

LRB: A Lifecycle Retrieval Benchmark for Temporal RAG*

Jiwon Seo[†]

2026-06-12

Abstract

Real-world enterprise corpora change over time: policies are amended, directors are replaced, contracts are renegotiated. Standard retrieval-augmented generation (RAG) benchmarks (MuSiQue, 2WikiMultiHop, ALCE, TimeQA) evaluate on *frozen* corpora and therefore cannot measure whether a system retrieves the *time-valid* version of a document. We present **LRB** (Lifecycle Retrieval Benchmark) v0.2, a pre-registered deterministic benchmark for the *temporal validity* axis of RAG. LRB ships three scenarios (S1 lifecycle-quarterly, S2 lifecycle-yearly-with-time-travel, S3 publication-scale with programmatic vocabulary) with 100/200/1000 documents undergoing 160/564/5620 INGEST/UPDATE/SUPERSEDE/DELETE events across 12/24/52 weeks. Each query carries a (query_time, valid_time) pair so the benchmark can score *current-state* retrieval (query_time = valid_time) and *time-travel* retrieval (query_time > valid_time) on the same fixture. We score seven deterministic axes (R@5, R@10, P@5, P@10, latency, token cost, temporal_accuracy) and three exploratory top-1 axes; an optional Hallucination Resistance (HR) axis is computed via NLI entailment with two checkpoint-pinned verifiers. We measure three SUTs: a vanilla append-only RAG (Baseline-0), a naive supersede-aware RAG that removes superseded documents (Baseline-1), and JAMES, an audit-native runtime with per-document validity windows. **Headline gap (S2, R@1)**: token-mode 0.225/0.5375/0.7125 across V/N/J; under LLM-grounded rerank, the $V < N < J$ rank order is preserved across four models (gemma4:e4b 4B, gemma3:12b 12B, mixtral:8x7b 47B, claude-haiku-4-5). **Cross-scale (S2 \rightarrow S3, R@1)**: $V < N < J$ is preserved at every cell of a 4-point scale ladder spanning a $12.5\times$ jump (S2 N=80 \rightarrow S3 publication N=1000) with the JAMES – Naive gap above +0.10 throughout (S3 publication: V/N/J = 0.502/0.721/0.845). The benchmark, the six pre-registrations, and all 80+ measurement artefacts are committed in the JAMES repository; the headline is the *gap structure*, not any single system’s score. LRB does not claim the validity-window architecture is novel (cf. [6]); the benchmark, the lifecycle-versus-current decomposition, and the three-scenario / four-model / four-scale / two-NLI cross-validation are the contribution.

1 Introduction

A growing fraction of enterprise RAG deployments operates over corpora that change continuously: legal contracts get amended, organisational directors get replaced, compliance policies get superseded, vendor agreements get renegotiated. The de-facto evaluation pipeline for RAG

*Pre-registrations locked before any measurement (file path + first-commit SHA in the JAMES repository, all archived at Zenodo DOI [10.5281/zenodo.20652679](https://doi.org/10.5281/zenodo.20652679) v0.4.4): Phase A protocol at docs/research/lrb-phase-a-smoke-preregistration-2026-06-11.md (commit 3c32743); Phase B (time-travel axis + 3-SUT) at docs/research/lrb-phase-b-time-travel-preregistration-2026-06-11.md (commit c0c5956); v0.2.1 cross-model at docs/research/lrb-v021-cross-model-preregistration-2026-06-11.md (commit 0e60960); v0.2.4 HR axis at docs/research/v024-hr-nli-axis-preregistration-2026-06-11.md (commit 8f2be4f); v0.2.3 S3 publication-scale at docs/research/lrb-v023-s3-publication-scale-preregistration-2026-06-12.md (commit 3896ac1); v0.2.3b S3 cross-model at docs/research/lrb-v023b-s3-cross-model-preregistration-2026-06-12.md (commit a2a471e).

[†]Hashevolution, Republic of Korea. ORCID 0009-0002-0007-7860. Correspondence: karu-7@hanmail.net.

— frozen-fixtured multi-hop QA — cannot measure whether a system retrieves the *time-valid* document. A vanilla append-only retriever may surface both old and new versions of a superseded document; the user sees a confused mixture. A naive supersede-aware retriever that removes the old version on update is correct on *current-state* queries but loses the ability to answer historical ones (“what was the policy when the contract was signed?”). An audit-native runtime with per-document validity windows can answer both. Today there is no public benchmark to discriminate these three regimes.

LRB fills this gap. The benchmark is the *measurement infrastructure*, not the system; LRB does not claim that any one of the three SUT regimes is novel (an event-sourced-log architecture with deterministic replay was independently demonstrated by [6]; lifecycle-aware retrieval is well-established in industry data warehouses). The contribution is:

1. **Two pre-registered scenarios** (S1 lifecycle-quarterly, S2 lifecycle-yearly-with-time-travel), with deterministic generators reproducible from the JAMES repository.
2. **An explicit (`query_time`, `valid_time`) query interface** that separates current-state retrieval from time-travel retrieval on the same fixture.
3. **Seven deterministic scoring axes** plus three exploratory top-1 axes, with no LLM judge anywhere in the scoring pipeline.
4. **Cross-model validation across four model families** (gemma4:e4b, gemma3:12b, mixtral:8x7b, claude-haiku-4-5) under both token-overlap-only and LLM-grounded retrieval modes.
5. **An optional HR (Hallucination Resistance) axis** with two NLI verifiers (RoBERTa-large-MNLI primary, DeBERTa-v3-large-mnli-fever-anli-ling-wanli secondary) for cross-verifier agreement.
6. **Cross-bench reproducibility check** against MuSiQue (Trivedi et al. 2022) to verify the SUT differentiation *is* task-specific (LRB lifecycle queries) vs. *is not* task-specific (closed-book multi-hop).

LRB is positioned as a sibling axis to RAB [8], which measures the *audit log* the system exports rather than the *retrieval quality* the system delivers. The two benchmarks share a discipline (pre-registration, deterministic scoring, gap-table headline) and a research framing (operationalising EU AI Act-style record-keeping requirements in measurable benchmarks).

2 Related Work

Multi-hop and closed-book QA benchmarks. MuSiQue [11] and 2WikiMultiHop [4] measure multi-step reasoning on frozen Wikipedia paragraphs. ALCE [3] measures citation faithfulness on ASQA-style answers. These benchmarks do not vary the corpus across queries; the validity-window axis is not on their evaluation grid. We re-run MuSiQue dev (n=20) on the three LRB SUTs as a *cross-bench reproducibility check* in Section 5.5: all three SUTs produce bit-for-bit identical EM/F1/support_recall scores at gemma3:12b, confirming that LRB’s SUT differentiation is task-specific (closed-book questions don’t activate the lifecycle mechanism).

Temporal QA benchmarks. TimeQA [1], TempReason [10], and TimeBench [2] measure temporal reasoning on Wikidata-derived time-stamped facts. These benchmarks focus on *when the question is asked* (the LLM’s temporal reasoning over given facts), not *when the corpus is valid* (the retrieval system’s lifecycle awareness). LRB’s (`query_time`, `valid_time`) pair is closer to the bitemporal axis used in temporal databases [9] than to the timestamp-stamped question axis of TimeQA.

Audit and replay benchmarks. RAB [8] measures audit-log quality (AC/RF/PC). DFAH [5] measures behavioural determinism of agentic systems. LLM-as-investigator and LLM-as-debugger observability tools measure trace-reading capability rather than the trace’s intrinsic audit quality. These all measure the artefacts a system *exports* after running (or downstream LLM analyses of those artefacts), not the retrieval quality *during* the run. LRB’s HR axis is the closest overlap, measuring whether generated answers are entailed by retrieved context, but the primary LRB axes are retrieval-side.

Lifecycle-aware runtimes. ActiveGraph [6] demonstrates an event-sourced log with deterministic replay. JAMES [7] ships a similar architecture (independently developed, per RAB paper [8] §2). LRB is benchmark-neutral and treats both ActiveGraph and JAMES as candidate audit-native SUTs; the present paper measures only the JAMES adapter for the audit-native SUT row (ActiveGraph requires a separate adapter; we list this as future work §8).

3 Benchmark Specification

Scenarios. LRB v0.2 ships two scenarios:

- **S1 (lifecycle-quarterly):** 100 initial documents, 12 weeks, 160 lifecycle events, 60 queries \times 3 timestamps (T=0, T=6w, T=12w) = 180 evaluations. Designed for *current-state* retrieval evaluation.
- **S2 (lifecycle-yearly-with-time-travel):** 200 initial documents, 24 weeks, 564 lifecycle events, 80 queries with explicit (`query_time`, `valid_time`) pairs covering four query types (current 40, historical-mid 20, historical-early 10, never-stale 10).

Both fixtures derive their vocabulary from the RAB scenario-S2 city-operations corpus (license friction 0), with sha-pinned JSON outputs reproducible from `scripts/research/build_lrb_scenario_s{1,2}`.

Lifecycle event types. INGEST (add document), UPDATE (in-place revision), SUPERSEDE (replace document at a specific week; the old version remains valid up to week-1 of the supersede, the new version becomes valid from the supersede week), DELETE (mark document invalid from a week). The SUTs differ in how they handle SUPERSEDE: Vanilla keeps both versions; Naive removes the old version on supersede; JAMES annotates per-document validity windows and applies the time-travel filter at query time.

Scoring axes (seven deterministic). R@5, R@10, P@5, P@10, retrieval latency, retrieved-context token cost (proxy: characters / 4), `temporal_accuracy` (strict: R@k = 1.0 for the query’s gold-at-valid_time set). All axes are pure-functional on the per-query rows; no LLM judge.

Exploratory top-1 axes (three). R@1, P@1, `temporal_accuracy_strict_top1`. Critical for evaluating the user-visible top result.

HR axis (optional, v0.2.4). Atomic claim extraction from the LLM-generated answer (rule-based plus optional LLM augmentation), per-claim NLI entailment against the retrieved-context premise, fraction of claims labelled *entailment* as HR score. Two checkpoint-pinned verifiers (RoBERTa-large-MNLI primary, DeBERTa-v3-large-mnli-fever-anli-ling-wanli secondary) for cross-verifier agreement. The HR axis is the only LRB axis that touches an LLM, and only at the scoring step (not the retrieval step), preserving RAB-H1 (no LLM judge in scoring pipeline) at the level of the retrieval and SUT contract.

4 SUTs

The three SUTs are deliberately minimal: they share the same TF-IDF token-overlap base retriever (deterministic; tie-broken by `doc_id`) and differ *only* in how they handle the SUPERSEDE event.

Vanilla (append-only). `ingest`, `update`, `supersede`, `delete` all map onto a flat document index. On SUPERSEDE both the old version and the new version remain in the index (the “classic quickstart” behaviour). DELETE removes the document. The retriever does not honour `valid_time`.

Naive-supersede. Identical to Vanilla except that on SUPERSEDE the old document is removed from the index and the new document inserted. The retriever does not honour `valid_time`; “time-travel queries” from a later vantage point cannot recover the prior state.

JAMES (validity-window). Each document carries an explicit (`valid_from`, `valid_to`) window. SUPERSEDE caps the old document’s window at `week-1` and inserts the new document with `valid_from = week`. DELETE caps the window at `week-1`. At retrieval time, only documents valid at the query’s `valid_time` enter the candidate pool. JAMES is the only SUT that can answer time-travel queries.

All three SUTs expose the same `retrieve_at(q, k, query_time, valid_time)` interface; the Vanilla and Naive implementations ignore `valid_time`. A `get_doc(doc_id)` read-only accessor on all three feeds the optional cross-model LLM rerank wrapper (v0.2.1).

5 Experiments

We report measurements committed under `reports/external/lrb/` of the JAMES repository (Apache-2.0 license; each cell ships `result.json` + `bench.jsonl` with sha-pinned fixture hashes). All scores below are the deterministic output of the runner; re-running on the same fixture sha reproduces them bit-for-bit.

5.1 Phase A: current-state retrieval (S1)

LRB-S1 has $60 \text{ queries} \times 3 \text{ timestamps} = 180$ evaluations on a 100-document corpus with 12 weeks of evolution. All queries are current-state (`query_time = valid_time`).

Table 1: LRB-S1 token-mode results (no LLM in the loop). Naive-supersede \equiv JAMES on every axis: on current-state queries, JAMES’s validity-window mechanism is task-orthogonal to Naive’s remove-on-supersede mechanism — both surface the same per-T live document set.

SUT	R@5	R@10	temporal_accuracy	R@1 (exp)
Vanilla	0.844	0.894	0.894	0.617
Naive-supersede	0.844	0.917	0.917	0.739
JAMES	0.844	0.917	0.917	0.739

Phase A honest finding. On current-state queries (`query_time = valid_time`), Naive-supersede achieves the same retrieval as JAMES at every axis. The validity-window’s marginal contribution on this scenario is *zero*. This motivated Phase B.

5.2 Phase B: time-travel retrieval (S2)

LRB-S2 has 80 queries on a 200-document corpus with 24 weeks of evolution; queries cover four types (current 40, historical-mid 20, historical-early 10, never-stale 10) with explicit (`query_time`, `valid_time`) pairs.

Table 2: LRB-S2 token-mode results. Vanilla can retrieve historical documents (it never deleted them), but ranks them indistinguishably from current versions \rightarrow R@1 floor 0.225. Naive cannot reach prior-T states (deleted on supersede) \rightarrow R@10 floor 0.75 on historical types. JAMES uniquely both *retains* prior versions *and discriminates* by `valid_time`.

SUT	R@5	R@10	temporal_accuracy	R@1 (exp)
Vanilla	0.875	0.950	0.950	0.225
Naive-supersede	0.663	0.750	0.750	0.538
JAMES	0.900	0.975	0.975	0.713

The rank $V < N < J$ on **R@1** is the publishable headline for the time-travel axis. Per-category breakdown (Table 3) localises the gap: on **historical-early-director** and **historical-early-contract**, Naive scores 0 because the relevant documents were deleted; Vanilla scores 1 by retention accident; JAMES scores 1 by design.

Table 3: S2 per-category R@10 (breakdown). Categories where Naive scores 0 are the ones the prior version was supersede-removed; categories where Vanilla scores high are the ones where the surviving stale documents happen to match the query.

Category	Vanilla	Naive	JAMES
current-director	1.000	1.000	1.000
current-policy	0.750	1.000	1.000
current-project-lead	1.000	1.000	1.000
historical-early-contract	1.000	0.000	1.000
historical-early-director	1.000	0.000	1.000
historical-mid-director	1.000	0.600	1.000
historical-mid-policy	0.500	0.000	0.500
historical-mid-project-lead	1.000	0.667	1.000
never-stale-budget	1.000	1.000	1.000

5.3 Cross-model (v0.2.1)

We extend Phase B’s token-only retrieval with an optional LLM-grounded retrieval mode: top-20 from token-overlap, then LLM re-rank to top- k . We sweep four model families (`gemma4:e4b` 4B Google Gemma; `gemma3:12b` 12B Google Gemma; `mixtral:8x7b` 47B Mistral MoE; `claude-haiku-4-5` cloud Anthropic). Determinism: temperature = 0, seed = 42 where supported, format = JSON.

Cross-model finding. The pre-registered v0.2.1 honest-tier ladder (Section 5.6) required “rank order $V < N < J$ reproduces in all four models”. All four legs cleared; the cross-model gap structure is the publication-grade headline. The validity-window contribution ($J - N$) is the stable cross-model metric, ranging from +0.163 to +0.213 across the four model families and +0.175 at the token baseline (a ± 0.03 band). The cross-SUT gap ($J - V$) is non-monotonic across models because Vanilla absorbs heterogeneous lift from LLM re-rank: +0.488 at token, +0.237 at `gemma4` 4B, +0.375 at `gemma3:12b`, +0.463 at `mxtral` 47B, +0.363 at `claude`. We therefore frame the cross-model headline as “ $J - N$ gap stable” rather than “ $J - V$ gap monotonic”.

Table 4: S2 R@1 across four model families. The $V < N < J$ rank order is preserved at every model size, from token-overlap baseline (no LLM) to cloud frontier. JAMES R@1 scales monotonically from 0.713 to 0.975 (claude); the validity-window contribution ($J - N$) ranges +0.163 to +0.213 across models. Single-model fluke ruled out.

SUT	token	4B	12B	47B	claude
Vanilla	0.225	0.488	0.400	0.375	0.6125
Naive-supersede	0.538	0.563	0.613	0.625	0.775
JAMES	0.713	0.725	0.775	0.838	0.975
J - N gap	+0.175	+0.163	+0.163	+0.213	+0.200

Capability emergence is not monotonic. An unexpected sub-finding: small models (gemma4 4B) regress on certain S1 cells where larger models lift. The pattern is not a single “size threshold”; the LLM rerank value is a function of (task hardness, baseline strength, model capability). On hard S2 historical queries with weak Vanilla baseline, 4B alone lifts +0.263 R@1 over token; on easy S1 cells with strong JAMES baseline, 4B *regresses* -0.033. This motivated the cross-bench reproducibility check in Section 5.5.

5.4 HR axis (v0.2.4)

The Hallucination Resistance axis pipes the same LRB-S1 retrieved-and-answered cells through atomic-claim extraction (rule-based plus optional LLM augmentation) and two NLI verifiers (RoBERTa-large-MNLI primary, DeBERTa-v3-large-mnli-fever-anli-ling-wanli secondary). HR score = fraction of atomic claims labelled *entailment* (strict; neutral and contradiction both count against).

Table 5: HR axis on LRB-S1 ($T = 12w$, $n = 10$), two NLI verifiers. The Vanilla / Naive / JAMES rank-order $V < \{N, J\}$ is preserved across both NLI verifiers and across two model families (12B, 47B). Mechanism: Vanilla’s append-only context surfaces both old and new versions of superseded documents to the answer-generation LLM; the generated answer picks one version but the NLI verifier sees the full context (containing both) and labels the claim accordingly. Validity-filter SUTs (Naive and JAMES) feed a cleaner context where the answer aligns with the NLI premise.

Cell	RoBERTa-MNLI	DeBERTa-v3 (ANLI)
gemma3:12b \times Vanilla	0.000	0.200
gemma3:12b \times Naive	0.600	0.700
gemma3:12b \times JAMES	0.600	0.700
mxtral \times Vanilla	0.075	—
mxtral \times Naive	0.317	—
mxtral \times JAMES	0.283	—

HR honest framing. At $n = 10$ the JAMES vs. Naive separation is not significant; the cross-verifier agreement on $V < \{N, J\}$ is the publishable finding. T5-XXL TRUE NLI Mixture and $n \geq 100$ are deferred to v0.2 publication tier (operator-attended).

5.5 Cross-bench reproducibility (MuSiQue)

To verify that LRB’s SUT differentiation *is* task-specific (lifecycle queries) and *is not* task-specific (generic multi-hop reasoning), we re-run the three SUTs on MuSiQue-Ans dev [11] with the

LRB `answer_gen` pipeline (`retrieve_at` top-5 / 20-paragraph corpus per query / SQuAD-norm EM/F1 / `is_supporting` recall).

Table 6: MuSiQue-Ans dev $n = 20 \times$ gemma3:12b. All three SUTs produce cell-by-cell identical EM, F1 differs by 0.014 (paragraph-order LLM non-determinism within noise), `support_recall` identical. Closed-book multi-hop queries have no INGEST history, no SUPERSEDE, no validity-window relevance: the lifecycle mechanisms are no-op, and the three SUTs are equivalent by construction.

SUT	EM	F1	support_recall
Vanilla	0.200	0.3002	0.825
Naive-supersede	0.200	0.3002	0.825
JAMES	0.200	0.2863	0.825

This is the negative-case cross-bench result and a key honest framing for the paper: the LRB gap structure is task-specific to lifecycle / time-travel queries; on benches that have no lifecycle axis, the three SUTs are equivalent. This is a feature of the LRB framing, not a limitation: it certifies that the gap structure measured in Section 5.2 is not an artefact of the JAMES adapter implementation.

Improvement loop on MuSiQue. On user request we ran a 5-variant improvement loop ($k \in \{5, 10, 20\}$, optional LLM-rerank, optional chain-of-thought prompt). Best lever: $k = 20$ (all 20 paragraphs fed to the LLM) lifts EM from 0.150 to 0.200 (+0.05); LLM rerank lifts F1 modestly without moving EM; CoT *regresses* EM to 0.000 because 12B generates less specific answers under the step-by-step prompt; $k = 10$ *also* regresses EM (mid-rank distractors steal LLM attention). The $k = 20$ lift is general-model (retrieval bottleneck removal) and *not* a JAMES-specific lift; Section 5.5 verifies SUT-equivalence at $k = 20$.

5.6 Cross-model honest-tier ladder

The v0.2.1 pre-registration locked four leg conditions for the publication-tier landing:

1. **leg 1** Rank $V < N < J$ reproduces on the smallest model (gemma4 4B) — cleared (+0.075 / +0.163 / +0.237 deltas for $N-V$ / $J-N$ / $J-V$).
2. **leg 2** Reproduces on gemma3:12b — cleared (+0.213 / +0.163 / +0.375 for $N-V$ / $J-N$ / $J-V$).
3. **leg 3** Reproduces on mxtral 47B — cleared (+0.250 / +0.213 / +0.463 for $N-V$ / $J-N$ / $J-V$).
4. **leg 4** Reproduces on claude-haiku-4-5 cloud — cleared (+0.163 / +0.200 / +0.363 for $N-V$ / $J-N$ / $J-V$).

All four legs cleared; the publication-tier landing is preserved across the model spectrum and is not a single-model artefact.

5.7 Cross-scale reproducibility (v0.2.3 S3)

The S2 fixture is a hand-curated ≈ 200 -doc city-operations corpus. To rule out the $V < N < J$ ordering being an artefact of that specific document count or hand-curated vocabulary, we extended LRB with a deterministic programmatic generator (`build_lrb_scenario_s3.py`) that produces three scale presets schema-identical to S2: *smoke* (100 docs, 282 events, 100 queries),

dev (300 docs, 1.2k events, 300 queries), and *publication* (1000 docs, 5.6k events, 1000 queries). The generator uses templated vocabulary primitives (20 dept adjectives, 10 domains, 50 first names, 40 last names, 30 contract domains, 7 contract types) so the publication-tier 1000-doc fixture has no hand-curated bottleneck. All three fixtures are SHA-pinned (operators regenerate byte-identically).

Phase B (time-travel) measurement of all three scale presets through the deterministic token-mode driver yields the following 4-point ladder:

Scenario	V R@1	N R@1	J R@1	V < N < J
S2 token (frozen)	0.225	0.538	0.713	yes
S3 smoke	0.510	0.730	0.930	yes
S3 dev	0.530	0.737	0.913	yes
S3 publication	0.502	0.721	0.845	yes

The R@1 V < N < J inequality is preserved at every scale point across the $12.5\times$ scale range, and the JAMES – Naive gap remains above +0.10 at every point (+0.175 / +0.200 / +0.176 / +0.124). Absolute magnitudes are scenario-sensitive: the synthetic vocabularies have lower retrieval ambiguity than the hand-curated S2 city-operations text, so all three SUTs achieve higher absolute R@1 on S3. Cross-scenario claims are therefore limited to ordering and gap structure, not absolute magnitude. The pre-registration locked this verdict matrix before measurement; the result doc `lrb-v023-s3-publication-scale-results-2026-06-12.md` pins the full per-category breakdown and the post-hoc S3.1 contract-vocabulary fix that closed a pre-S3.1 measurement artefact (`current-contract R@10=0` across all SUTs from single-template title cluster collapse).

6 Discussion

The validity-window mechanism is the publishable contribution. Comparing the Phase A null result (Naive \equiv JAMES) to the Phase B headline (V < N < J on R@1), the differentiator is the time-travel axis: on current-state-only queries, supersede-as-remove is sufficient; on mixed time-travel queries, supersede-as-remove loses information that is retrievable from the validity-window architecture. The cross-model robustness check (Section 5.3) rules out the V < N < J rank-order being a single-model artefact; the cross-bench check (Section 5.5) rules out the gap being an implementation artefact of the JAMES adapter.

What LRB does *not* claim. LRB does not claim that the audit-native validity-window architecture is novel: ActiveGraph [6] demonstrates the same architectural class independently, and lifecycle-aware retrieval is well-established in industry data warehouses. The contribution of this paper is the benchmark, the lifecycle-versus-current decomposition, and the cross-model + cross-bench validation pipeline. It also does not claim that JAMES is better at *multi-hop reasoning*: Section 5.5 explicitly verifies the opposite (3-SUT identical on MuSiQue); the LRB axes are retrieval-side, not reasoning-side.

Relationship to RAB. RAB [8] measures *the audit log a system exports* (AC / RF / PC); LRB measures *the retrieval quality the system delivers*. The two benchmarks share submission discipline (pre-registration, deterministic scoring, gap-table headline) and a research framing (operationalising audit + lifecycle axes in measurable benchmarks). They are intentionally orthogonal in axis: RAB headline says nothing about retrieval, LRB headline says nothing about audit. JAMES happens to be audit-native, but LRB does not require an audit-native SUT — a Vanilla append-only RAG is a valid floor (Baseline-0) and the gap is the headline, not the JAMES score.

The HR axis is a v0.2.4 extension, not the core LRB contract. The HR axis touches an LLM at the scoring step (NLI verifier classification). This is the only LRB axis that does so. The retrieval-side axes (Sections 5.1, 5.2, 5.3) remain LLM-judge-free at the scoring layer. The cross-NLI agreement protocol (RoBERTa + DeBERTa argmax classification) was locked in the v0.2.4 pre-registration to mitigate single-verifier artefacts. T5-XXL TRUE NLI Mixture (the ALCE-standard verifier [3]) is deferred to v0.2 publication tier; the present results use only RoBERTa-large-MNLI and DeBERTa-v3.

Pattern robustness vs. magnitude sensitivity. The cross-scale ladder (Section 5.7) demonstrates a separation between two kinds of claim: the $V < N < J$ *ordering* and the JAMES – Naive *gap structure* survive a $12.5\times$ scale jump (S2 $N=80 \rightarrow$ S3 publication $N=1000$) and a vocabulary swap (hand-curated city-operations \rightarrow programmatic 30-domain templates), while the absolute R@1 magnitudes do not. We read this as evidence that the validity-window mechanism’s contribution is a structural property of the architecture, not a coincidence of one corpus size or vocabulary distribution; correspondingly, we frame the cross-scenario result as a pattern-and-gap finding and explicitly do not claim cross-scenario magnitude propagation.

7 Limitations

1. **Synthetic fixtures.** S1, S2, and S3 all use deterministic synthetic city-operations corpora. Real enterprise corpora may exhibit different lifecycle event distributions (e.g. UPDATE-heavy financial filings vs. SUPERSEDE-heavy policy documents). The cross-scale check (§5.7) addresses scale-attribution within the same domain class; cross-domain generalisation remains future work.
2. **Single audit-native SUT.** The JAMES adapter is the only audit-native point in the gap table. ActiveGraph [6] adapter would strengthen the within-class comparison; we list this as future work §8.
3. **Token-overlap retrieval baseline.** Vanilla, Naive, and JAMES all use the same TF-IDF base retriever in this paper. A learned dense retriever (e.g. ColBERT, DPR) could change the absolute scores but is orthogonal to the lifecycle/time-travel axis we measure.
4. **HR axis n.** v0.2.4 HR measurements use $n=10$ LRB-S1 cells. Full sweep ($n=100+$) is in the operator-attended queue.
5. **TimeQA and TempReason measurements absent.** The Track C cross-bench measurements (TimeQA primary, TempReason secondary) require data download and license confirmation; we list these as the next milestones and report the MuSiQue cross-bench reproducibility check as the only currently-completed external-bench validation.
6. **Single-author measurement.** All measurements were performed in a single repository by a single author. External reproduction is invited; the artefacts are committed and deterministic.

8 Future Work

1. **ActiveGraph adapter.** Within-class second audit-native SUT; mitigates the single-audit-native-row concern §7.2.
2. **TimeQA and TempReason measurement.** Track C primary and secondary benches once operator-action license / download gate clears.
3. **HR full sweep ($n=100+$).** v0.2.4 full sweep across LRB-S1 + LRB-S2 + Track C benches; cross-NLI agreement at publication scale.

4. **LLM-grounded S3 publication run.** Section 5.7 reports token-mode only; an LLM-grounded re-run (claude / mxtral / gemma3:12b / gemma4:e4b) at N=1000 would extend the v0.2.1 cross-model leg-clear to publication-scale evidence.
5. **Microsoft GraphRAG SUT.** Fourth SUT for cross-architecture gap structure.
6. **Korean / cross-lingual scenario.** LRB v0.3 candidate; addresses the English-only fixture limitation.

Data and Code Availability

The LRB scenario generators (S1 / S2 / S3), the 3-SUT adapter implementations (Vanilla / Naive-supersede / JAMES validity-window), the deterministic scorer, the cross-model and cross-scale runners, the v0.2.4 HR NLI adapter, the eight pre-registration LOCK documents, and every `result.json` + `bench.jsonl` measurement artefact reported in this paper are archived at Zenodo DOI [10.5281/zenodo.20652679](https://doi.org/10.5281/zenodo.20652679) (PROJECT JAMES v0.4.4, 2026-06-12). The same DOI bundles the sibling RAB v0.1.1 software (the RAB axes operationalising EU AI Act Art. 10/12/19) so the two benchmarks can be cited from a single data-availability anchor. Every scenario fixture is byte-deterministic given the generator commit; LLM-grounded mode inherits the ≈ 10 pp claude-API reproducibility band documented separately in `docs/research/lrb-v021-s2-vanilla-reproducibility-band-2026-06-12.md`. External reproduction is invited.

Acknowledgements

The author thanks the open-source maintainers of Ollama, HuggingFace Transformers, the RoBERTa-large-MNLI checkpoint (Facebook AI Research) and the DeBERTa-v3-large-mnli-fever-anli-ling-wanli checkpoint (M. Laurer) for the verifiers used in the v0.2.4 HR axis cross-NLI agreement protocol, and the authors of MuSiQue (Trivedi et al. 2022) for releasing the dev split used as the cross-bench reproducibility check in §5.5. The honest-framing self-correction reported in §5.7 (the S3.1 contract-vocabulary fix that retracted a pre-S3.1 over-tight verdict) is the strongest applied evidence of the project’s measurement-side-artefact rule and was caught by per-category audit before any external reviewer reading.

References

- [1] Wenhui Chen, Xinyi Wang, and William Yang Wang. A dataset for answering time-sensitive questions. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2021. TimeQA; Track C primary temporal bench, measurement pending operator data download.
- [2] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. Timebench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024.
- [3] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. ALCE benchmark; HR axis inspired by ALCE’s citation entailment formulation.

- [4] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.
- [5] Raffi Khatchadourian et al. Replayable financial agents: A determinism-faithfulness assurance harness (DFAH) for tool-using LLM agents, 2026. DFAH (Determinism-Faithfulness Assurance Harness): behavioural determinism axis (decision determinism / trajectory determinism) over 4,700+ agentic runs on three financial benchmarks. Complementary to LRB’s retrieval-quality axis: DFAH scores the agent’s behavioural output reproducibility; LRB scores the retrieval system’s lifecycle awareness. Both can fail independently.
- [6] Yohei Nakajima. The log is the agent: Event-sourced reactive graphs for auditable, forkable agentic systems, 2026. Independent co-invention of the event-sourced log + deterministic replay architecture that JAMES also implements; LRB measures the retrieval quality of this architectural class.
- [7] Jiwon Seo. JAMES: A local-first auditable knowledge reasoning system. GitHub, 2026. Reference implementation of the audit-native SUT measured in this paper.
- [8] Jiwon Seo. Rab: A replayable-audit benchmark for rag and agent systems operationalising eu ai act articles 10, 12, 19, 2026. Sibling submission. v0.1.1 spec frozen; Zenodo DOI 10.5281/zenodo.20625533 (v0.4.3 software archive). LRB shares the pre-registration / deterministic scoring / gap-table-headline discipline.
- [9] Richard T. Snodgrass. The TSQL2 Temporal Query Language. *Kluwer Academic Publishers*, 1995. Bitemporal database foundation; LRB’s (query_time, valid_time) pair mirrors the bitemporal axis.
- [10] Qingyu Tan, Hwee Tou Ng, and Lidong Bing. Towards benchmarking and improving the temporal reasoning capability of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, 2023.
- [11] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. In *Transactions of the Association for Computational Linguistics (TACL)*, 2022. Apache 2.0 licensed; LRB cross-bench reproducibility check uses MuSiQue-Ans dev split.