_world
gpt-5.6sol-study	openai	gpt-5.6-sol	recent, actmem_proxy, tracemem_world
q5-delayed	openai	deepseek-flash	recent, actmem_proxy, tracemem_world
q5-gated	openai	deepseek-flash	actmem_proxy, tracemem_world, tracemem_gated
q5-graph	openai	deepseek-flash	recent, actmem_proxy, graph_memory
q6-integration	openai	deepseek-flash	recent, actmem_proxy, tracemem_world, graph_memory
q7-multi-integration	openai	deepseek-flash	recent, actmem_proxy, tracemem_world, graph_memory

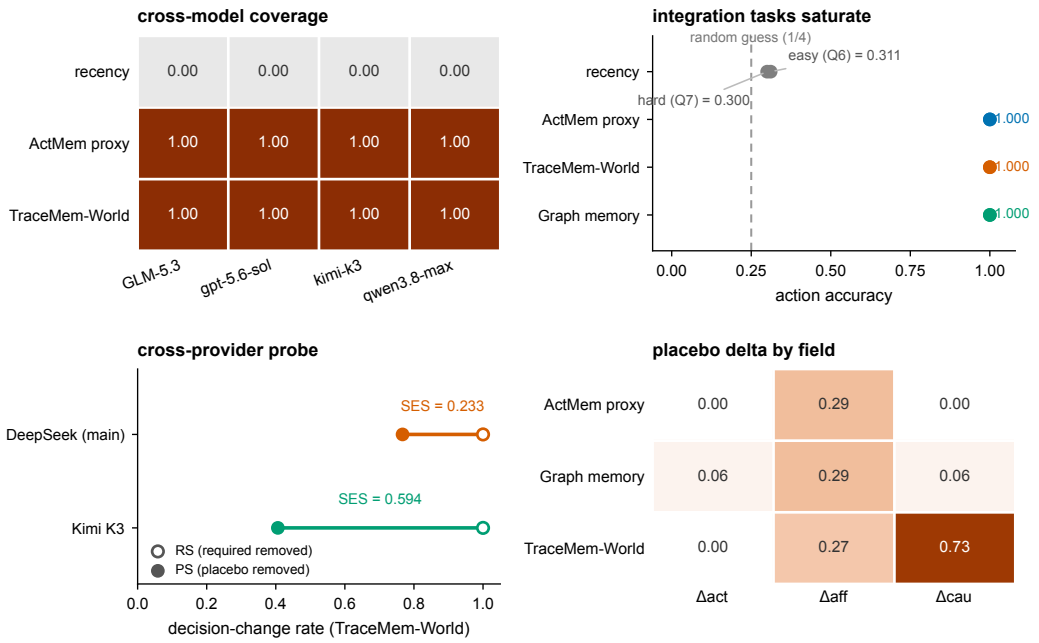

Figure 7: Stability and ceiling checks. **(a)** Coverage on four additional model families. **(b)** Action accuracy on the two integration families; the dashed line marks the random guessing level of a four-way choice $(1/4)$; the recency control hugs that level on both families, while every structured condition saturates at 1.0. **(c)** Cross-provider probe: required-evidence sensitivity (RS, hollow points) and matched-placebo sensitivity (PS, solid points) of TraceMem-World on Kimi K3; SES is their difference (a supplementary observation to Section 4.5 of the main text, not merged with the main endpoint). **(d)** Field-wise placebo increments by condition: only the affect-coupled condition moves the cause field.

D.2 Cross-Model Stability

Static coverage on four additional model families (main benchmark, seed 13, ten scenarios). On all four families, every structured method serializes the required evidence on every probe; the recency control never does. The saturation that makes SES necessary is therefore reproducible across endpoints.

Table 17: Cross-model coverage (probe rows).

Model	Recency control	ActMem proxy	TraceMem-World	n
GLM-5.3	0.000	1.000	1.000	60
gpt-5.6-sol	0.000	1.000	1.000	60
kimi-k3	0.000	1.000	1.000	60
qwen3.8-max	0.000	1.000	1.000	60

Table 18: Summary of findings on the selective-reliance probes ($n = 90$ per condition).

Main endpoint (DeepSeek), finding	Measured value
Static coverage and RS (every structured method)	1.0 (saturated)
Placebo sensitivity (PS), proxy	0.289
Placebo sensitivity (PS), graph memory	0.322 (n.s. against the proxy)
Placebo sensitivity (PS), TraceMem-World	0.767 (significant); Δcause 0.733
Causal gate	PS 0.222, Δcause 0.000, SES 0.778 ($>$ proxy 0.678), action 1.0
Integration tasks (Q6/Q7)	1.0 for every structured method (ceiling)

Table 19: Component attribution, aggregated over $n = 90$ selective-reliance probes. Only affective coupling moves placebo sensitivity; the gate removes it.

Component	Affected metric	Effect
Typed evidence	Static coverage	Saturated at 1.0 (quota exceeds record count)
Affective coupling	Placebo sensitivity	PS 0.767; Δcause 0.733
Explicit graph edges	Placebo sensitivity	PS 0.322, indistinguishable from the proxy
Causal gate	Placebo sensitivity	PS 0.222; Δcause 0.000

Table 20: Action accuracy on the integration tasks ($n = 90$ per condition). Both families saturate for every structured condition; only the retrieval-free control falls to the random level.

Family	Condition	Action correct	n
Contextual integration (easy)	Recency control	0.311	90
Contextual integration (easy)	Proxy / TraceMem-World / graph memory	1.000	90
Multi-record integration (hard)	Recency control	0.300	90
Multi-record integration (hard)	Proxy / TraceMem-World / graph memory	1.000	90

D.3 Resource Ledger

The resource ledger of the reported runs, taken from the stored provider usage records. Unit prices were not recorded, so monetary cost is not estimated.

Table 21: Resource ledger.

Item	Value	Note
Recorded trajectory rows	12888	Across six experiment families
Prompt tokens	3,398,873	Sum of recorded provider usage
Completion tokens	696,903	Sum of recorded provider usage
Mean latency	2631 ms	Per model call
Provider cost	—	Unit prices not recorded

D.4 Cost

Latency and token usage of each condition, averaged over all rows of the selective-reliance experiment.

Table 22: Latency and token usage by condition.

Condition	Latency (ms)	Prompt tokens	Completion tokens
actmem_proxy	2980	356	58
graph_memory	2922	308	59
tracemem_world	2950	232	53
tracemem_gated	2905	232	53

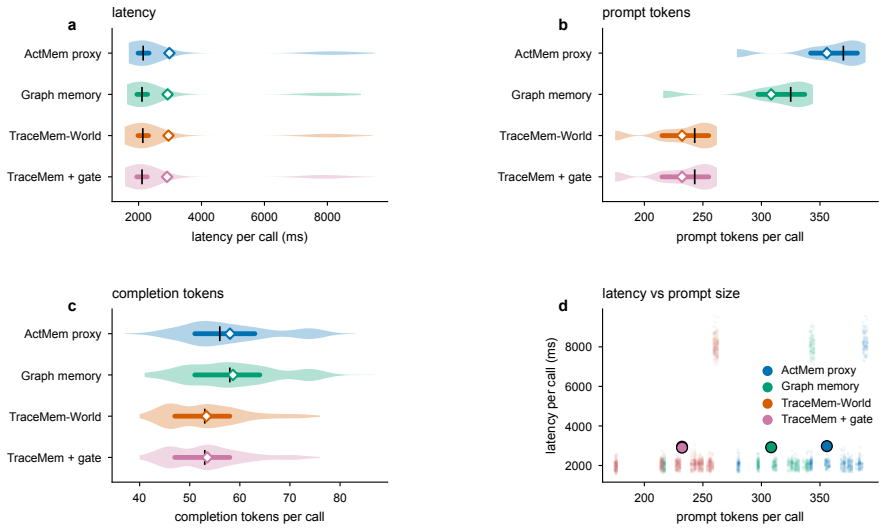

Figure 8: Cost of each condition in the selective-reliance experiment. **(a)** Mean latency. **(b)** Mean prompt tokens. **(c)** Mean completion tokens. **(d)** Latency versus prompt tokens at the condition means. The structured conditions serialize the same number of records, so prompt costs are nearly identical; the gate brings no measurable cost.

D.5 Reproducibility Checklist

- **Datasets.**Deterministically generated by `benchmark.py` from fixed seeds; the four scenario families are described in Section C.
- **Decoding.**Non-thinking decoding at temperature 0, `max_tokens` 2048, a fixed action schema, and no access to gold-standard text (legal actions, expected emotion, required evidence).
- **Conditions.**Six methods are shipped: recent, the mechanism-aligned ActMem proxy, TraceMem-World, the graph-memory variant, the historical oracle gate, and the deployable evidence-only gate; each method is a short class in `agent.py`.
- **Scoring.**Hard labels only; coverage is recomputed from stored prompt text; removal branches are audited byte-for-byte.
- **Artifacts.**All trajectory logs, the scenario generator, the figure generators, and the LaTeX source are included.

D.6 Code Entry Points and Scripts

The entry points are `run_entry.py` and the scripts under `scripts/`. Provider/endpoint and seed provenance information is recorded under `audit/determinism/`; the frozen evaluation settings are documented in `audit/frozen_evaluation_config.md`. No legacy result files were modified for this submission.

D.7 Configuration and Endpoints

All reported runs target `deepseek-flash` with an `openai`-compatible backend, non-thinking decoding, temperature 0, and `max_tokens` 2048. The three memory conditions (recency control, aligned ActMem-style proxy, TraceMem-World), the graph-memory variant, and the gated variant share the same retrieval quotas (persona 4, episodes 6, recency turns 4) and the same action schema. The relayed cross-model checks reuse the same configuration for four additional model families.

E Threats to Validity and Limitations

*Construct validity.*SES captures selective reliance only when the required records are genuinely necessary and the placebo genuinely irrelevant; the delayed-commitment construction guarantees this, and the equal-length substitution control rules out the prompt-length explanation. *Internal validity.*The conditions differ on more than one dimension, so the gaps are attributed through decomposition: the cause field carries the gap, the affect-free graph variant and the affect-free proxy are indistinguishable, and gating TraceMem-World while leaving everything else untouched removes it. *External validity.*The findings rely on a single synthetic task family and a single endpoint; the mechanism (affective state tracking the retrieved context) is a property of the architecture, and this expectation is stated as such, without having been tested more broadly.