S3-RAG-01 Phase D Findings: Bounded Null-Hysteresis Under Saturated Current-Record Retrieval

Jeremy C. Jones

HoldingLight LLC

ORCID: 0009-0007-2515-3774

contact@universalcollapse.com

v1.0 · 2026-05

---

§1. Headline finding

Under saturated current-record retrieval, S3-RAG-01 found no detectable post-swap stale-pack hysteresis. Across 420 fact-classifications in the swap condition (C2), zero matched the pre-swap pack. This establishes a bounded clean-record baseline; generalization to partial retrieval, longer pre-swap exposure, or memory-enabled settings is left to planned stress variants (see §6).

Scope within the T16 program. S3-RAG-01 is the slot-4 demonstration within a broader T16 5-substrate study (parent paper: Jones 2026, DOI 10.17605/OSF.IO/JPXCU) that tests S_1/S_2/S_3 signatures across pre-specified AI substrate classes: frontier OpenAI, frontier Anthropic, frontier Google, search-grounded/RAG, and local open-weights. The mini-RAG harness occupies the search-grounded/RAG slot. Cross-substrate findings from the other four slots will be reported in the parent T16 paper; the present paper is the focused empirical demonstration on the RAG slot. Single-generator execution (`gpt-5.4-2026-03-05`) is by design — the question here is RAG-harness behavior under controlled retrieval conditions, not cross-vendor behavior. Multi-vendor extension of the RAG probe specifically is mapped as a stress-test item.

Three findings follow from the data: null hysteresis under complete retrieval (C1, C2); setup-conditioned prior suppression when the decisive chapter is absent (C3); and source-gap commitment to the non-prior outcome as a separate failure class from stale-pack residue. Section 3 develops each in turn.

---

§2. Methods

§2.1–§2.4 describe the corpus, closed-book calibration, harness, and three retrieval conditions. §2.5 documents the scoring procedure, including the classifier-tightening pass between v1 and v2 of the analysis.

§2.1 Corpus

The corpus comprised three custom narrative-fiction stories — *Salt-Marsh* (S1), *Marrow Run* (S2), and *Containment at Auchenval* (S3) — each authored in two outcome variants (Pack A and Pack B) that diverged only at Chapter 5. Chapters 1–4 of each pack were identical between variants; Chapter 5 contained the seven decisive facts probed by the test battery. The full corpus comprised 30 chapter files (3

stories × 2 packs × 5 chapters). Chapter content, the seven-fact test battery per story, and the answer-option-to-pack mapping are documented in the design doc (§3–§5).

The stories were custom-written to avoid known-title contamination from training data. Closed-book baselines (§2.2) verified that the model did not recognize the stories by name and did not produce prior-aligned content from titles alone.

§2.2 Calibration

Three closed-book baselines were executed against the generator without retrieval. **T1** verified that the model declined to answer story-specific questions when prompted only with story titles (the title-contamination gate); it passed on all three stories after the renaming of S3 to *Containment at Auchenval*. **T2** measured the model's tendency to commit to outcome picks given minimal context (character names plus binary framing); it surfaced strong narrative-completion priors and was reframed in audit as a minimal-context prior measurement rather than a contamination test. **T3** produced the load-bearing closed-book baseline: 20 repetitions per topic at temperature 0.7, scoring per-fact A/B picks across the seven-fact test battery.

T3 yielded 100% unanimity in 20 of 21 fact-positions. The single exception — S3 F7 (Nyota's behavior post-retirement) — split 17A/3B (85% A), providing the only fact with a non-saturated baseline for graded directional measurement, and the load-bearing per-fact diagnostic for Phase D analysis. Full protocols and results are in the calibration doc.

§2.3 Harness

The mini-RAG harness combined a cosine-similarity vector store over `sentence-transformers/all-MiniLM-L6-v2` embeddings (one chunk per chapter file) with a generator slot occupied by `gpt-5.4-2026-03-05` (SYSTEM_REGISTRY slot 4). Retrieval returned `retrieval_k=5` chunks per query, scored by cosine similarity. The k=5 setting was adjudicated during Phase C smoke after a k=3 retrieval failure on the consolidate prompt; the k=5 setting reframes C2 from a selective-retrieval test to a full-context residue test (rationale and adjudication in the smoke findings doc).

Pack-switching was implemented as a `load_pack` / `clear` interface on the vector store: at any step, exactly one pack's chapter files were indexed; swapping cleared the index and re-indexed from the target pack's chapter files. Conversation history persisted across the swap; only the retrieval state changed. Full harness implementation is documented in the harness sketch.

§2.4 Conditions

Three retrieval conditions were probed, each across all three stories (9 topics total) at 20 sessions per topic. The design produced **180 sessions, 420 conversational turns, and 1,260 fact-classifications** (7 facts × 20 sessions × 9 topics); each condition contributed **420 fact-classifications** (7 facts × 20 sessions × 3 stories).

C1 — baseline retrieval. Pack B loaded throughout. A single-turn session: the test battery was administered with Pack B's five-chapter content fully retrievable. C1 measures the baseline rate at which the model produces pack-aligned answers when current retrieval is complete and consistent.

C2 — pack swap. Five-turn session. Pack A loaded at steps 1–2: the model summarized the story across two consolidate prompts, producing assistant turns referencing Pack A's Chapter 5 content into the conversation history. At step 3, the harness swapped to Pack B; steps 3–4 produced further consolidate summaries from the new retrieval state. At step 5, the test battery was administered. C2 measures whether

prior conversational exposure to Pack A leaves detectable residue in test-battery answers when the active retrieval state is Pack B.

C3 — setup-only retrieval. Single-turn session. Chapters 1–4 loaded (these are shared between packs); Chapter 5 omitted from the index. The test battery was administered with the decisive ending chapter unavailable to retrieval. C3 measures the model's behavior when the directly relevant source content is absent but related (shared) setup context is present.

§2.5 Scoring

Each test-battery answer was scored on the model's bolded pick — the bolded text following the answer prompt, interpreted as the model's committed choice between Pack A and Pack B outcomes per fact. The answer-option-to-pack mapping was fixed in the probe specification and identical across conditions.

Per-condition scoring categories:

- **C1**: pack-aligned (matches Pack B's Chapter 5) vs misaligned vs ambiguous.
- **C2**: matches_current_pack (Pack B, the swap-to state) vs matches_other_pack (Pack A, the swap-from state) vs ambiguous. The matches_other_pack count is the operational measure of stale-pack hysteresis.
- **C3**: prior_aligned (matches the T3 modal pick for the fact) vs source-gap non-prior commitment (committed pick contradicting the T3 modal — emitted as `invent_other` in the score output) vs ambiguous vs abstain. C3 was scored on output content rather than retrieval coverage; the `target_chunk_retrieved` field was found degenerate at k=5 (with 4 chapters loaded against k=5, top-k trivially returns all loaded chunks) and is not used as a scoring gate. The `invent_other` code label predates the present analysis; the paper uses "source-gap non-prior commitment" (or, in shorthand, "source-gap commitment") to describe the category since the picks are explicit A or B commitments (the binary choice space) opposite to the closed-book prior — not novel content outside both packs.

A classifier-tightening pass was executed between v1 and v2 of the analysis. The v2 classifier reduced ambiguous picks from 24.0% to 4.1%. In regression checks against v1, only three substantive deltas appeared in the headline tables: two A↔B reversals and one A/B→? change. All three were attributable to the extraction-bug correction (the v1 classifier had been matching bolded question text rather than bolded answer text in some C3 records) rather than true regressions. Pattern extensions in v2 included comma-not contrast structure, semicolon-split negation, comparative phrasing, and one corpus-keyword pattern (S2 water-rights narrative); the corpus-keyword pattern was verified against T3 baselines and produced zero baseline regressions. Both v1 and v2 outputs are preserved for audit; the classifier extension is documented in `classifier_v2_notes.md`.

---

§3. Results

The three-part finding architecture: null hysteresis under complete retrieval (§3.1), setup-conditioned prior suppression when the decisive chapter was absent (§3.2), and source-gap commitment to the non-prior outcome as a separate failure class (§3.3).

§3.1 Null hysteresis under complete retrieval

The C1 baseline established that current-record retrieval is operationally clean. Across 420 fact-classifications (3 stories × 7 facts × 20 sessions), every committed pick aligned with the active retrieval

pack. Under the v2 classifier (see §2.5), C1 yielded 100% pack-aligned picks, 0% misaligned picks, and 0% ambiguous picks.

The C2 swap condition produced the central hysteresis result. Across 420 fact-classifications spanning three stories × seven facts × 20 sessions, every committed pick matched the swap-to pack (current_pack); zero picks matched the swap-from pack (other_pack). The per-story split was identical across all three stories:

Table 3.1 — Per-story C2 hysteresis split (v2 classifier).

Story	n picks	matches_current (Pack B)	matches_other (Pack A)	ambiguous
Salt-Marsh (S1)	140	140 (100%)	0 (0%)	0 (0%)
Marrow Run (S2)	140	140 (100%)	0 (0%)	0 (0%)
Containment at Auchenval (S3)	140	140 (100%)	0 (0%)	0 (0%)
Total	**420**	**420 (100%)**	**0 (0%)**	**0 (0%)**

The result is consistent with retrieval dominance under saturated current-record retrieval. Under the conditions tested — complete Pack B chapter coverage at retrieval_k=5, two prior consolidate summaries from Pack A in conversation history, swap at step 3 of a five-turn session — the model's committed answers aligned with the post-swap retrieval state in every observed case. No detectable residue from prior conversational exposure to Pack A was observed in any fact-classification.

Statistical bound on the stale-pack rate. At the fact-classification unit, the rule-of-three approximation gives an event-level one-sided 95% upper confidence bound of approximately 0.71% on the underlying stale-pack rate under this tested condition.[1] The point estimate is 0.0%; the methodological claim is that the event-level rate is at most a small fraction of one percent under this configuration.

The boundary conditions of this finding — k=5 saturation against five-chapter packs, two pre-swap exposures, single model, single corpus class — are detailed in §4 and §6.

§3.2 Setup-conditioned prior suppression

C3 — setup-only retrieval, with Chapters 1–4 loaded (shared across packs) but Chapter 5 omitted — produced the only condition in which a graded prior-suppression measurement is possible. Twenty of twenty-one facts in the dataset were saturated at T3 baseline (100% unanimity for one direction) and therefore cannot support a graded prior-suppression estimate — although they can still depart from those saturated priors, as the appendix data show. The remaining fact — S3 F7 (Nyota's behavior post-retirement) — provides the only non-saturated baseline (17A/3B = 85% A) and thus the load-bearing per-fact diagnostic for directional movement.

[1] The rule of three gives an approximate one-sided 95% upper confidence bound for a binomial event rate after zero observed events: approximately 3/n. For n = 420, this gives 0.714%; the exact Clopper-Pearson upper bound is approximately 0.711%. Because observations are nested within stories, facts, and sessions, the bound is used descriptively at the fact-classification unit rather than as a population-level estimate for RAG systems generally.

At T3 baseline (closed-book, no retrieval), the model picked Pack A's outcome on 17 of 20 sessions (85% A). Under C3, with Chapters 1–4 retrieved but Chapter 5 unavailable, the model picked Pack A on 3 of 20 sessions (15% A) — a 70-percentage-point shift away from the closed-book prior.

Table 3.2 — S3 F7 across conditions (v2 classifier).

Condition	A picks	B picks	Ambiguous	A-rate	Δ from T3 baseline (85% A)
T3 baseline	17	3	0	85%	—
C1 (Pack B retrieval)	0	20	0	0%	−85pp
C2 (post-swap Pack B retrieval)	0	20	0	0%	−85pp
C3 (setup_only)	3	17	0	15%	−70pp

The result is consistent with setup-conditioned prior suppression: when the decisive ending chapter is absent from retrieval but the shared Chapters 1–4 are present, the model's committed answers shift away from its closed-book prior. The data cannot uniquely identify the mechanism. Because Chapters 1–4 are shared across packs and do not themselves contain Pack-specific outcome signals, contributing factors may include shared-setup priming, prompt-construction effects (the harness instruction to give a best inference given retrieved sources), question-framing effects from the test battery, retrieval-state metadata if exposed to the model, or some interaction of these. Disentangling these mechanisms is left to planned stress variants (see §6).

The finding is distinct from retrieval hysteresis. It does not measure prior persistence under conflicting retrieval; it measures prior suppression under partial retrieval — a different operational condition with different implications.

§3.3 Source-gap commitment to the non-prior outcome

C3 produced a third signal distinct from §3.1 and §3.2: under conditions where the decisive source content (Chapter 5) was absent and the harness instructed the model to commit to a best inference, the model produced explicit picks that contradicted its closed-book prior — committing to the alternative pack's outcome — at a measurable rate. Because the test battery presents a binary A/B choice per fact, these are not novel-content fabrications outside both packs; they are explicit commitments to the pack outcome opposite to the T3 modal pick.

Across the 420 C3 fact-classifications, the v2 scoring yielded 76.2% `prior_aligned`, 11.4% source-gap commitment (`invent_other` in the score output), 11.4% `ambiguous` (the residual hedge floor), and 1.0% `abstain`. The 76.2% prior-aligned majority reflects the twenty facts with unanimous T3 priors: when the model lacks the decisive chapter and is asked to commit, it most often falls back on the closed-book prior. The source-gap commitment rate is the signal of interest.

Table 3.3 — C3 condition rollup (v2 classifier).

Category	n picks	Rate
`prior_aligned`	320	76.2%
source-gap commitment (`invent_other`)	48	11.4%
`ambiguous` (non-abstain)	48	11.4%
`abstain`	4	1.0%
Total	**420**	**100.0%**

Source-gap commitments were absent in S2 and concentrated in S1 and S3 — specifically S1 F2 (9 picks), S1 F6 (17 picks), and S3 F7 (17 picks), with small contributions from other facts (per-fact breakdown in Appendix A.3). S2's seven facts produced zero source-gap commitments; S1 produced 26 across two facts; S3 produced 22 across multiple facts including the F7 load-bearing diagnostic. The pattern is consistent with source-gap commitment under active setup context: when the decisive ending was absent, the model sometimes committed to the non-prior pack outcome rather than abstaining or preserving its closed-book modal pick. Because Chapters 1–4 are shared across packs, this pattern should not be interpreted as Pack-B-specific evidence unless retrieval metadata or prompt framing exposed the active pack state. The mechanism remains unresolved: possible contributors include shared setup text, prompt construction, question framing, retrieval-state metadata, and forced-inference instructions. The stress-test program isolates these contributors by varying retrieval completeness, prompt framing, abstention permission, and A↔B counterbalancing.

The methodological implication for AIP: S3-RAG distinguishes stale-pack hysteresis from source-gap commitment as separate failure modes. The two are operationally distinct: stale-pack hysteresis (§3.1's null result) measures residue from prior conversational exposure to a previously-retrieved pack; source-gap commitment measures contradiction of the closed-book prior when decisive retrieval is absent. A finding of null hysteresis does not entail safety against source-gap commitments; the failure modes are independent and would manifest under different production conditions.

---

§4. Limitations

The following limitations bound the interpretation of §3.

1. **retrieval_k=5 saturation against five-chapter packs.** With retrieval_k=5 against packs containing five chapters, every retrieval query returned the full pack. C2 therefore tested whether complete current-record retrieval overrides prior conversational exposure — a strong-retrieval-completeness condition — rather than whether partial or ambiguous retrieval produces hysteresis. Planned stress variants will test lower k against larger packs (§6).

2. **C3 `target_chunk_retrieved` predicate degeneracy.** With four chapters loaded against retrieval_k=5 in C3, top-k retrieval trivially returned all four loaded chunks; the `target_chunk_retrieved` field carried no information in C3 and was excluded from scoring (§2.5). C3 results were scored on output content rather than retrieval coverage. This is a side effect of the k=5 setting and does not affect the §3.3 source-gap-commitment finding; the stress-test program will restore a meaningful predicate by varying k against pack size.

3. **C3 abstain instrumentation artifact.** The C3 abstain rate of 1.0% (4 of 420) is an artifact of the harness system prompt's "give your best inference" instruction, which suppresses explicit abstention. The current abstain category is not interpretable as a stable model trait; the stress-test program will revise the prompt to explicitly permit "not available in retrieved sources" and rescore (§6).

4. **No A↔B counterbalance.** In all C2 sessions, the initial pack was Pack A and the swap-to pack was Pack B. The null hysteresis finding has not been verified against the reverse swap direction (B → A); confirmation that the result is not pack-valence-specific or story-trajectory-specific is a planned stress-test item (§6).

5. **Single generator.** All Phase D turns were executed against `gpt-5.4-2026-03-05`. Cross-vendor variation of the RAG probe specifically is not addressed in this sub-paper. The broader T16 program (parent paper, Jones 2026) tests $S_1/S_2/S_3$ across five substrate classes including

frontier Anthropic and frontier Google, but those probes operate on different signature targets — not the RAG-harness hysteresis target tested here. Multi-vendor extension of the RAG probe is stress-test material.

6. **Single corpus class.** The corpus comprised three custom narrative-fiction stories with structurally parallel five-chapter packs. Generalization to other domain content (technical documentation, encyclopedic content, conversational corpora) is not tested. The harness is corpus-agnostic in architecture; corpus extension is a planned stress-test item.

7. **Nested observation structure.** Fact-classifications are nested within stories, facts, and sessions. Statistical bounds reported (e.g., the rule-of-three upper bound in §3.1) are descriptive of event-level frequency at the fact-classification unit and should not be interpreted as population-level estimates for RAG systems generally.

8. **Shared setup chapters.** Chapters 1–4 are identical between Pack A and Pack B; only Chapter 5 diverges. C3 retrieval therefore loaded shared (not Pack-specific) setup content. Mechanism claims involving "retrieved setup context" cannot, on this design alone, attribute pack-direction effects to setup content; planned stress variants address this via weaker source framing and counterbalanced swap direction.

---

§5. What this finding does and does not claim

The following enumerates the specific scope of the v1.0 finding and the claims explicitly disclaimed.

§5.1 Claims

1. **Null stale-pack hysteresis under saturated current-record retrieval.** In this harness, with this model, for this corpus class, when active retrieval recovered the complete current source pack and prior conversational exposure consisted of two consolidate summaries, no detectable post-swap stale-pack residue appeared in 420 fact-classifications. Rule-of-three bounds the event-level rate at approximately 0.71% (one-sided 95% upper bound), descriptive of fact-classification frequency.

2. **Setup-conditioned prior suppression under setup-only retrieval.** When the decisive ending chapter was absent from retrieval but the shared Chapters 1–4 were present, the model's committed answers on the load-bearing dynamic-range fact shifted by 70 percentage points away from its closed-book prior. The data is consistent with setup-conditioned prior suppression; the specific mechanism (shared-setup priming vs. prompt-construction vs. question-framing vs. retrieval-state metadata) is not uniquely identified.

3. **Source-gap commitment as a separate failure class.** Under conditions where the decisive source content was absent and the harness instructed the model to commit to a best inference, the model produced explicit picks contradicting its closed-book prior — committing to the alternative pack's outcome — in 11.4% of C3 picks. This failure mode is operationally distinct from stale-pack hysteresis and would manifest under different production conditions.

§5.2 Disclaimers

1. **No general claim about RAG robustness against conversation-history bias.** The finding is bounded to the tested retrieval-completeness condition. Behavior under partial retrieval, weaker source coverage, longer pre-swap exposure, ambiguous source convergence, or memory-enabled architectures is not addressed.

2. **No claim that retrieval always dominates conversational exposure.** What the data supports is that complete current-record retrieval dominated two-turn prior exposure in this harness. The asymmetry between retrieval depth and conversation-history depth in this test does not transfer to settings where pre-swap exposure is longer, retrieval is weaker, or both.

3. **No claim that the model lacks hysteresis in general retrieval architectures.** Production RAG systems vary in chunking strategy, retrieval depth, prompt construction, and memory configuration. The null result in §3.1 applies to the specific configuration tested; other configurations may produce different results.

4. **No claim that production RAG systems are safe from stale-source contamination.** The result indicates a clean boundary condition (complete current-record retrieval; null residue), not a general safety property. Practitioners should treat the bound as a baseline against which to characterize their own systems, not as a license to assume residue-freedom.

5. **No claim that conversation history is irrelevant to retrieval-grounded generation.** The null result indicates that the tested two-turn pre-swap exposure did not produce detectable residue against saturated current-record retrieval. It does not indicate that conversation history is operationally inert; under different ratios of conversation-history to retrieval evidence, history effects may emerge.

6. **No mechanism claim for §3.2 or §3.3.** The C3 shifts (graded F7 suppression in §3.2 and binary-commitment movement in §3.3) are characterized as patterns consistent with setup-conditioned behavior. The data cannot identify whether the driving factor is shared setup text, prompt construction, question framing, retrieval-state metadata, or some interaction. Mechanism resolution is a stress-test objective.

---

§6. Stress-test roadmap

The bounded result in §3 establishes a clean-record baseline; the stress-test program extends this baseline by stress-testing the conditions that produced it. Each item addresses a specific boundary identified in §4 or surfaces a generalization question the v1.0 design did not test.

1. **Lower retrieval_k.** Test k=2 or k=3 against the existing five-chapter packs. Decouples retrieval coverage from k saturation; surfaces residue, if any, under partial current-record retrieval. Addresses §4 item 1.

2. **Larger source packs.** Extend to 8–10 chapter packs with distractor chapters so retrieval_k=5 no longer trivially returns the full pack. Establishes a non-saturated retrieval condition. Addresses §4 items 1 and 2.

3. **Longer pre-swap exposure.** Increase the C2 pre-swap consolidate count from 2 to 5+ summaries of Pack A. Tests whether the null hysteresis result holds against stronger conversational residue.

4. **Partial / changing evidence.** Test conditions where some facts are retrieved and others are not within a single test battery. Probes the boundary between §3.1 (null hysteresis when retrieval is complete) and §3.3 (source-gap commitment when retrieval is absent).

5. **Abstention-enabled C3.** Revise the harness system prompt to permit "not available in retrieved sources" as an explicit option, separating evidence-grounded answering from forced inference. Addresses §4 item 3 and surfaces a meaningful abstain rate.

6. **Weaker current-source framing.** Test whether the explicit framing of retrieved content as the authoritative current source is doing part of the work in §3.1, and whether retrieval-state metadata exposure contributes to §3.2 and §3.3 shifts. Addresses §4 item 8 and reduces the prompt-construction contribution to retrieval dominance.

7. **Same-session vs fresh-session split.** Run parallel sessions where the pack swap occurs in a fresh session (no prior conversation history) versus in the current session (per the v1.0 protocol). Distinguishes retrieval-index correctness from conversation-history residue as independent mechanisms.

8. **A↔B counterbalance.** Reverse the swap direction in half the sessions (initial Pack B → swap to Pack A) to verify the null hysteresis result is not pack-valence-specific or story-trajectory-specific. Addresses §4 item 4.

The stress-test program is designed to locate the conditions under which residue, if it exists, becomes detectable — establishing the boundary between the clean-record baseline of v1.0 and the regime where stale-pack effects emerge, and to isolate the §3.2/§3.3 mechanism contributors. Items 1–4 stress the retrieval conditions; items 5–6 stress the prompt construction; item 7 isolates the conversation-history mechanism; item 8 controls for valence.

---

§7. Discussion

This paper contributes a bounded clean-record baseline for stale-pack hysteresis under saturated current-record retrieval and a demonstration that S3-RAG can separate two distinct failure classes — stale-pack hysteresis and source-gap commitment — that are often conflated in informal RAG evaluation. The null result in §3.1 is not the absence of a finding; it is a finding about the operational conditions under which conversation-history bias is not detectable. Establishing that boundary is methodologically valuable: it tells practitioners what conditions are operationally clean and lets them characterize whether their own systems sit inside or outside that boundary.

This is also the role of AIP as a structural protocol. AIP is not designed to produce one kind of result. It is designed to keep the result attached to the condition that produced it. In one setting, that may mean identifying a failure: stale retrieval, memory contamination, unrecorded hysteresis, pseudo-redundancy, or unsupported fabrication. In another setting, it may mean establishing that a failure was not observed under a stated access tier, corpus, retrieval depth, prompt condition, and scoring rule. S3-RAG-01 v1.0 is useful for that reason. It does not dramatize the absence of stale-pack hysteresis into a general safety claim, and it does not ignore the source-gap commitments observed in C3. It separates them. Under saturated current-record retrieval, stale-pack residue was not detected. Under setup-only retrieval with abstention suppressed, the model produced commitments contradicting its closed-book prior. Those are different structural conditions and therefore different findings.

That distinction matters for production auditing. Many RAG evaluations collapse all wrong or unsupported answers into the same category of "hallucination" or "model error." AIP treats that as too coarse. A stale answer after a source swap is not the same failure as a non-prior commitment after decisive evidence is absent. One points toward retrieval hysteresis or update failure; the other points toward source-gap commitment under incomplete records and forced inference. The practical value of the protocol is that it preserves those distinctions in the ledger. A clean baseline is not a pass in the unlimited sense. A failure finding is not a condemnation in the unlimited sense. Each is a bounded record of what the system did under the conditions actually tested.

On the deployment side, the result offers bounded reassurance to production RAG practitioners. Under saturated current-record retrieval — where the retrieval layer recovers the complete relevant source content — two-turn conversation-history exposure to alternative source content did not produce detectable residue in committed answers. This applies most directly to RAG architectures where retrieval is reliable, document coverage is high relative to k, and conversation-history depth is short compared to retrieval evidence. The asymmetry between retrieval depth and history depth in this test is consistent with these conditions; results may differ under degraded retrieval, longer pre-swap exposure, or memory-enabled architectures that amplify history weight.

The stress-test program (§6) is designed to map the boundary between the clean-record baseline established here and the regime where stale-pack effects emerge. Stress variants will reduce retrieval completeness, lengthen pre-swap exposure, introduce architectural variation in the prompt construction, and isolate the mechanisms underlying §3.2 and §3.3. The combination of this v1.0 bounded null and the planned stress-tests positions S3-RAG as a maturing structural protocol — one that establishes baseline conditions, locates the boundaries of those baselines, and characterizes failure classes when boundaries are crossed.

---

§8. Conclusions

S3-RAG-01 v1.0 finds no detectable post-swap stale-pack hysteresis under saturated current-record retrieval. Across 420 fact-classifications in the swap condition, every committed pick matched the active retrieval state; zero matched the pre-swap state. The result establishes a bounded clean-record baseline for this harness, this model, and this corpus class, with rule-of-three bounding the event-level rate at approximately 0.71%.

Two additional findings emerged from the C3 setup-only condition: when the decisive ending chapter was absent from retrieval but the shared Chapters 1–4 were present, the model's committed answers shifted strongly away from its closed-book prior; and under a system prompt that suppressed abstention, the model produced source-gap commitments — explicit picks contradicting the closed-book prior — in 11.4% of C3 picks. The two phenomena are operationally distinct from each other and from the §3.1 null result, identifying source-gap commitment as a separate failure class. The mechanism underlying the C3 shifts remains unresolved and is a stress-test objective.

The important point is not that the system "passed." AIP does not need that language here. The important point is that the protocol found a boundary. With complete current records in retrieval, stale-pack hysteresis did not appear. When the decisive record was removed and inference was encouraged, the system did something different: it committed against its closed-book prior on a subset of facts, by a mechanism the v1.0 design cannot uniquely identify. Those are not the same failure mode. A structural protocol earns its keep by keeping that difference visible. S3-RAG-01 v1.0 gives AIP a clean baseline to carry forward: not a claim that RAG is safe, not a claim that conversation history cannot matter, but a record of the conditions under which current records held. The next step is to weaken those conditions and see where the boundary breaks.

Appendix A — Full per-topic per-fact data

Three per-condition sub-tables (63 rows total). Column key: A/B/? = Phase D pick counts (n=20 per cell); Abst = abstain-detected subset of `?`; T3-A / T3-B = T3 baseline counts (n=20 per cell, per Calibration v0.2 §5); Score = condition-specific score string as emitted by `analyze_phase_d.py`. The `invent` field in C3 Score strings is a legacy score-output label; in this paper it is interpreted as source-gap non-

prior commitment (per §2.5 and §3.3), not as novel content outside both packs. Source: `runs/27_phase_d/analysis/per_fact_table_v2.csv`.

Appendix A.1 — C1 (Pack B loaded, baseline retrieval)

Topic	Fact	A	B	?	Abst	T3-A	T3-B	Score
T-RAG-S1-C1	F1	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S1-C1	F2	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S1-C1	F3	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S1-C1	F4	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S1-C1	F5	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S1-C1	F6	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S1-C1	F7	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S2-C1	F1	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S2-C1	F2	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S2-C1	F3	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S2-C1	F4	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S2-C1	F5	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S2-C1	F6	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S2-C1	F7	0	20	0	0	20	0	pack_b-aligned: 20/20
T-RAG-S3-C1	F1	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S3-C1	F2	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S3-C1	F3	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S3-C1	F4	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S3-C1	F5	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S3-C1	F6	0	20	0	0	0	20	pack_b-aligned: 20/20
T-RAG-S3-C1	F7	0	20	0	0	17	3	pack_b-aligned: 20/20

Appendix A.2 — C2 (Pack A → Pack B swap; test under Pack B)

Topic	Fact	A	B	?	Abst	T3-A	T3-B	Score
T-RAG-S1-C2	F1	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S1-C2	F2	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S1-C2	F3	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S1-C2	F4	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S1-C2	F5	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S1-C2	F6	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S1-C2	F7	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S2-C2	F1	0	20	0	0	20	0	curr=20/20 other=0/20
T-RAG-S2-C2	F2	0	20	0	0	20	0	curr=20/20 other=0/20
T-RAG-S2-C2	F3	0	20	0	0	20	0	curr=20/20 other=0/20
T-RAG-S2-C2	F4	0	20	0	0	20	0	curr=20/20 other=0/20
T-RAG-S2-C2	F5	0	20	0	0	20	0	curr=20/20 other=0/20

Topic	Fact	A	B	?	Abst	T3-A	T3-B	Score
T-RAG-S2-C2	F6	0	20	0	0	20	0	curr=20/20 other=0/20
T-RAG-S2-C2	F7	0	20	0	0	20	0	curr=20/20 other=0/20
T-RAG-S3-C2	F1	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S3-C2	F2	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S3-C2	F3	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S3-C2	F4	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S3-C2	F5	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S3-C2	F6	0	20	0	0	0	20	curr=20/20 other=0/20
T-RAG-S3-C2	F7	0	20	0	0	17	3	curr=20/20 other=0/20

Appendix A.3 — C3 (setup_only — Ch1–4 loaded, no Ch5)

Topic	Fact	A	B	?	Abst	T3-A	T3-B	Score
T-RAG-S1-C3	F1	0	3	17	0	0	20	abst=0 prior=3 invent=0
T-RAG-S1-C3	F2	9	0	11	0	0	20	abst=0 prior=0 invent=9
T-RAG-S1-C3	F3	0	20	0	0	0	20	abst=0 prior=20 invent=0
T-RAG-S1-C3	F4	0	20	0	0	0	20	abst=0 prior=20 invent=0
T-RAG-S1-C3	F5	0	20	0	0	0	20	abst=0 prior=20 invent=0
T-RAG-S1-C3	F6	17	0	3	0	0	20	abst=0 prior=0 invent=17
T-RAG-S1-C3	F7	0	19	1	0	0	20	abst=0 prior=19 invent=0
T-RAG-S2-C3	F1	20	0	0	0	20	0	abst=0 prior=20 invent=0
T-RAG-S2-C3	F2	20	0	0	0	20	0	abst=0 prior=20 invent=0
T-RAG-S2-C3	F3	20	0	0	0	20	0	abst=0 prior=20 invent=0
T-RAG-S2-C3	F4	19	0	1	0	20	0	abst=0 prior=19 invent=0
T-RAG-S2-C3	F5	17	0	3	0	20	0	abst=0 prior=17 invent=0
T-RAG-S2-C3	F6	20	0	0	0	20	0	abst=0 prior=20 invent=0
T-RAG-S2-C3	F7	20	0	0	0	20	0	abst=0 prior=20 invent=0
T-RAG-S3-C3	F1	1	19	0	0	0	20	abst=0 prior=19 invent=1
T-RAG-S3-C3	F2	0	19	1	0	0	20	abst=0 prior=19 invent=0
T-RAG-S3-C3	F3	2	16	2	0	0	20	abst=0 prior=16 invent=2
T-RAG-S3-C3	F4	1	12	7	3	0	20	abst=3 prior=12 invent=1
T-RAG-S3-C3	F5	1	13	6	1	0	20	abst=1 prior=13 invent=1
T-RAG-S3-C3	F6	0	20	0	0	0	20	abst=0 prior=20 invent=0
T-RAG-S3-C3	F7	3	17	0	0	17	3	abst=0 prior=3 invent=17 (split-prior)

---

§9. End matter

References

Internal documents:

- Design doc: `scope/UCT_T16_S3RAG_Design_v0_1_2026_05.md`

- Calibration doc: `scope/UCT_T16_S3RAG_Calibration_v0_2_2026_05.md`
- Phase C smoke findings: `scope/UCT_T16_S3RAG_Phase_C_Smoke_Findings_v0_1_2026_05.md`
- Harness sketch: `scope/UCT_T16_S3RAG_Harness_Sketch_v0_1_2026_05.md`
- Probe specification: `probes/s3_rag_01_retrieval_hysteresis.yaml` (v0.2)
- Classifier v2 notes: `runs/27_phase_d/analysis/classifier_v2_notes.md`
- Per-fact data: `runs/27_phase_d/analysis/per_fact_table_v2.csv`
- AI Empirical Demo (T16 parent): `UCT_T16_AI_Empirical_Demo_v1_0_2026_05.docx`
- AIP v1.0: `AIP_v1_0_2026_05.docx`

Parent program reference:

- Jones, J. C. (2026). *Empirical Demonstration of S_1, S_2, and S_3 in Deployed AI Systems: A Tier C Application of UCT's Portable Structural Signatures to Five AI Substrates* (UCT T1.6). HoldingLight LLC. DOI: 10.17605/OSF.IO/JPXCU.

Sibling UCT empirical demonstrations:

- Jones, J. C. (2025). *Rice Hysteresis.* HoldingLight LLC. DOI: 10.17605/OSF.IO/KZ8TP
- Jones, J. C. (2026). *COGITATE iEEG Reanalysis.* HoldingLight LLC. DOI: 10.17605/OSF.IO/MXYU2

Related standards-layer papers:

- Jones, J. C. (2026). *Records Across Nature, Life, and Mind.* HoldingLight LLC. DOI: 10.17605/OSF.IO/7H6DY
- Jones, J. C. (2026). *Update Integrity Standard.* HoldingLight LLC. DOI: 10.17605/OSF.IO/DWM29

UCT library note

This paper is part of the Universal Collapse Theory library, published by HoldingLight LLC. It is the slot-4 sub-demonstration within the broader T16 AI Empirical Demonstration (Jones 2026, DOI 10.17605/OSF.IO/JPXCU), which sits alongside the Rice Hysteresis and COGITATE iEEG Reanalysis demonstrations in the T1.6 empirical corpus. For a reading guide and full architecture, visit universalcollapse.com/roadmap.

AI Disclosure

AI tools were used to assist with manuscript preparation, drafting, organization, and editorial refinement. The underlying theory, structural decisions, analysis, and conclusions are the author's own.

Citation

Jones, J. C. (2026). *S3-RAG-01 Phase D findings: Bounded null-hysteresis under saturated current-record retrieval.* HoldingLight LLC. DOI: 10.17605/OSF.IO/5QMVS

End of v1.0.