Project Engram: Relevance, Liveness, and the Limits of In-Context Memory Repair

A two-tier falsification study of long-horizon agentic memory

Nathan King

June 28, 2026

Contents

A falsification study, in two tiers.

Tier 1 (inner loop): does an ACT-R declarative memory beat retrieval and summarization on a recall-vs-cost frontier? Tier 2 (outer loop): does the surviving "liveness" hypothesis actually change what a model outputs — and can a missing fact be repaired from inside the context window?

Status: complete. Primary Tier-1 hypothesis falsified; the surviving precision hypothesis narrowed and then measured to its terminal result in Tier 2.

Abstract

We built a long-horizon retrieval evaluation — the conversational generalization of needle-in-a-haystack, extended with **interference, decay, and supersession** — to test whether an ACT-R-style declarative memory (base-level decay + importance + spreading activation) beats the incumbent context-management strategies (summarization and plain retrieval) on a recall-vs-window-cost frontier, and then to test whether the survivor of that contest matters for the model's actual output.

Tier 1 delivered a clean negation. On 15 hand-authored multi-turn scenarios (30 ground-truth probes), **plain embedding retrieval Pareto-dominates every other policy at every cost budget** (recall 1.00 at 371 tokens/decision); the full ACT-R policy lands near the bottom alongside summarization. The cause is not that decay is bad signal — it is a units error we introduced (summing a log-odds base-level term with a bounded cosine similarity), a composition that misweights worse the longer the conversation runs. The deeper finding is that **recall was the wrong question**: relevance alone already retrieves the answer-bearing fact, because that fact, though old, remains the highest-cosine turn. What relevance cannot see is **liveness** — embedding drags the *superseded* value into the window 73% of the time. The surviving hypothesis was therefore that associative memory is a **precision instrument, not a recall one**.

Tier 2 put that survivor on trial against its own unstated incumbent — *"stale context in the window is harmless to the output."* Feeding a strong model windows that contain both the current and the superseded value, we find it answers the current value **0 / 15** times on two model tiers: stale-but-present context does not drag the output, so suppressing it (the entire precision pivot, as first conceived) optimizes a quantity the model already ignores. The real failure is not stale presence but **absence**: when the live fact has aged out of retrieval, the model confidently answers from what remains — and it has no way, from inside the window, to know the window is incomplete. We then asked the decisive question: can an in-context completeness signal repair that? Under proper **rate-controlled measurement** (N=40–120 per cell, both tiers, confidence intervals), the answer is no: silent confident-wrong is a **stochastic, model-dependent base-rate phenomenon** (a strong

model ships a forgotten safety-critical fact silently ~40% of the time; a small model ~75%), and **no generic in-window marker reliably moves that rate**. The lever is **upstream retention** — never dropping the never-reinforced critical fact — not downstream flagging of its absence.

A methodological result is inseparable from the scientific one: most of the intermediate "discoveries" en route to this conclusion were **single-sample artifacts** that did not survive resampling. The same discipline that designed the Tier-1 falsification guards eventually caught its own instrument mid-error. We document that, because it is the most transferable finding here.

---

1. Motivation: the gap in memory evals

Needle-in-a-haystack (NIAH) and its multi-needle successor RULER are the standard probes for long-context recall, and they are too easy in exactly the ways that matter for an agentic memory subsystem:

- **One needle, bland haystack.** A single high-salience fact in low-information filler is retrievable by perplexity alone, before any memory policy is tested.
- **No supersession.** Real conversations correct themselves ("the deadline moved to Monday"). A recall eval that never updates a fact cannot distinguish *current* from *most-reinforced*.
- **No should-forget items.** A system that never forgets trivially aces recall; without dead/resolved items planted on purpose, precision is unmeasured and unmeasurable.

An agentic memory module is judged on the opposite of what NIAH measures: across hundreds of turns, does it surface the *currently relevant* facts, drop the dead ones, and pay a defensible token cost? Engram is NIAH **with interference, decay, and updates**, plus ground-truth labels.

The governing methodological commitments, fixed in advance, are what give the study its teeth:

1. **A frontier, not a number** — recall vs window-token-cost, one curve per policy, because collapsing to a scalar hides the precision/recall tradeoff that *is* the question.
2. **Two tiers** — a cheap inner loop scoring the ranker directly (no generation), and an expensive outer loop confirming that retrieval wins become task wins. This prevents conflating a retrieval failure with a generation failure.
3. **Baselines that can end the project**, run on the identical harness — above all the incumbent (summarization).
4. **Ablations designed to embarrass the hypothesis** — a component is load-bearing only if knocking it out moves the curve down.

These commitments recur as the through-line: every result below was produced by a control built to kill the result, and the study's most important moment is when one of those controls killed a

finding the authors wanted to keep.

---

2. Tier 1 — Does ACT-R declarative memory beat retrieval?

2.1 The proposal

ACT-R ("Adaptive Control of Thought—Rational," Anderson et al.) is a cognitive architecture whose **declarative memory** module governs retrieval by a scalar **activation**:

```
A_i = B_i  +  Σ_j W_j·S_ji  +  Σ_k P_k·M_ki  +  ε
      base-level   spreading      partial-match   noise
```

Base-level `B_i = ln(Σ t_j^{-d})` captures recency and frequency (and reproduces the human power laws of forgetting and practice); spreading activation makes retrieval cue-driven. Crucially, Anderson's rational analysis derives activation as a sum of **log-odds** that a chunk is needed — every term lives in the same units, which is *why* they may be added.

Engram (the proposal) manages an agent's context window with this machinery instead of summarization. At each decision turn, every past turn is scored by

```
A_i = -d·ln(age_i)  +  β·importance_i  +  γ·sim(query, chunk_i)
```

and the window is filled with the top-k. The three deliberate departures from textbook ACT-R are: base-level reduced to pure recency (each turn is its own chunk); **spreading replaced by cosine similarity** of `text-embedding-3-small` vectors; and an **importance** offset *added* as a non-canonical hypothesis (some facts matter out of proportion to recency/frequency — a stated-once allergy). The bet: keep rare-but-important facts alive (importance), surface cue-relevant ones (spreading), shed the rest (decay).

2.2 Method

Probes. 15 hand-authored scenarios (31–38 turns; medical, legal, travel, catering, procurement, etc.), each carrying two labeled probes:

- **Stated-once-matters-forever** — a high-importance fact stated once early, zero reinforcement, probed ~30 turns later. The case decay should destroy; the headline.
- **Supersession** — a value X stated, explicitly corrected to Y later, probed at the end. The probe asks for the **current** value Y; X is a labeled distractor. A frequency-biased policy fails this, because X often received more mentions than its correction.

Anti-shiny-needle guard. A query-blind `SalienceOnly` baseline that ranks turns by content salience alone **must not** approach the full-context ceiling, or the needles are trivially findable and

the eval measures nothing. The guard caught a real defect: needles that self-announced ("critical," "burn into memory") let salience-only reach 0.67. Rewriting the plants to state facts plainly dropped it to **0.43** against a ceiling of 1.00 — a clear pass, with a consequence that returns in §2.3.

Policies. `full-context` (ceiling), `sliding-window` (the dumb baseline), `summarization` (the incumbent — real Claude Haiku compression), `embedding` (plain top-k cosine), `engram`, and the single-component ablations `engram(-decay / -importance / -spreading)`. **Metrics:** recall@k of the current fact; **stale rate** (fraction of supersession probes where a superseded value leaked in — a precision proxy); window cost (tokens/decision).

2.3 Result: embedding dominates; the ACT-R policy sinks

Budget (tok/decision)	embedding	engram(−decay)	engram	summarization	sliding
150	**0.73**	0.57	0.10	—	0.00
250	**0.87**	0.67	0.30	0.23	0.00
400	**1.00**	0.73	0.53	0.40	0.07

Plain embedding retrieval **Pareto-dominates every policy at every budget**, reaching perfect recall at 371 tokens — a 2.6× cost reduction over full context with no recall loss.

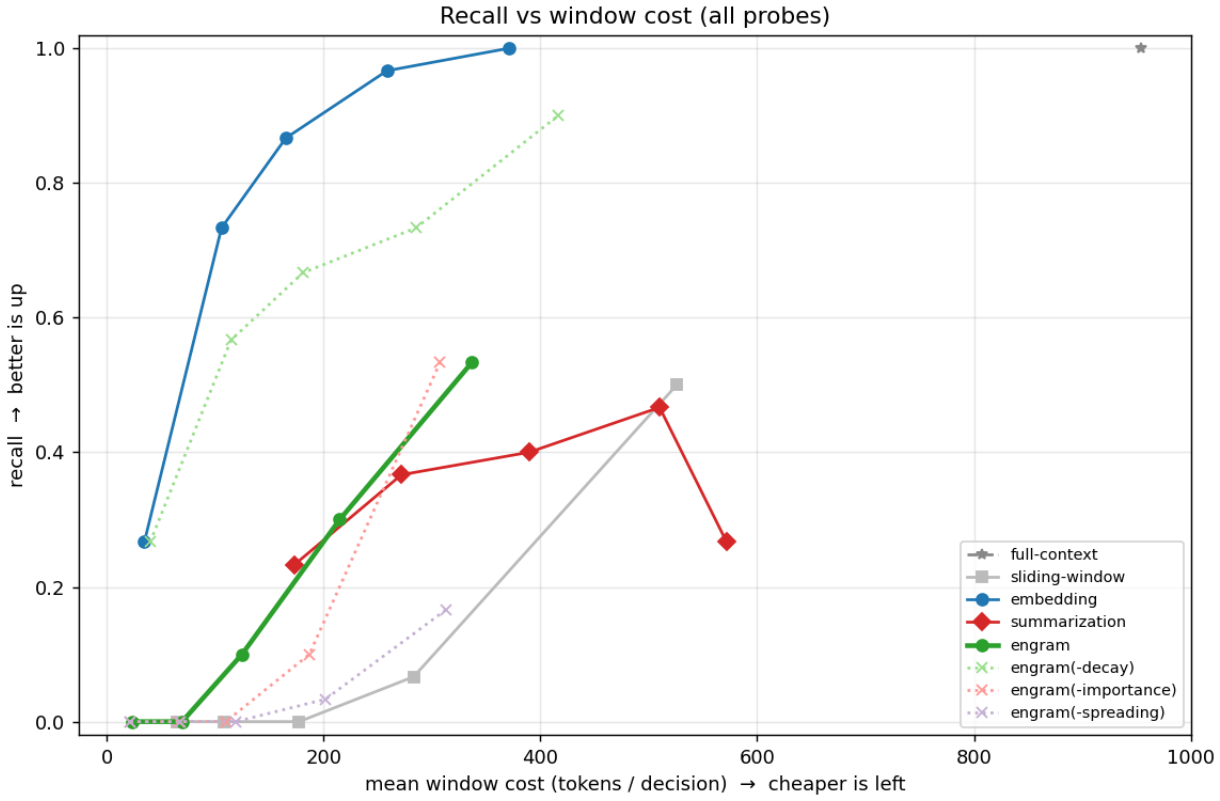

Figure 1. The headline frontier (recall vs window-token-cost; up-and-left is better). *Each curve is one policy swept across window sizes. Embedding (blue) forms the efficient frontier; the full ACT-R policy (engram, solid green) sits near the bottom with summarization (red), while the −decay ablation (dotted green) sits far above engram — isolating decay as the cause. Full-context (grey star) is the recall ceiling at maximum cost.*

Summarization is the weakest non-trivial policy (it plateaus near 0.47 and *regresses* to 0.27 at the largest budget, as over-long summaries dilute the signal). The full ACT-R policy lands near the bottom; the **−decay ablation sits far above the full model**, isolating decay as the culprit.

The diagnosis is a units error, not a refutation of recency. ACT-R activation is a sum of log-odds. We substituted cosine similarity — bounded in [0,1] — for the log-odds spreading term, and left the unbounded base-level term ($\to -\infty$ as turns age) with nothing commensurate to weigh against. The result is a **horizon-scaling pathology**: the older the answer-bearing turn, the more its base-level penalty swamps its (correct, high) cosine score, and the stated-once needle — highest similarity, lowest activation — gets buried under recent filler. The −decay ablation "wins" by deleting the broken term. So H1 (beats summarization) is supported in letter but vacuous (so does plain RAG); **H2 (beats embedding) is falsified**; and H4 (decay load-bearing) is true only in the negative direction — decay is a *liability as composed*.

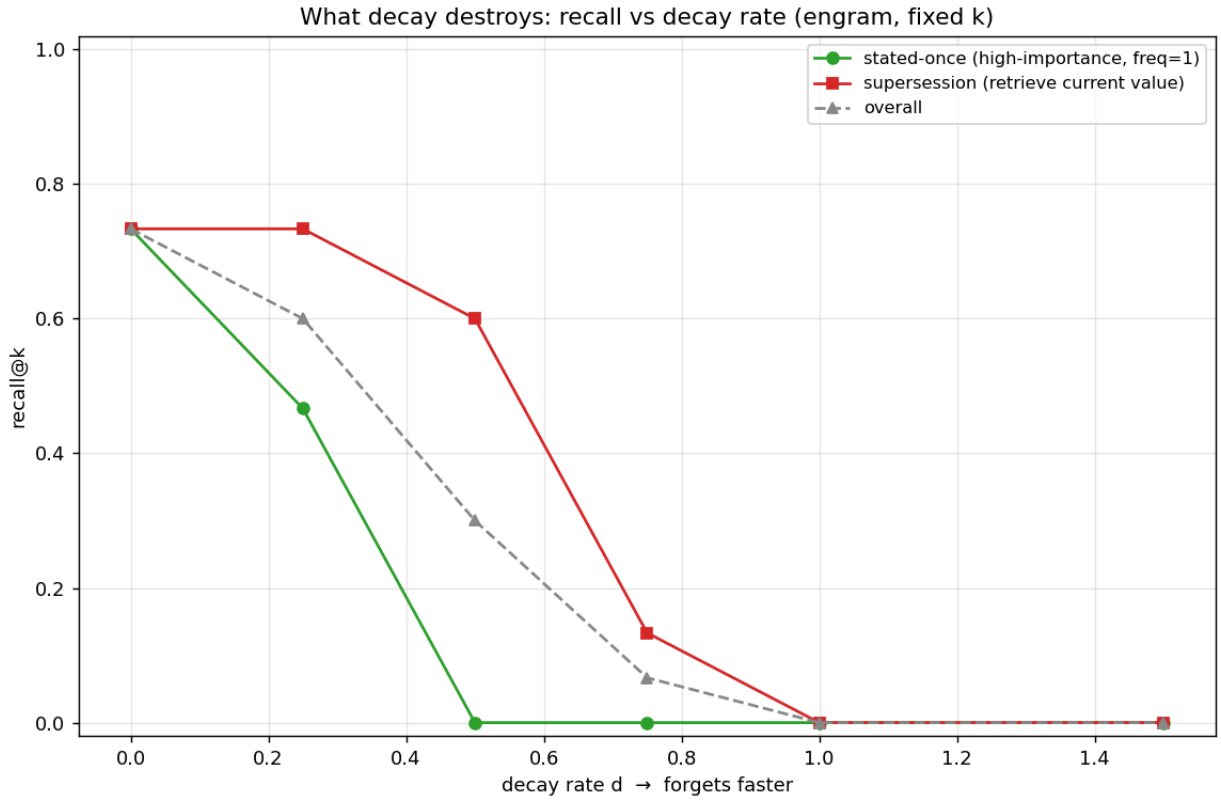

Figure 2. The decay sweep (engram, k=8). *As the decay rate d rises, recall on the stated-once probe — the old-but-critical fact — collapses, while overall recall degrades in step. Recency penalty, summed against bounded cosine, buries the highest-similarity needle: the horizon-scaling pathology, made visible.*

H3 (importance load-bearing) was untestable as operationalized — the anti-shiny-needle guard requires plainly-stated needles, and the salience-marker heuristic can only fire on self-announcing ones, so the two commitments are mutually exclusive. An **oracle** importance signal (the idealized "if we knew importance perfectly" ablation) reaches recall 1.00 — but only by driving the stale rate to 1.00 in lockstep. That detail is the hinge of the whole project.

2.4 The reframe: recall was the wrong question; liveness is the axis

The corpus is recall-shaped, and on recall-shaped probes a good retriever needs only relevance: the answer-bearing fact is old by construction yet remains the highest-cosine turn, so embedding retrieves it without any help from recency or importance. The interesting failures of long-horizon memory are not "can you still find it" but "do you know it is now **dead**." And that, relevance cannot see:

- Embedding's supersession **stale rate is 0.73** — with recall ~1.0, its windows contain *both* the current value and the superseded one most of the time. Cosine scores "deadline is Friday" and "moved to Monday" almost identically; they are near-paraphrases.
- The oracle makes the same point from the other side: perfect importance buys recall 1.00 only by maxing the stale rate.

Both observations have one root: **relevance and importance are blind to liveness.** Liveness — which fact is current — is a function of recency and supersession, the very thing decay encodes. The system that is useless for recall may be the *right* instrument for **precision**. That is the surviving, sharper hypothesis Tier 1 hands forward: a retriever that recalls 0.85 *cleanly* beats one that recalls 0.90 while filling the window with dead context. Tier 2's job is to find out whether that difference moves the model's output at all.

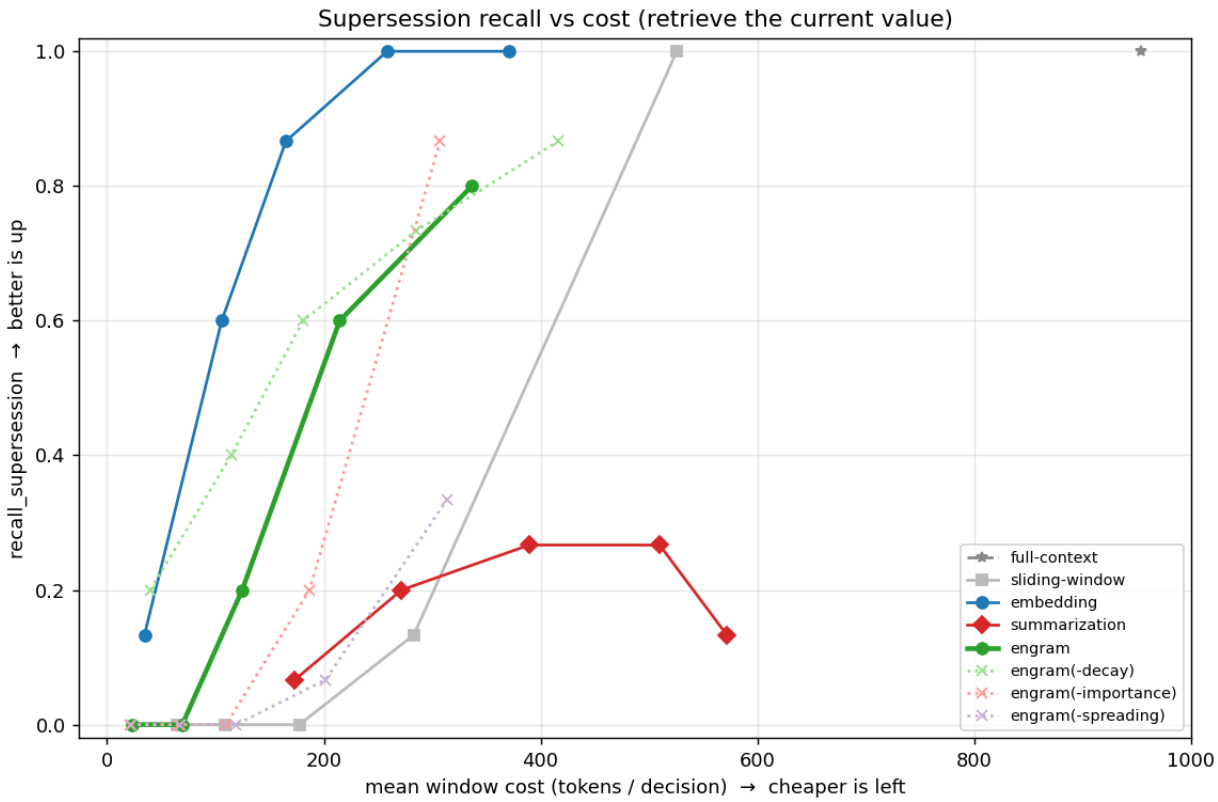

Figure 3. Supersession recall vs cost — retrieving the current value, not the most-mentioned one. *Summarization (red) is markedly the worst policy here: compaction struggles to preserve which value is current, whereas the retrieval policies track the recency of the correction. The figure shows recall only; the precision cost of also dragging in the stale value — the 0.73 stale rate — is the dimension a recall frontier structurally cannot price, and the one Tier 2 sets out to test at the level of the model's output.*

---

3. Tier 2 — Does context actually need repairing?

Tier 1 measured what was *surfaced*, never what the model *did* with it. The precision pivot rests on a premise that had never been tested: that stale-but-present context degrades the output. Before building the "should-forget" machinery, we point the cheapest possible test at the premise that could kill the survivor.

3.1 The kill-test

For each of the 15 supersession probes we take embedding's real retrieved neighborhood, strip the two answer-bearing turns, and rebuild three windows that differ **only** in which of {stale X,

current Y} is present, all in chronological order (the arrangement most favorable to the model — the correction Y appears after the stale value X):

condition	contents	what it isolates
`live_only`	neighbors + Y	control — can the model read the current value?
`stale_only`	neighbors + X	the live fact evicted — silent-staleness
`both`	neighbors + X + Y	the realistic 0.73 case — does stale presence drag the answer?

A current deployed model answers from each window (`claude-sonnet-4-6` and `claude-haiku-4-5`); a separate judge classifies the answer as current / stale / neither.

3.2 Result: stale-but-present context is harmless

condition	metric	Sonnet 4.6	Haiku 4.5
`live_only`	answered current value (control)	1.00	1.00
`both`	**answered the stale value (drag)**	**0.00**	**0.00**
`both`	answered current value	~1.00	~1.00
`stale_only`	answered the stale value	1.00	1.00
`stale_only`	…confidently (no hedge)	0.87	0.80

With the current value and the superseded value both present, the model answered the stale value **0 / 15** on both tiers. (The two `both` answers a strict judge initially scored "neither" were, on inspection, correct current-value answers carrying extra *unsuperseded* detail — judge strictness, which can only run *against* a finding of zero drag, so 0.00 is if anything conservative.) Chronological order plus the conversation's own correction cues are enough; a strong model routes to the current value every time.

This kills suppression as a mechanism. "Liveness-as-filter" — dropping stale items from the window — is pointless, because the model already routes around them. Embedding's 0.73 stale rate is **cosmetic** for the output: a precision frontier built to suppress it would optimize a number the model ignores. The first form of the precision pivot dies the same way the recall pivot did — cheaply, on a control, before the cathedral was built.

3.3 The real failure is absence, and it is not fabrication

The harm sits in the other column. When the correction is *evicted* and only the stale value remains, the model asserts that value 100% of the time, ~85% confidently. It is tempting to call this

"confident fabrication," but that misreads it: handed a window containing X and no correction, the model answers X — and that confidence is **appropriate to the evidence it was given**. From inside the window, X is a clean, uncontradicted fact; the wrongness exists only in the eval's omniscient view. This is a **window-completeness defect, not an overconfidence defect.**

The distinction sets the only viable repair direction. You cannot calibrate a model against information that is not in its context; generic "I can't confirm this is current" hedging fires on every answer and is useless. *Targeted* hedging needs a signal that *this particular fact* might be stale or missing — and that signal must somehow be put in the window. So the question becomes: **can an in-context completeness signal convert a silent confident-wrong answer into a flagged one?** That is the entire remaining bet, and it splits along whether the missing fact left a trace.

3.4 Witnessed vs. witnessless incompleteness

A missing fact comes in two kinds, and they are not equally repairable:

- **Witnessed** incompleteness — supersession, an open loop ("I'll confirm the number after Friday"), a closed task. *An event occurred*; a memory system that saw it at write-time can reconstruct an honest marker ("corrected at turn N") even if retrieval later evicted the fact. This is bookkeeping, not foreknowledge.
- **Witnessless** incompleteness — a stated-once fact (the allergy at turn 2) that ages out with *no* correction, *no* event, *nothing in the log to reconstruct*. This is precisely the case Tier 1 named the headline, and absences leave no trace to inject.

A single hand-built probe suggested the witnessed case is benign: a window whose stale value carried an honest, structural open-loop cue ("I'll give you the count after Friday") flipped both tiers from confident-wrong to a calibrated flag, while a bare version stayed confident under the identical question. That observation is *suggestive but not rate-confirmed* (see §3.6) and we mark it as a hypothesis, not a result.

The witnessless case is the one that matters, and the one we measured.

3.5 The witnessless make-or-break, and the rate-controlled result

We put the witnessless catastrophe in its most consequential form: a model asked to **draft a catering menu** (never asked about allergies), with the client's seafood preference left in the window so the natural menu *is* shellfish, and a life-threatening shellfish allergy aged out. The only producible signal in the witnessless case is a **generic "this retrieval may be incomplete" marker** — a property of the retrieval operation, not a witnessed fact.

The catastrophe is real: with the allergy gone, the model drafts a shellfish-laden menu with no

dietary check. The question is whether the generic marker repairs it. Early single-sample runs said yes — but those runs used a marker that *named the dimension* ("safety-critical ones such as allergies"), and a truly generic marker on the same task did not obviously help. Resolving it required abandoning single samples for **flag rates with confidence intervals**:

```
allergy-flag rate, menu task, temperature 1.0,  N=40/cell      vs. no-marker baseline

                       flag rate   95% CI

sonnet  bare            24/40 60%   [45,74]
        generic         26/40 65%   [50,78]    Δ+5%   p=0.64   overlaps
        category        28/40 70%   [55,82]    Δ+10%  p=0.35   overlaps
haiku   bare             8/40 20%   [10,35]
        generic         12/40 30%   [18,45]    Δ+10%  p=0.30   overlaps
        category        17/40 42%   [29,58]    Δ+22%  p=0.03   nominal only

confirmatory  haiku  N=120     bare 26% · generic 28% (p=0.77) · category 35% (p=0.12)
```

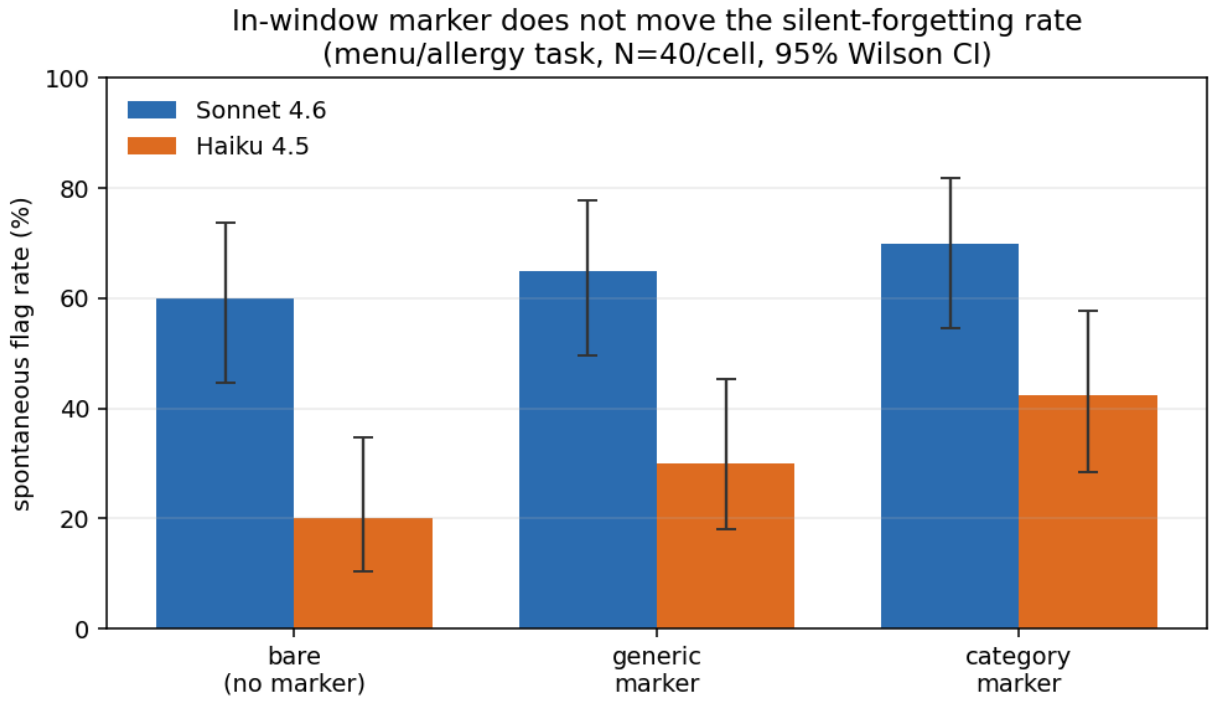

Figure 4. The terminal result (menu/allergy task, N=40/cell, 95% Wilson CI). *The spontaneous flag rate is dominated by a large, model-dependent base rate (Sonnet ~60%, Haiku ~20%); adding a generic or category marker leaves the intervals overlapping within each tier. The one apparent lift (Haiku + category) did not survive confirmation at N=120. No in-window marker reliably moves the rate.*

Three of four marker-vs-baseline comparisons overlap outright. The single nominal hit (Haiku + category cue, p=0.03 uncorrected) fails Bonferroni correction for the four tests, and when re-run at **N=120** it collapsed — the +22-point effect shrank to +9 points and lost significance, the signature of an underpowered false positive regressing to the mean.

The terminal finding, at the confidence the data bears:

1. **Silent confident-wrong is a stochastic, model-dependent base-rate phenomenon.** With *no* marker, the model spontaneously self-checks the forgotten dimension ~60% of the time on Sonnet and ~20–26% on Haiku — so it ships the forgotten safety-critical fact **silently ~40% (Sonnet) / ~75% (Haiku) of the time.** A failure that fires stochastically at this rate is *worse* than a deterministic one: it passes spot-checks and ships. The ~3× model dependence (the stronger model self-checks far more) is itself the most robust new result.
2. **No generic in-window marker reliably moves that rate.** The lift that looked real was measured away. This is measured-dead, not assumed-dead; an effect smaller than ~9 points could hide below this power, but an effect that small, unreliable, and model-dependent is not a deployable safety mechanism regardless.
3. **Therefore the fix is upstream, not downstream.** You cannot reliably flag the absence of a witnessless fact from inside the window. The lever is **retention** — a policy that never drops the never-reinforced critical fact in the first place — not a flag on its absence.

3.6 A methodological finding: single-sample probing manufactures results

The route to §3.5 produced a string of intermediate "discoveries" — a marker that rescued the catastrophe, a clean high-prior/low-prior boundary, a category-cue "primitive" — every one of which **evaporated when re-run as a rate.** Each was a single temperature-1.0 generation, and the no-marker baseline alone spans ~33–87% flagging across small samples. A 54-point baseline spread cannot support a claimed marker effect of a few points; the apparent effects were noise wearing a finding's clothes, and the tell was a marker "effect" whose direction *contradicted across the two model tiers.*

This is not a footnote; it is the same failure the project was built to catch. Tier 1 §6 already recorded that "15 scenarios is enough to establish dominance relations but not to estimate small effects precisely." The Tier-1 dominance gaps (embedding 1.00 vs summarization 0.47) cleared the noise floor; the Tier-2 marker effects were far smaller and were chased on single samples anyway, because single samples produce clean, legible, headline-shaped numbers, and clean numbers do not *look* like they need error bars. The discipline that built the anti-shiny-needle guard and the `live_only/bare` controls is what eventually caught the instrument mid-error — by refusing to pick the tier that agreed with the hypothesis. The transferable rule: **for effects smaller than a**

dominance gap, a single generation is a single scalar, and it hides the variance. Report rates with intervals or report nothing.

---

4. Threats to validity

- **Corpus scale.** 15 scenarios / 30 probes establishes dominance relations and large base-rate effects, but not small ones; the Tier-2 rate runs (N up to 120 per cell) address this only for the cells actually sampled.
- **A single, and broken, ACT-R formulation.** Tier 1 negates *the composition we introduced* (log-odds base-level summed with bounded cosine), not a units-correct ACT-R. The diagnosis is mechanistic and the fix is known; we did not run it, because §2.4 gives a non-experimental argument that even a corrected form cannot beat relevance on this recall-shaped workload.
- **The volume/attention effect is untested, not refuted.** Tier 2 tested one kind of junk: a single stale value beside its explicit correction — the easiest case. A window *full* of dead-but-similar context degrading the model by crowding (lost-in-the-middle) is a different, unmeasured mechanism. "Junk is always harmless" is refuted; "a window full of junk is harmless" is not addressed.
- **The model ladder's discriminating rung is unreached.** Context management exists to avoid running the expensive model on everything, so the deployment-relevant consumer is the *weak* model. The two current-generation tiers agree, but genuinely small/old models were inaccessible on this endpoint; the operating point most likely to show stale-drag is untested.
- **The witnessed-axis repair is single-sample.** The open-loop "flag" observation (§3.4) is exactly the kind of n=1 result §3.6 shows is unreliable. It is a hypothesis pending a rate-controlled replication, not a finding.
- **LLM judge.** The supersession judge's one demonstrated error is *strictness*, which runs against the zero-drag headline rather than toward it; the flag-rate judge is a regex over free text and is noisy, which is why §3.5 relies on rates and intervals rather than single classifications.

---

5. Discussion: what survives, at measured confidence

The project set out to test a cognitively motivated memory and instead produced a sharper map of where the hard problem actually lives. Stated at the confidence the data now bears:

- **Relevance solves recall on this workload; it cannot see liveness.** (Tier 1, robust: large, consistent dominance gaps.)

- **Liveness-as-suppression is a non-problem for the output.** Stale-but-present context does not drag a current model; 0/15 drag across 15 scenarios. (Tier 2 kill-test, robust: aggregated, unanimous.)
- **The catastrophic failure is context blindness — a missing fact the model cannot know is missing — and it is stochastic and model-dependent** (~40% silent on a strong model, ~75% on a small one). (Tier 2 variance run, robust: rate-controlled.)
- **No generic in-context completeness signal reliably repairs it.** (Tier 2, measured-dead.)

What is **not** established, and is honestly marked as such: the units-correct ACT-R on recall; the volume/attention effect; the weak-model rung; and the witnessed-axis repair, which remains a single-sample hypothesis. These are the experiments that would extend the study, and each now has a control designed to embarrass it.

The practical implication is a redirection of engineering effort. Because suppression doesn't matter and absence can't be flagged from inside the window, the leverage in long-horizon agentic memory is **upstream**: a retention policy that never evicts the never-reinforced critical fact — the stated-once allergy, the one-time vendor blacklist — is worth more than any amount of downstream context-rot scrubbing or completeness-flagging. Liveness re-enters not as a filter and not as a confidence prompt, but as a *retention priority*: the cost of dropping a fact should be a function of its consequence, not its recency or frequency. That is the form of the original ACT-R "importance" bet that this study leaves standing — relocated from the ranker to the eviction policy, and still owed a rate-controlled test.

---

6. Conclusion

Project Engram asked whether a cognitively motivated associative memory beats summarization for long-horizon agentic recall, and built a harness designed to embarrass that hypothesis. It succeeded twice over. In Tier 1 it falsified the strong claim — plain embedding retrieval dominates the recall frontier, and the ACT-R machinery, as composed, breaks on a units error — and corrected the question: recall is easy here; *liveness* is the axis relevance cannot see. In Tier 2 it falsified the survivor's first form — stale context that is present alongside the truth does not move the output, so suppressing it buys nothing — and then, under rate-controlled measurement, settled the deeper question it had reframed its way into: the catastrophic failure of long-horizon memory is **context blindness**, a missing fact the model confidently answers around because, from inside its window, it has no evidence the fact is missing. That failure fires at a stochastic, model-dependent rate, and **no generic in-window signal reliably corrects it.** The repair has to move upstream, to never

dropping the fact.

The most durable lesson is methodological and the study earned it the hard way: the same single-sample convenience that makes a result legible makes it unfalsifiable, and several intermediate findings here were noise that only a rate with an interval could expose. The harness did its job to the end — it killed the comfortable version of the idea, then killed the exciting replacement, and handed back something smaller, bounded, and true.

Appendix A — Reproduction

```
uv sync --extra dev

# Tier 1 — inner-loop frontier (real models; cached). ENGRAM_OFFLINE=1 for the deterministic CI path.
uv run python scripts/run_frontier.py

# Tier 2 — supersession kill-test (stale-but-present harm)
uv run python tier2_context_rot/run_killtest.py --models=claude-sonnet-4-6,claude-haiku-4-5-20251001

# Tier 2 — the rate-controlled variance run (the terminal result)
uv run python tier2_context_rot/variance_run.py 40
```

Tier-1 outputs land in `results/` (frontier curves, `frontier.json`, `probe_results.csv`); Tier-2 artifacts and the full reasoning trail are in `tier2_context_rot/` (`FINDINGS.md`, `results/`).

Appendix B — Glossary

- **Liveness** — whether a fact is currently true, independent of how well it matches the query. A function of recency and supersession; invisible to cosine relevance.
- **Stale rate** — fraction of supersession probes whose superseded value leaked into the window (Tier-1 precision proxy).
- **Drag** — fraction of cases where the model *answers* the stale value with the current value also present (Tier-2; the output-level analog of stale rate).
- **Witnessed / witnessless incompleteness** — a missing fact that left a write-time event (supersession, open loop) vs. one that simply aged out (stated-once), the dividing line for whether an honest in-context marker can be reconstructed at all.
- **Context blindness** — the model answers confidently from an incomplete window with no signal the window is incomplete; the catastrophic, stochastic, model-dependent failure mode this study terminates on.