# EpisodicRAG: A Controlled Interference Benchmark and a Diagnosed Retrieval Bottleneck in Graph RAG

Eugenio Noyola Leon (Keno)

## Abstract

Graph RAG systems retrieve without regard to whether a memory's original context resembles the context it is now being retrieved under. Six independently read prior-art sources converge on the same gap: no existing system combines graded, context-distance modulation of retrieval strength with a controlled evaluation of context-mismatched distractors on a typed entity-relation graph.

We build a three-tier controlled interference benchmark (correct, natural-distractor, and engineered-distractor episodes; 60 queries across four classes) and run four baselines at full scale.

A distance-based suppression mechanism designed against this gap is disconfirmed at full scale, along with seven further diagnostic signals, each failing for a distinct, identified reason. A decisive follow-up experiment isolates why: generation given oracle retrieval solves 80% of the interference benchmark; generation given the same systems' real, noisy retrieval solves 27–40%. Retrieval quality, not generation, is the dominant bottleneck (the system's memory search fails, not its ability to write a good answer once it has the right memory), and no retrieval-time signal tested here closes it.

A representation-level alternative (versioned, timestamped graph edges) is evaluated against two pre-registered gates and found unsupported by the underlying graph even on its own best-case example. We release the benchmark, the baseline sweep, and the full diagnostic record.

## 1. Introduction

Static graph RAG systems index once and freeze: retrieval carries no representation of whether the situation in which a memory was encoded resembles the situation it is now being retrieved under: the encoding-specificity gap (Tulving & Thomson, 1973), stated in engineering terms. Six prior-art sources read directly and in full, spanning cognitive science, agentic ML, industry memory infrastructure, and software fault taxonomy (Section 2), converge on the same absence:

none combines (a) continuous, graded context-distance modulation of retrieval strength, (b) a controlled evaluation with deliberately injected, context-mismatched distractors, and (c) a typed entity-relation graph substrate.

This paper tests that gap directly. A three-tier controlled interference benchmark is built (Section 3.1) and four baselines are run at full scale (Section 4.1). A distance-based suppression mechanism designed against the gap is disconfirmed at full scale, along with seven further diagnostic signals, each failing for a distinct, identified reason (Section 4.2). The decisive finding is a diagnosed retrieval bottleneck: generation given oracle retrieval solves 80% of the interference class; generation given real, noisy retrieval solves 27–40% (Section 4.3). Retrieval quality, not generation, is the dominant source of error, and no retrieval-time signal tested here closes the gap.

Three contributions: (1) a three-tier controlled interference benchmark, usable independent of any mechanism tested against it; (2) a four-baseline reference sweep at full 60-query scale; and (3) a diagnosed retrieval bottleneck: generation given oracle retrieval solves 80% of the interference class, generation given the same baselines' real, noisy pools solves 27–40%, the clearest evidence in the paper of where the actual error lives. A distance-based suppression mechanism and seven further diagnostic signals, each disconfirmed or reversed for a distinct, identified reason, and a representation-level substrate proposal evaluated against two pre-registered gates and closed before being built (Section 5), support and root-cause this finding.

**Background, for readers outside RAG/memory research**

RAG (retrieval-augmented generation) systems answer questions by retrieving relevant text from a document collection, then generating an answer from what's retrieved, rather than relying only on a language model's own trained knowledge. Graph RAG builds a structured knowledge graph from that text (entities and the relationships between them) instead of retrieving raw text chunks, and answers by traversing that graph. Episodic memory, in cognitive-science terms, is memory of specific, dated events (what happened, when) as distinct from general factual knowledge. The failure this paper studies, encoding specificity (Tulving & Thomson, 1973), is the finding that a memory is easiest to retrieve in a context resembling the one it was formed in, and hardest when a newer, more available memory about the same topic gets recalled instead, even though it's the wrong one for the question being asked. That's the interference this paper's benchmark is built to catch.

## 2. Related Work

Six sources were read directly and in full (not by abstract, not by proxy citation) spanning cognitive science, academic agentic ML, industry memory infrastructure, and software fault taxonomy:

| Source | Field | What it does | What it's missing |
|---|---|---|---|
| Stern & Nadel, "Drawing on Memory" (2026) | Cognitive science | Binary scene-trace presence/absence, real controlled ablation, +20pp on LongMemEval-S | No continuous modulation; no interference test, flagged as their own limitation |
| AUGUSTUS (2025) | Agent systems | Tag-context graph, cosine-similarity ranking | "Graded" is ordinary top-k ranking; untyped tag co-occurrence graph, not entity-relation; no interference test |
| SHIELDA (2025) | Software fault taxonomy | Names "Misaligned Memory Recall" as an exception type | No mechanism, no evaluation at all for this exception |
| Xiong et al. (ACL 2026) | Agentic ML | Threshold-based deletion rules | Discrete; observational proxy, not a designed distractor test |
| Salama et al. / MemInsight (Amazon AWS) | Industry | Attribute filtering / embedding top-k | Discrete gating; reuses a pre-existing benchmark slice, not a controlled test |
| SEEM (2026) | Agent systems | Typed entity-relation-temporal graph, PageRank + provenance retrieval | Closest graph-type match found; still no decay/potentiation/context-distance mechanism, no interference test |

Zero exceptions were found: none combines (a) continuous or graded context-distance modulation, (b) a controlled evaluation that injects a similar-but-context-mismatched distractor and measures precision/recall/false-positive rate, and (c) a typed entity-relation graph substrate. Most sources are flat; the two closest graph-structured attempts, AUGUSTUS and SEEM, use an untyped tag co-occurrence graph and a PageRank/provenance system, neither matching this project's typed extraction graph. Three sources (AUGUSTUS, SHIELDA, and SEEM) independently named this failure mode from engineering instinct alone, with no contact with the encoding-specificity literature motivating this project; convergent evidence is treated here as stronger than a purely theory-derived hypothesis alone.

This project's original question (does a purpose-built recency or frequency mechanism beat vanilla RAG) is separately answered already: MemBench (ACL Findings 2025) found it condition-dependent, not re-tested here; the recency/frequency classes (Section 3.1) are retained as a harness check, not a claim this paper makes. HippoRAG, Larimar, and AriGraph, read in the same pass but not among the six gap-defining sources, are consulted as prior art for the mechanism design in Section 3.3.

# 3. Method

## 3.1 Corpus and Evaluation Design

The evaluation is built to test a specific conjunctive claim, not episodic memory for RAG in general: whether a mechanism can (a) modulate retrieval strength continuously by context-distance, (b) under a controlled interference condition, (c) on a typed entity-relation graph substrate. This scoping follows directly from the gap identified against six independently read prior-art sources (Section 2), and is deliberately narrower than re-testing whether a purpose-built recency or frequency mechanism beats vanilla RAG: a question already answered, condition-dependently, by existing work (MemBench, ACL Findings 2025).

Corpus. The evaluation extends a pre-existing 41-episode synthetic biography corpus (E01–E41), unchanged aside from one targeted 3-sentence completeness fix to a single episode (E30), closing a gap between a dataset-notes claim and the text actually available to a retriever. Fifteen new episodes (E42–E56) are added as an interference-distractor tier. The corpus totals 56 episodes.

Query classes. Four classes span 60 queries: Recency (R01–R15) and Frequency (F01–F15), reused unmodified as a harness sanity check (Section 2); Context-shift/interference (CS01–CS15), the paper's central evidentiary class, with a new three-tier design (below); and Sequence/order recall (SEQ01–SEQ15), a secondary, orthogonal check that the mechanism doesn't regress an unrelated capability.

Interference class structure. Each of the 15 context-shift queries has three tiers: (Tier 0) one or more correct episodes; (Tier 1) a natural distractor, an existing episode with genuine surface overlap and mismatched framing, present for 14 of 15 queries; and (Tier 2) an engineered distractor, one newly authored episode (E42–E56) built to maximize surface similarity while remaining genuinely context-mismatched. Three Tier 1 pairs (CS07, CS13, CS14) are empirically confirmed (a prior evaluation's real runs actually substituted these episodes for the correct answer) rather than inferred, and this distinction is preserved through scoring rather than flattened. One engineered pair, CS09, is a documented weak or forced fit (its distractor, E50, is a genuine second true memory rather than an adversarial construction) and is reported separately rather than pooled into an aggregate metric.

**Worked example: CS01**

*Query: "Who is John closest to emotionally right now?" Ground truth: Maya and Ethan both, the Ethan relationship fully repaired by 2024.*

**Tier 0 (correct, E39, 2024):** "...He called Maya on Sunday mornings. He called Ethan on Thursday evenings... He was sixty-four years old and his life had the quality he had been reaching toward without knowing its name for most of his adult life."

**Tier 1 (natural distractor, E26, 2010):** "The relationship with Ethan had been difficult since the divorce and by 2010 it had settled into something that was not quite estrangement but was close enough that the distinction felt mostly semantic... Ethan was sixteen and angry in a way that had not softened with time."

**Tier 2 (engineered distractor, E42, 2004):** "One section asked each parent to describe, in their own words, which child they felt closest to and why... John wrote that Maya, at thirteen, had always been easier for him to read."

The engineered distractor's design is visible directly in its text: E42 reuses the query's own phrase, "closest to," inside a 2004 custody-evaluation scene, giving a retriever a lexical match the correct 2024 episode does not share. The natural distractor is topically adjacent (same relationship, same person) but describes a superseded emotional state. This is the shape every one of the 15 interference queries follows.

Sequence/order class. Fifteen pairwise "which came first" queries are built from 25 of the 41 original episodes; five, including E30 and E41, are excluded for ambiguous dating. Gap sizes span sub-year to 41 years; reported as a single results-table line, not a headline finding.

Metric. The interference-class metric is precision, recall, and false-positive rate on which episode's chunks reach the retrieval-stage context, not what the generated answer emphasizes, requiring episode-keyed retrieval logging (Section 3.2) in place of a prior entity-keyed log that couldn't distinguish the two.

Baselines. Four systems are compared: LightRAG naive (vector-only, isolating graph structure); full LightRAG local/global (no episodic mechanism, the system being improved on); Xiong et al.'s deletion rule, reimplemented from published formulas after their repository proved unusable as released; and MRAgent, a structurally different graph substrate (Cue-Tag-Content), as a second comparison arm. MemInsight was considered and deferred; AUGUSTUS was excluded, having no available implementation.

## 3.2 Baselines: Implementation

All four baselines run at full 60-query scale from the start, scored by the same harness: a rule-based scorer for the interference and sequence classes, computed from episode-keyed retrieval logs, and an LLM-as-judge scorer for the recency and frequency classes' free-text answers.

Naive uses LightRAG's built-in vector-only mode. Local and global run full LightRAG with no episodic mechanism attached, through a single reranker-hook patch on process_chunks_unified (replacing a three-site patch structure used previously). Xiong et al.'s threshold-based deletion rule is reimplemented from its published formulas, since the original repository is a patch set against an unrelated autonomous-driving codebase, not a standalone method. MRAgent runs via an isolated, adapted clone (three compatibility issues fixed during adaptation), its Cue-Tag-Content graph substrate structurally distinct from LightRAG's.

The sequence-recall judge prompt is fixed prior to these runs: an earlier version requested a verdict before its supporting rationale, letting the model commit to a verdict the rationale didn't actually support; the fix reorders the prompt so rationale precedes verdict.

### 3.3 Mechanism Attempts: Design

A graded, distance-based suppression mechanism is designed, modulating retrieval strength continuously by embedding distance between a query and each candidate episode, informed by the distance diagnostic in Section 4.1 but described independent of that result. The design cites HippoRAG, Larimar, and AriGraph as prior art for decay-, potentiation-, and encoding-specificity-style mechanisms, avoiding re-derivation of an existing functional form. Two parameters, a distance threshold (d0) and steepness (s), are tuned against five held-out triples not overlapping the 15 real interference queries, then frozen (tune-then-freeze-then-evaluate-fresh). A suppression-exclusion check verifies the weighting can change exclusion decisions in global mode, not just compute weights correctly. The mechanism is evaluated at full scale (60 queries, 3 modes) against one pre-registered criterion: at least 25 points of engineered-tier false-positive-rate improvement.

Three further signals are tested directly against the real 15-query set: chunk-level embedding distance (the distance diagnostic at finer grain), entity overlap (named-entity overlap between query and episode), and temporal drift (recency gap between correct episode and each distractor).

A structurally different approach (cross-encoder joint query-passage attention rather than independently compared embeddings) is tested three ways: a curated 14-query, three-way (correct/natural/engineered) comparison; an offline episode-to-episode confusability check with no query, comparing known-confusable pairs against random unrelated pairs; and a pool-level re-ranking test against the actual candidate pools already logged during the baseline sweep (20–82 chunks/query), recomputing recall and false-positive rate with the harness's existing scorer.

A final diagnostic isolates retrieval from generation directly: for each of the 15 queries, an answer is generated under an oracle context (correct episode only) and again under the real naive and global candidate pools, regenerated fresh. The judge is calibrated against three hand-written cases (clearly correct, incorrect, partial) before scoring real data. Each generation runs as a separate operating-system process, after a direct test found LightRAG's response cache persisting across fresh build_rag() calls within one process: two calls with identical query text but different forced contexts otherwise returned identical cached text.

## 4. Results

### 4.1 Baseline Sweep Results

All five baseline arms were run at full 60-query scale. Results:

| Baseline | CS precision | CS recall | CS natural FPR | CS engineered FPR | R/F correct/partial/incorrect | SEQ correct/incorrect/unclear |
|---|---|---|---|---|---|---|
| Naive | 0.40 | 0.80 | 0.643 | 0.60 | 21/6/3 | 15/0/0 |
| Local | 0.385 | 0.667 | 0.50 | 0.60 | 22/5/3 | 14/1/0 |
| Global | 0.40 | 0.80 | 0.643 | 0.60 | 22/6/2 | 15*/0/0 |
| Xiong et al. | 0.40 | 0.80 | 0.643 | 0.60 | 21/5/4 | 13/2/0† |
| MRAgent | 0.444 | 0.267 | 0.071 | 0.267 | 21/8/1 | 15/0/0 |

**Global's SEQ score is the manually corrected figure; the raw automated judge score was 14/1/0, with one verdict (SEQ08) overridden as a confirmed judge misfire rather than a real answer error. †Xiong's SEQ score is the raw, uncorrected automated figure: both of its two INCORRECT verdicts (SEQ08, SEQ09) read as judge misfires on direct inspection, consistent with the same pattern corrected for Global, but no manual override was applied here.*

Naive and global convergence. Under this configuration (no dedicated rerank model, the episodic-weight patch left neutral), naive mode and global mode returned identical retrieved chunks for 9 of 15 interference queries (a later, directly re-verified recount during a subsequent diagnostic found 8 of 15; the discrepancy was flagged at the time as minor counting variance and not resolved), consistent with process_chunks_unified falling back to plain vector-similarity ranking whenever no real reranker is active. "Naive versus graph-based retrieval" isn't yet a clean comparison under this configuration; that requires a real mechanism or reranker in the loop.

Xiong et al. deleted zero natural and zero engineered distractors across all 15 queries; its one correct-episode deletion (CS06/E24) only avoided hurting score because that query had a backup correct source; any query without that redundancy would face a real risk. Its interference-class numbers are numerically identical to naive.

A sequence-recall judge misfire (SEQ08) recurred across two baselines' scoring, meeting the project's own bar for revisiting the prompt rather than treating it as isolated; reordering to reasoning-before-verdict fixed 5 of 6 test cases, and MRAgent's later, structurally different scoring (15/0/0, zero misfires) is corroborating evidence the fix generalizes.

MRAgent's natural and engineered false-positive rates (0.071, 0.267) are far lower than every other baseline (0.50–0.643 natural, 0.60 engineered), but so is its recall (0.267 vs. 0.667–0.80), with 6 of 15 queries returning nothing; the same underlying cause, not two findings: an aggressively narrow retrieval loop, not a solved interference class.

Ahead of any mechanism design, embedding distance was checked as a candidate signal. Per query, strict correct<natural<engineered ordering held only 4 of 14 times (29%), including one large inversion (CS13, natural distractor 0.28 closer than correct). Cross-referenced against which of the five real baseline arms actually surfaced a distractor as a false positive, the picture sharpened: queries closer than ~0.69 were surfaced by 3–5 of 5 arms; farther queries by 0–1 of 5, most pronounced on the engineered tier (one clear exception: CS08, close but never surfaced). MRAgent's low FPR was not distance-driven: it stayed low in both buckets alike, i.e., blunt conservatism, not distance-awareness.

## 4.2 Mechanism Results

The baseline sweep's distance-bucketed pattern (Section 4.1) motivated a mechanism narrower than an original four-mechanism pitch (graded, distance-based suppression), tuned separately (Section 3.3) rather than reusing the exploratory ~0.69 threshold directly. Seven further diagnostic signals were then tested against the same 15-query set:

| Signal | Test | Result |
|---|---|---|
| Distance suppression (Phase 6 mechanism) | Full 60-query × 3-mode eval, frozen d0=0.65, s=0.10 | Disconfirmed: engineered FPR 0.60→0.60 (target ≥25pt drop); natural FPR improved only in global (−7.1pt), more than engineered's zero movement, inverting the predicted asymmetry; global cost 13.3pt recall (CS07/CS14 suppressed) |
| Chunk-level distance | Correct-vs-natural gap, chunk grain | Confirmed negative (−0.0025), same direction as episode level |
| Entity overlap | Named-entity overlap, natural tier | Structurally 0/14: queries omit disambiguating entities by design |
| Temporal drift | Recency gap, correct vs. distractor | Engineered tier weak positive (+2.15yr); natural tier wrong-direction (−0.54yr), CS09/CS13 favor a newer wrong answer |
| Cross-encoder, curated (correct vs. natural) | 14-query, 3-way comparison | 9/14 (64.3%, p=0.424): partial positive, beats bi-encoder (8/14) |
| Cross-encoder, curated (correct vs. engineered) | Same 14 queries | 5/14, below chance: audited, real, traced to CS09/CS13 |
| Episode confusability (offline) | Known-confusable vs. random pairs, no query | p=0.062→p=0.031 (mean aggregation): sole signal to survive every audit; encoding-time only |
| Cross-encoder pool re-rank (real gate) | Actual logged 20–82-chunk pools | Regression: engineered FPR +6.7pt; natural FPR +21.4pt (local)/+7.1pt (global); recall gain indiscriminate |

The Phase 6 disconfirmation's root cause, verified against logged query embeddings: the real 15-query set's correct/natural/engineered distance tiers overlap almost completely at d0=0.65 (means within 0.034 of each other); a parameter/corpus mismatch, not a broken implementation. The mechanism's build-stage checks (Section 3.3) already confirmed it excludes chunks correctly, and no improvement concentrated in the close-distance bucket. For CS09 and CS13 specifically, a recency- or surface-similarity-favoring signal would actively suppress the correct answer for a newer, wrong-context one; the CS07/CS14 suppressions in Phase 6's full run (table, row 1) are this same failure mode at scale. CS15 behaved atypically in two of the four distance/drift diagnostics, traced to two separate, unrelated causes.

The episode-confusability result (table, row 7) is the only signal to survive every robustness check applied to it, but remains an encoding-time signal: converting it into a retrieval-time

mechanism was not attempted. The pool-scale regression (row 8) is tentatively root-caused to the cross-encoder model (ms-marco-MiniLM-L-6-v2) being trained for general passage relevance rather than correct-vs-plausible-distractor discrimination specifically, a distinction the narrower three-way tests (rows 5–6) never actually exercised.

### 4.3 The Retrieval Bottleneck (Oracle vs. Noisy-Pool Generation)

A final diagnostic (Section 3.3) separated retrieval from generation directly: each of the 15 interference queries was answered under an oracle context (correct episode only) and under the real, unmodified naive and global candidate pools. Aggregate accuracy:

| Condition | CORRECT | PARTIAL | INCORRECT | Accuracy |
|---|---|---|---|---|
| Oracle | 12 | 0 | 3 | 80.0% |
| Naive pool | 4 | 5 | 6 | 26.7% |
| Global pool | 6 | 2 | 7 | 40.0% |

Both noisy-pool conditions fell well below the 80.0% oracle ceiling (naive 53.3 points below, global 40.0 points below), clearing a pre-registered 30-point bottleneck-confirmed threshold, naive by more than 3×. Of the 12 oracle-CORRECT queries, only 3 (CS04, CS10, CS12) held CORRECT under both noisy pools; 9 degraded under at least one; 3 (CS05, CS07, CS09) collapsed cleanly to INCORRECT under both. CS01 and CS08 were wrong under all three conditions including oracle (a different, milder pattern, discussed below).

Global (40.0%) beat naive (26.7%) by 13.3 points but with a different error shape: more polarized (6/7/2) versus naive's more hedged middle (4/6/5). A larger pool (43–82 chunks vs. naive's 20) and graph-structured context may explain this, but it wasn't pursued further.

CS09 and CS13 carry the same "known confusion case" label but resolved to opposite root causes. CS09 was solved cleanly under oracle context and failed under both noisy pools by substituting E50, its own documented weak-fit distractor, for the correct episode E01; this directly confirms, in the generation domain, a risk already flagged in the retrieval domain (Section 3.1):

**Worked example: CS09, same query, two conditions**

**PASS (oracle context, E01 only):** “John and his father shared a memorable trip to Candlestick Park in June 1968 to watch a San Francisco Giants game... his father placed his hand on John's shoulder and encouraged

him to pay close attention to Willie McCovey at the plate." *Judge: CORRECT — "clearly and primarily states that they shared a trip to a 'San Francisco Giants game'... a formative bonding experience."*

**FAIL (naive pool, real 20-chunk retrieval):** "John and his father shared a reliable routine of listening to San Francisco Giants baseball games on the radio during the summer of 1967 [4]... he valued this activity as a meaningful form of closeness." *Judge: INCORRECT — "directly contradicts the ground truth that they shared 'Giants games' by attending them at 'Candlestick Park.'"*

Both answers are fluent, well-cited, and internally coherent: the FAIL case is not a hallucination, it is a correct answer to a different, wrongly retrieved episode (E50). The failure is visible in the answer text itself, not only in the verdict: this is what "retrieval, not generation, is the bottleneck" looks like in practice.

CS13 failed even under oracle context: neither correct episode (E04, E05) explicitly connects history-reading to technology-hype skepticism. The closest passage, from E04, is thematically adjacent but not the claim itself — "how the same mistakes kept appearing in different costumes across centuries" — never mentioning hype, dot-coms, or AI. The ground-truth claim was asserted in the dataset design notes but never written into either correct episode, the same category of gap already found and fixed once elsewhere in the project. This is a corpus-completeness gap, not a generation failure: the model correctly reported the text didn't support the claim rather than fabricating a connection.

CS01 and CS08 form a third, milder pattern: wrong in all three conditions, but each judge rationale describes the model declining to commit to a superlative framing the ground truth required, even though the underlying facts were plausibly present across four correct episodes each: a synthesis-across-many-documents difficulty, no single episode states the superlative claim outright, distinct from CS09's substitution pattern or CS13's flat content gap.

Two checks rule out infrastructure artifacts. The judge was calibrated beforehand against three hand-written test cases (clearly correct, clearly incorrect, clearly partial); all three matched their expected verdict. And because LightRAG's response cache was found to persist across fresh build_rag() calls within one process (two calls with identical query text but different forced contexts otherwise returned identical cached text), each generation ran as a separate OS process; a post-hoc check confirmed zero identical-text collisions and zero empty answers across all 45 generations.

Both noisy-pool conditions land 30-plus points below the oracle ceiling, and CS09 shows noise alone can flip a correct generation to incorrect with the right episode present in the pool; this undermines the premise of every retrieval-time signal in Section 4.2, since even a perfectly ranked correct episode leaves the rest of a noisy pool in context. The 80.0% ceiling itself matters: solving retrieval completely wouldn't solve this class outright, and one of its three failures (CS13) is a corpus gap, not a generation limitation.

## 5. Discussion and Future Work

Section 4.2's diagnostics share a structural limitation: distance, entity overlap, temporal drift, and the cross-encoder all operate on the same representation (text embedded independently and compared after the fact), which can't express "this fact used to be true, now a different fact is true" as a first-class property, since LightRAG's extraction attaches time only implicitly, to documents, not relationships. Seven signals failing for structurally different reasons motivated asking whether the representation itself needed to change, rather than testing an eighth signal on top of it. This motivated a pitch, evaluated but not built: attach the source episode's date to every extracted graph edge; mark an edge superseded when a new extraction contradicts it; and for relation types where current state matters, traverse to the most recent non-superseded edge rather than the nearest one by embedding distance: a change to what the graph stores, not what's compared at query time. Against the six prior-art sources (Section 2), the closest, SEEM, already has a typed entity-relation-temporal graph with PageRank/provenance retrieval, but no graded modulation under controlled interference; that's the axis this project claims, and this pitch doesn't resolve it either, since it proposes a discrete rule.

Two pre-registered checks ran before any implementation commitment. A shape audit classified all 14 interference queries with a natural distractor against the mechanism's target shape (a clean temporal contradiction): only 1 of 14 (CS01) fit cleanly, 2 more (CS05, CS11) as soft cases: a generous count of 3 of 14, exactly at the pitch's own ≤4-of-15 concern threshold. Worse, 7 of 14 would have the wrong answer actively favored by a naive most-recent-wins rule (CS02, CS03, CS04, CS09, CS10, CS13, CS15), overlapping all 6 of an independent temporal-drift inversion set found weeks earlier by a different method (Section 4.2). A second check tested whether the graph structure could support the mechanism at all, on the pitch's own best-case example: CS01's

John–Ethan edge retains real per-episode provenance (file-path spanning all 8 contributing episodes, 9 source chunks) but its description was already synthesized into two narrative clusters, not eight separable, dateable versions: no way to query what was true in 2010 versus now, confirming at the edge level a node-level consolidation pattern found earlier. The other two checked queries (CS05, CS11) revealed a more basic problem: the extraction never linked the relevant episodes into a comparable relation at all. Only 1 of 3 checked queries had testable structure, and that one confirmed the merging concern. A third gate (re-reading SEEM's own mechanism to confirm whether it already closes this gap) was not run, since neither open question it would answer changes that this project's graph doesn't currently support the mechanism regardless.

Phase 8 does not proceed as scoped. If pursued, the realistic scope is not a timestamp field on an existing edge but re-extraction with contradiction-preserving prompting so distinguishable, dateable edge versions exist in the first place, closer to a new ingestion pipeline than a schema addition. This audit also tested the mechanism's ceiling only against a query set authored before the idea existed; a fair test of whether versioned-edge supersession helps in principle would need its own purpose-built, contradiction-shaped test set, the same discipline used for Phase 6's tuning triples (Section 3.3), applied here in the opposite direction.

## 6. Limitations

Sample size. The central evidentiary class is 15 interference queries; the decisive diagnostic (Section 4.3) draws its conclusion from those 15 under three generation conditions each. Reported statistics at this scale (the episode-confusability result, $p=0.031$; its offline pairing, $p=0.062$) should be read as suggestive on a small, fixed sample, not confirmatory. Per-query findings like CS09's and CS13's root causes are case studies at this scale, not statistically powered results.

Single corpus, substrate, and embedding model. All results come from one 56-episode synthetic biography corpus, one language, built on LightRAG's specific pipeline and a single embedding model. Whether the baseline sweep's pattern, the mechanism failures, or the retrieval bottleneck transfer to real-world corpora, other graph-RAG systems, or other embedding models was not tested.

Judge scoring and mechanism scope. Free-text answers are scored by an LLM judge calibrated against three hand-written cases, not measured against human inter-rater agreement; the sequence-recall judge's known residual gap for transitive-inference answers (Section 3.2) is corroborated but not directly re-tested since the fix shipped. Phase 6's negative result applies to one functional form (a two-parameter sigmoid, tuned against five held-out triples) and does not rule out distance-based suppression generally: the root-caused parameter/corpus mismatch (Section 4.2) suggests a different tuning procedure remains untested, not that the signal is unusable. The ~0.69 distance threshold (Section 4.1) is explicitly descriptive, found by inspecting the same 15 reporting queries, and was never reused as a design parameter; any future use needs independent justification. A shared 500/day request quota shaped what could be run in a single pass (tuning-set size, diagnostic variants) rather than reflecting an unconstrained design choice throughout.

Evidentiary weight. Section 4.3's bottleneck finding rests on one diagnostic: 15 queries, three conditions each, one calibrated judge. If that alone isn't sufficient grounds for its conclusion, the paper's contribution doesn't collapse: the interference evaluation design (Section 3.1) and the four-baseline sweep (Section 4.1) stand independently as a validated benchmark, consistent with how this project treated Phase 5 as a legitimate standalone stopping point throughout its own decision gates, not only in retrospect.

## 7. Conclusion

This paper tested a specific, conjunctive gap in graph RAG's treatment of episodic interference and reports what testing it found. The mechanism motivated by that gap (graded, distance-based suppression) was built, tuned, and disconfirmed at full scale; seven further diagnostic signals failed or reversed, each for a distinct reason; a representation-level alternative was pitched, gated, and closed before being built. None of that is the headline result. The headline result is that retrieval quality, not generation, is the dominant bottleneck: generation given perfect retrieval solves 80% of the interference class; generation given the same baselines' real, noisy pools solves less than half of that. Every retrieval-time signal tested here attacked the wrong layer; this is visible only because the negative results before it were tracked and reported honestly rather than discarded on the way to a positive one.

What remains is a validated three-tier interference benchmark, a four-baseline reference sweep at full scale, and an evidenced diagnosis of where the next attempt should go: not another retrieval-time re-ranking signal, but either the representation-level change Section 5 costs out and declines to build prematurely, or an explicit state-tracking layer that survives noisy retrieval rather than depending on ranking it away. Both are real next projects; neither is attempted here. The underlying protocol (build controlled interference, measure retrieval, measure oracle generation, and separate the two) generalizes beyond this corpus and this specific failure mode, independent of whether any particular mechanism built on top of it succeeds.

## References


Anokhin, P., Semenov, N., Sorokin, A., Evseev, D., Kravchenko, A., Burtsev, M., & Burnaev, E. (2025). AriGraph: Learning knowledge graph world models with episodic memory for LLM agents. IJCAI 2025. https://arxiv.org/abs/2407.04363

Das, P., Chaudhury, S., Nelson, E., Melnyk, I., Swaminathan, S., Dai, S., Lozano, A., Kollias, G., Chenthamarakshan, V., Navrátil, J., Dan, S., & Chen, P.-Y. (2024). Larimar: Large language models with episodic memory control. ICML 2024. https://arxiv.org/abs/2403.11901

Gutiérrez, B. J., Shu, Y., Gu, Y., Yasunaga, M., & Su, Y. (2024). HippoRAG: Neurobiologically inspired long-term memory for large language models. NeurIPS 2024. https://arxiv.org/abs/2405.14831

Jain, J., Maheshwari, S., Yu, N., Hwu, W., & Shi, H. (2025). AUGUSTUS: An LLM-driven multimodal agent system with contextualized user memory. LAW 2025 Workshop @ NeurIPS 2025. https://arxiv.org/abs/2510.15261

Lu, Z., Li, D., Shi, Y., Wang, B., Wang, L., & Hu, B. (2026). Structured episodic event memory (SEEM). arXiv preprint. https://arxiv.org/abs/2601.06411

Salama, R., Cai, J., Yuan, M., Currey, A., Sunkara, M., Zhang, Y., & Benajiba, Y. (2025). MemInsight: Autonomous memory augmentation for LLM agents. arXiv preprint. https://arxiv.org/abs/2503.21760

Stern, B., & Nadel, P. (2026). Drawing on memory: Dual-trace encoding improves cross-session recall in LLM agents. arXiv preprint. https://arxiv.org/abs/2604.12948

Tan, H., Zhang, Z., Ma, C., Chen, X., Dai, Q., & Dong, Z. (2025). MemBench: Towards more comprehensive evaluation on the memory of LLM-based agents. Findings of the Association for Computational Linguistics: ACL 2025, 19336–19352. https://arxiv.org/abs/2506.21605

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review, 80(5), 352–373.

Xiong, Z., Lin, Y., Xie, W., He, P., Liu, Z., Tang, J., Lakkaraju, H., & Xiang, Z. (2026). How memory management impacts LLM agents: An empirical study of experience-following behavior. ACL 2026. https://arxiv.org/abs/2505.16067

Zhou, J., Chen, J., Lu, Q., Zhao, D., & Zhu, L. (2025). SHIELDA: Structured handling of exceptions in LLM-driven agentic workflows. arXiv preprint. https://arxiv.org/abs/2508.07935