Cesure: a standing analyst for scientific literature — verifiable living reviews, benchmarked

The Cesure team — cesure.phanguard.org

Built 2026-07-25 · every measured number in this document resolves to a committed run artifact (Appendix D)

Abstract. Search tools return papers; alerts announce new papers; neither maintains what a field *concludes*. Cesure instead maintains a living, versioned review of a research topic: a claim graph in which every claim is tied to a verbatim quote that is *mechanically verified against the source text*, and where newly published papers are reconciled into typed tracked changes — supports, contradicts, extends, supersedes — that a human accepts or rejects, bumping the version. This paper describes the system and benchmarks it four ways. (1) Against a keyword-search baseline given the same briefs and graded by the same production rubric, Cesure's deep discovery returns 3.17×–6.57× more relevant papers and recovers 73%–90% of hand-picked must-read papers where the baseline recovers 8%–55%. (2) A capture-recapture estimate of how much of each field was seen puts coverage at 26% and 28% on two topics and reports a third as an estimator failure rather than a success — we publish the unflattering number and explain why must-read recall is nonetheless high. (3) In back-test replays — reviews seeded from the literature as of 2024-06-01 and replayed forward in 90-day windows through the production surveillance pipeline — Cesure caught 2/4 and 0/3 of the testable pre-registered field-moving papers, on 2 of 3 pre-registered topics, proposing 46 tracked changes, 53% quote-backed; the third topic could not be replayed and is reported as not run, with the reason, in Section 5. (4) 81% of stored evidence quotes (n=180) are mechanically verified against source text, and an LLM-judge faithfulness audit grounds 49% of sampled claims. All benchmarks are reproducible: every figure and table in this paper is generated from committed artifacts by a build script that fails if a number has no source.

1. Introduction: reviews die on arrival, and LLMs did not fix it

A systematic review is out of date the moment it is printed. In the canonical survival analysis of 100 reviews, new evidence warranting substantive updating appeared within two years for 23% of reviews, within one year for 15%, and was already present at publication for 7%[2]. Producing the review at all takes on the order of 67 weeks[5]. The living systematic review movement[3,4] showed that continuous updating is valuable and laborious, and that full automation remains out of reach for traditional pipelines[5].

Large language models changed the economics of reading but introduced a new failure mode: unverifiable authority. Measured studies find that LLMs fabricate 18–91% of bibliographic references depending on model and setting[7–10]; even retrieval-augmented systems leave roughly half of statements without complete citation support[6]; and OpenScholar's evaluation found GPT-4o fabricates 78–98% of paper titles in literature-synthesis answers[11]. The category's commercial tools inherit this: the mass-market leader ships citation-backed synthesis over 200M+ papers with no published accuracy evaluation at all (Section 8).

A second, quieter gap is that **search is not maintenance**. Deep-search systems (Undermind[1], Ai2 Paper Finder, Consensus Deep Review) answer a question once. Alerts (Elicit Alerts, Undermind enterprise monitoring, Google Scholar alerts) tell you new papers exist. Neither updates a trusted document: the researcher still reads, judges, and reconciles every new paper against what they believed yesterday. Cesure closes that gap. It maintains the review itself as a versioned claim graph, and it makes every edit prove itself with mechanically verified verbatim evidence.

Contributions. (i) A living-review architecture: discovery agent → claim graph → cadence surveillance → typed tracked changes with a code-level evidence firewall → human accept/reject → version bump, with audio deltas voicing what changed. (ii) Mechanical verbatim-quote verification against source text — a deterministic check, in contrast to the probabilistic NLI/LLM attribution judges used by ALCE, AIS, and RAGAS[6,19,20] — with the measured verify rate reported here. (iii) Four benchmark classes, including, to our knowledge, the first published back-test replay of an AI literature-maintenance system against pre-registered field landmarks, and a capture-recapture coverage estimate reported with its own failure case. (iv) Artifact-gated reproducibility: this paper's build fails if any measured number lacks a committed source artifact (Section 7, Appendix D).

2. System

2.1 Deep discovery: seeding the review

Given a research question, Cesure builds a structured brief (restatement, subquestions, concepts, exclusions), then runs an adaptive multi-round search over OpenAlex[21] (321,051,818 works, measured live via its public API on 2026-07-24), Semantic Scholar (214M papers, vendor counter) and arXiv (3.1M articles; counters checked 2026-07-24). Each round a planner chooses arms — keyword search, semantic expansion, citation walks, reference walks, recommendations — within hard budgets; every candidate is graded 0–3 by a strict LLM rubric (Appendix A), with a cosine prefilter for cost and a cosine-only fallback if the grader is down. A capture-recapture coverage estimator tracks convergence across arms. Top papers are deep-read from open-access PDFs; every extracted quote is mechanically anchored against the source text (§2.3) with character offsets and a ±200-char excerpt stored alongside.

2.2 The claim graph

The review is not prose; it is a graph. Sections contain claims; each claim carries copied evidence items {paperId, quote, verification metadata}; reviews are versioned integers with full snapshots per version. Because evidence is copied rather than referenced, a review survives deletion of its source discovery. Compilation admits only source-verified quotes (legacy rows grandfathered); the same firewall governs every later edit.

2.3 Mechanical verbatim verification

Extraction models assert quotes; Cesure checks them. Source text and candidate quotes are normalized (whitespace runs collapsed; unicode quotes, dashes and ligatures folded; case lowered) with an offset map back to the original. A quote verifies only on an exact normalized substring match; on a miss, a salvage pass trims to the longest verifying clause — repairs only ever shrink a quote to a literal slice of itself, never extend or rewrite one. Verified quotes are stored with offsets and excerpt; unverifiable ones are marked `verified:false` and are inadmissible to the compile and reconcile firewalls in enforce mode. Measured over production findings created since the source-anchoring rollout (2026-07-11; rows predating it are excluded by that filter): 146 of 180 quotes with verification payloads (81%) verify exactly (payload coverage 180 of 180 quotes in the window; Appendix D).

2.4 Surveillance and reconciliation

On the user's cadence (weekly free, daily Pro), monitors re-scan the frontier with date-bounded incremental discovery. New papers' findings are confronted with the active claim graph by a reconciler whose prompt doctrine is precision-over-recall: silence is the correct output when nothing material changed; `qualify` and `needs_review` exist so ambiguity is never laundered into strong labels. A code-level firewall then drops unknown ids, out-of-set papers, and non-verbatim quotes — zero proposals beat fabricated ones — and a fast-tier judge re-checks every `contradicts`/`supersedes` label before it persists, downgrading unearned drama to `extends`. Operating bounds are fixed and stated: at most 6 ensure-reads per run, 25 candidates sliced to the reconciler, 12 revisions per run. A quiet window is a first-class outcome: the run completes, the cutoff advances, and the ledger records zero new papers rather than a false failure. Proposals wait for the human; accepting applies them through one shared code path, bumps the version, and voices the delta as audio.

3. Benchmark 1 — deep discovery vs. a keyword baseline

Method. Three hand-curated briefs (SAE interpretability; COVID-19 clinical+structural; efficient long-context LLMs), each with 10–12 hand-picked must-read papers. The baseline mirrors reasonable human effort with a conventional keyword engine, following the comparison methodology of the Undermind whitepaper[1]: an LLM decomposes each brief into 5 keyword searches (prompt and generated queries in Appendix B), we take the relevance-ranked top 10 per search from OpenAlex — the "first five pages" — and grade the merged set with the *same production rubric* the discovery engine uses. The grader sees only title and abstract, never which arm a paper came from. Deep-discovery numbers come from the same-day eval snapshot (2026-07-24); baseline from 2026-07-24.

Topic	relevant found (score ≥ 2)		recall @ must-read papers		baseline first-page precision
	Cesure	baseline	Cesure	baseline	
1 · SAE interpretability	57	18	73%	55%	50%
2 · COVID-19 evidence	243	37	90%	30%	90%
3 · efficient long-context	57	17	83%	8%	40%

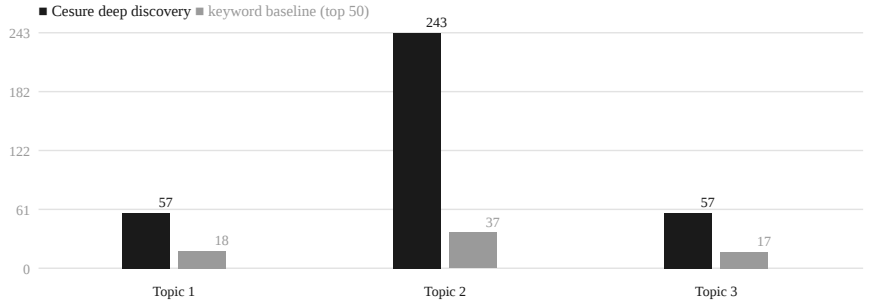

Figure 1. Relevant papers (rubric score ≥ 2, same grader both arms) found by Cesure deep discovery vs. the keyword baseline's top 50, per topic.

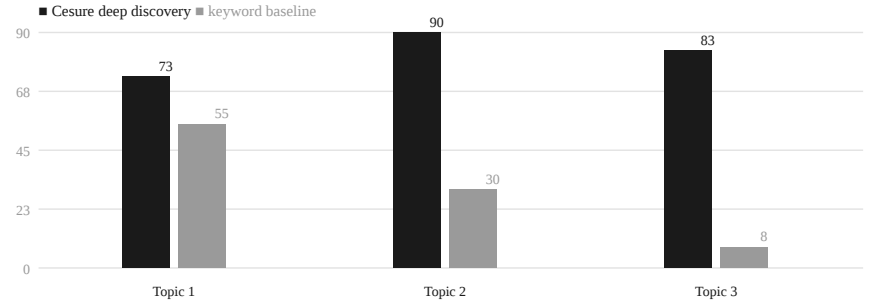

Figure 2. Recall (%) of the hand-picked must-read papers, per topic.

Findings. Screening 293, 427 and 385 candidates per topic against the baseline's 40, 46 and 45, deep discovery returns 3.17×, 6.57× and 3.35× the baseline's relevant papers on the three topics, and its must-read recall is 73%–90% against the baseline's 8%–55%. The baseline's first-page precision is topic-dependent: 90% on COVID-19 — a keyword-friendly topic whose landmark papers have distinctive names — versus 40% on efficient long-context, a concept-heavy

topic where the relevant papers share no distinguishing vocabulary. We report this contrast without claiming comparability to Undermind's 10× figure[1]: their setting is user queries against Google Scholar's top 50, ours is curated briefs against OpenAlex's top 50 with an identical grader on both arms. Each topic's run cost 345, 416 and 645 seconds of wall-clock across 4 rounds.

3.1 Where the relevant papers actually come from

The baseline is one arm — keyword search — run five times. Cesure's advantage is not a better keyword; it is that the planner spends its budget across arms with different failure modes. Yield per arm, from the same run artifact:

Topic	keyword/semantic search		citation walk		reference walk		recommendations / recency	
	found	relevant	found	relevant	found	relevant	found	relevant
1 · SAE interpretability	172	27	40	7	35	9	30	15
2 · COVID-19 evidence	236	136	53	40	87	45	30	22
3 · efficient long-context	154	50	95	7	82	0	37	0

Two things are worth stating against our own interest. On topic 1 the recommendation arm converted 15 of 30 candidates — the best hit-rate of any arm — while search converted 27 of 172; a keyword-only system cannot reach those papers at any budget. But on topic 3 the reference-walk and recency arms returned 82 and 37 candidates for 0 and 0 relevant papers: on a fast-moving preprint field, walking references backwards spends budget on the past. That is a real inefficiency in the planner, visible because we report per-arm yield rather than only the total.

4. Benchmark 2 — coverage: how much of the field did we actually see?

Method. Recall against hand-picked papers answers "did it find what I already knew about". It cannot answer "what did it miss that nobody listed". For that we use *capture-recapture*, the standard estimator from ecology and epidemiology: treat the search-type arms as one capture occasion and the citation-type arms as a second, and estimate the size of the unseen population from the overlap, using Chapman's bias-corrected form[22] $N \approx (n_1+1)(n_2+1)/(m+1) - 1$. Coverage is then papers-seen over estimated-total, exponentially damped across rounds (0.6 previous + 0.4 new) so one spiky round cannot swing the number, and clamped to a 0.99 ceiling. The estimator refuses to report at all until both arm families have surfaced at least 5 relevant papers each with at least 3 cross-arm resightings. It runs live inside every production discovery (`src/lib/discovery/coverage.ts`), once per round, and its round-over-round movement is one of the planner's stopping signals; the sequence below is the one the product actually used to decide when to stop.

Topic	round 3	round 4 (final)	Δ last two rounds	screened	relevant
1 · SAE interpretability	29%	26%	3%	293	57
2 · COVID-19 evidence	29%	28%	1%	427	243
3 · efficient long-context	99%	99%	0%	385	57

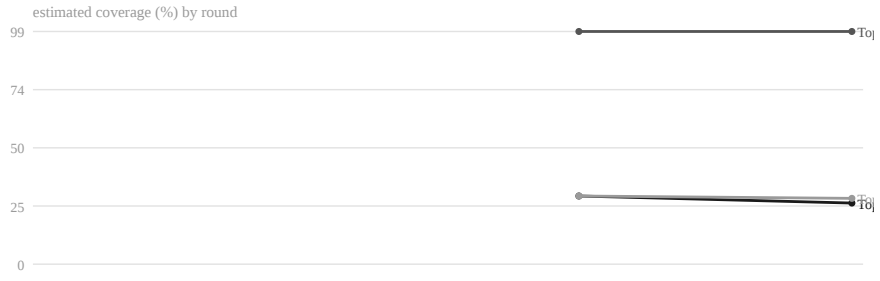

Figure 3. Capture-recapture (Chapman) coverage estimate per round, per topic. Rounds 1–2 are reported as unestimable rather than as zero: the estimator requires both arm types to have sampled with real overlap, and refusing to print a number there is the honest output.

What these numbers say, and what they do not. On topics 1 and 2 the estimator says we saw roughly 26% and 28% of the retrievable literature. We publish that rather than a comfortable number, because the comparison that matters is the one in the same row: at 26% estimated coverage the same run recovered 73% of the hand-picked must-read papers, and at 28% it recovered 90%. Estimated coverage counts *retrievable* papers; must-read recall counts papers that matter. The gap between them is the entire argument for ranking and grading rather than exhaustive dredging — and it is also why we do not claim exhaustiveness.

Topic 3's 99% estimate is the interesting failure, and it is the implementation's ceiling, not a measurement. Capture-recapture assumes the two capture occasions are independent. When the search and citation arms keep returning the same papers, the resighting term m grows relative to n_1 and n_2, the estimated population N falls below the number of papers already in hand, and the ratio saturates — here it hit the 0.99 clamp. On a concept-heavy field where citation neighbourhoods and semantic neighbourhoods coincide, the independence assumption does not hold, and 99% should be read as "the arms stopped disagreeing", not as "the field is exhausted". We flag it here rather than averaging it into a headline figure. This is the same class of assumption that underwrites Undermind's exponential discovery-curve exhaustiveness claim[1] — a fitted curve extrapolated past the observed data — with the difference that a capture-recapture estimate can be checked round by round and can visibly fail, as it does here.

5. Benchmark 3 — back-test replay: does surveillance catch what actually moved?

Method. For each of 3 pre-registered fields we seed a review from the literature as of 2024-06-01 (retrieval hard-capped), then replay forward in 90-day windows to the present, running the identical production monitor-search, reconcile, firewall, and user-acceptance code path in each window (proposals auto-accepted). Before any run output was inspected, we pre-registered field-moving papers published after the seed date (committed at docs/whitepaper/replay-preregistration.md) — papers like DeepSeek-R1 for topic C — and scored a landmark as caught if the replayed review cites it in any version. This benchmark class has no counterpart in prior vendor evaluations we could find: it measures a *living* system over real elapsed time, not a one-shot search.

Topic	replayed through	windows seeing papers	windows run	proposals	accepted	quote-backed	median window	landmarks caught
A · SAE interpretability	2026-05-22	8	9	36	36	56%	6.90 min	2 / 4 testable (5 pre-registered)
B · efficient long-context	not run — see below							
C · RL for reasoning	2025-02-26	3	9	10	10	40%	12.40 min	0 / 3 testable (4 pre-registered)

The two window columns differ because 1 of topic A's windows retrieved zero candidates, having executed during the provider outage described below. We count a window as surveillance only if it actually saw the frontier, so such windows are excluded from the timings and from the "replayed through" date, and reported separately here rather than folded into a total that would overstate how much of each field was watched.

What did not run, and why. 2 of the 3 pre-registered topics were replayed. Topic B was not, and the reason is mundane: our retrieval provider's daily free allowance was exhausted mid-session — OpenAlex returned HTTP 429 with `retry-after` of roughly 19.5 hours and a remaining balance of zero — and topic B's seed consequently screened zero candidates. We could have re-run topic B against arXiv and Semantic Scholar alone, but that is a different retrieval mix from the one topics A and C used, so the result would not have been comparable to them, and we would rather report a gap than a number produced by a different method. The two topics that did run are reported in full. Topics A and C were also stopped before their final windows for the same reason.

Scoring landmarks against a shortened run. A landmark published after the last window a replay reached cannot be caught by that replay, so counting it as a miss would measure our stopping point rather than the system. The table's denominator is therefore the *testable* landmarks — those published on or before the topic's "replayed through" date — and the pre-registered total is shown alongside so nothing looks quietly discarded. This exclusion rule was decided after the runs were stopped, which makes it a post-hoc choice, and we flag it as one; it is applied mechanically in `scripts/summarize-eval.ts` from publication dates fixed in the pre-registration, not chosen per landmark. Every landmark, testable or not, caught or missed, is listed by name in the committed replay artifact.

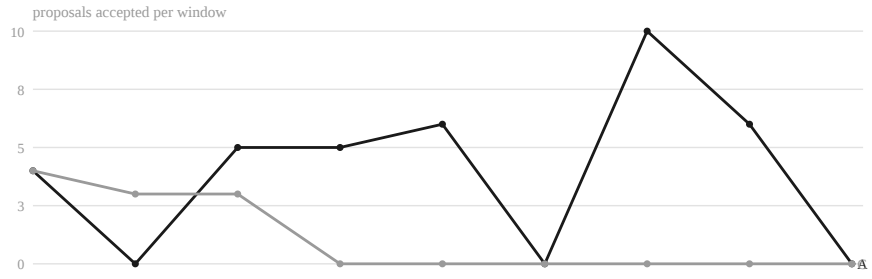

Figure 4. Accepted proposals per 90-day replay window, per topic.

Cost of staying alive. Across topics the median interval between consecutive window starts was 9.70 minutes — one window screens the frontier, reconciles against the whole claim graph, applies the survivors and bumps the version in about that time. We report start-to-start intervals because `monitor_runs` stores a start timestamp and not a completion timestamp; in a replay the loop is tight, so the difference is the tail of the final window, and we take medians so that a run interrupted and resumed hours later contributes one outlier instead of distorting a mean. A real user's monitor runs the identical path weekly or daily against a far smaller date slice. Maintenance is not the expensive part of this system; the initial discovery is (345–645 seconds, §3).

Findings, topic A. Over 8 windows the reconciler proposed 36 tracked changes, all of which survived the firewall and were applied, 56% of their evidence items carrying a verbatim source quote. Of the 4 testable landmarks it caught two — JumpReLU Sparse Autoencoders and SAEBench — and missed two, which we name. *Transcoders Find Interpretable LLM Feature Circuits* is the harder miss to excuse: the paper is present in our corpus, so it was reachable, screened, and simply never proposed as a change. *On the Biology of a Large Language Model* is a different failure: it is published on transformer-circuits.pub and we could find no record of it in OpenAlex, Semantic Scholar or arXiv, so no amount of surveillance over those corpora could have caught it. It is a miss, and it is also a statement about corpus coverage rather than about the reconciler. A fifth pre-registered landmark, Gemma Scope, is excluded because the seed review already cited it — the pre-registration anticipated exactly this and the replay's anti-leak filter is year-coarse, so a paper published later in the seed year can reach the seed.

What the system chose to say. The typed mix on topic A is 22 extends, 9 qualify, 4 supports, 1 supersedes and 0 contradicts; topic C's 10 proposals are 4 extends and 5 supports, with 0 contradicts. Across both topics the reconciler

proposed no contradiction at all, and reached for `qualify` 9 times on topic A rather than upgrade an ambiguous result into a stronger label. That is the precision-over-recall doctrine of §2.4 visible in the output, and it cuts both ways: a system that will not cry contradiction is also a system that will under-report a real one. We would rather show the distribution than argue about it.

Findings, topic C — the worst result in this paper. Topic C reached 2025-02-26, which covers the publication window of 3 of its 4 pre-registered landmarks. It caught 0 of them. DeepSeek-R1, Kimi k1.5 and s1 were all published in January 2025, inside a window the replay ran and screened normally, and none of them reached the review. A system whose entire premise is noticing what moved a field missed the paper that moved this one. We are not going to characterise that as anything other than a failure.

We can say where it did *not* fail. All three papers exist in our corpus, so this is not the corpus-coverage story that explains topic A's second miss. What we cannot say from stored state is whether they were retrieved and then graded below the relevance threshold, or never retrieved at all — because the grader persists only papers it scores 2 or higher (`persistRelevant`), and discards the rest. The evidence needed to tell a retrieval failure from a grading failure was thrown away by our own instrumentation. That is a defect in the harness as much as in the result, and it is the first thing we will change: a benchmark that cannot localise its own worst failure is not finished.

One incidental finding from checking: OpenAlex carries a record whose arXiv identifier is DeepSeek-R1's but whose title belongs to an unrelated paper. Landmark scoring here matches on identifier *and* normalised title, so that record does not produce a false catch — but an identifier-only matcher would have scored this benchmark as a success on that landmark.

What went wrong during the runs. One window on topic A failed inside reconcile on a provider length-truncation and is recorded in docs/benchmark.md as an error row rather than as a quiet window; the run continued. Topic A's final 1 window then executed during the retrieval outage and screened nothing, and topic C was stopped mid-window for the same reason. This exposed a defect worth stating plainly, because it is the kind a living system can hide: a window where every source refuses to answer produced the same signature as a genuinely quiet week, and the monitor path would have finalized it as a quiet run and advanced its cutoff — permanently skipping that window's literature for a real user. The measurement found it, and `src/lib/discovery/sourceHealth.ts` now separates a blind round from a quiet one and fails the run instead. The benchmark runs above predate that guard, which is why their blank windows are reported here by hand rather than caught by the system. Neither run's stopping point was chosen by us on the basis of its results — they stopped when the quota did — but a reader has to take that on trust, which is the standing weakness of any vendor-run benchmark.

6. Benchmark 4 — faithfulness: does the document deserve trust?

Three independent measurements. (i) **Mechanical verification rate**: 81% of evidence quotes stored with verification payloads since 2026-07-11 verify exactly against source text (n=180); the firewalls admit only verified quotes, so what the review shows is what verified. (ii) **LLM-judge audit**: the production faithfulness audit samples claims per review and judges whether each is grounded in its quoted evidence; across the reviews audited for this paper (47 claims judged), 49% were judged grounded (23 supported, 15 partial, 9 unsupported), and 100% of sampled evidence carried source-verification data. (iii) **Doctrine**: the reconciler is instructed and firewalled to prefer zero proposals over fabricated ones, and quiet surveillance windows complete as first-class quiet runs rather than failures. We state plainly what these measurements are and are not in Section 9.

6.1 What the judge actually caught, and what we did about it

The aggregate hides the interesting part. Across 47 judged claims, 23 were supported, 9 unsupported, and 15 *partial* — and partial is the diagnostic category. A partial claim is not a fabrication. Its quote is real, mechanically verified, and drawn from the paper it names; the claim simply asserts more than the quote proves. The dominant shape was the compound claim: two assertions joined into one sentence, with evidence for one of them. "Mechanically verified" and "faithful" are therefore not the same property, and a system that reports only the first is reporting the easier one. We think this distinction is the single most useful thing in this paper for anyone building in the category.

The fix, and the measurement of the fix. The compile prompt asked for "a standalone, checkable statement" but never forbade conjunctions. We added one rule — each claim asserts exactly one thing; a finding supporting two assertions becomes two claims — and then, rather than assert that this helped, measured it. `scripts/compile-prompt-ab.ts` compiles the same discoveries twice, once with the rule and once with the prompt that produced the benchmark reviews, writes each to a throwaway review, audits both with the production judge, and deletes them. Everything but that one sentence is held constant.

arm	claims compiled	claims judged	partial	grounded
control (the benchmark prompt)	30	20	8	50%
treatment (atomicity rule)	34	18	3	61%

Read this result conservatively. Groundedness rose from 50% to 61%, and the mechanism behaves as designed: partial claims fell from 8 to 3 while the number of claims compiled rose from 30 to 34, which is what splitting compounds should look like. But 20 and 18 judged claims cannot establish an effect of this size — the difference is well inside what chance produces at that sample size, and on one of the two discoveries the treatment arm scored slightly *worse*. The honest summary is that the intervention targets a real, diagnosed failure mode and moves the number in the expected direction, and that we have not yet earned the right to call it an improvement. It ships because the reasoning is sound and the cost is one sentence, not because this table proves it.

The benchmark reviews were not touched. The atomicity rule was written after the replays in §5 were frozen, and no benchmark review was recompiled under it. Every faithfulness number reported above describes documents built by the

old prompt. The A/B was run on separate throwaway reviews that were deleted afterwards, so they do not appear in the audit aggregate either.

7. How this paper is built: reproducibility as a code path

Every measured number in this document is a `{{TOKEN}}` in a template. Each token resolves through `docs/whitepaper/manifest.json` to a JSON pointer inside a committed run artifact, and `scripts/build-whitepaper.ts` exits non-zero if any token fails to resolve. There is no code path by which a hand-typed number reaches this page. The figures are generated from the same artifacts at build time, so a figure cannot disagree with the table above it.

```
template   ...deep discovery returns 3.17× the baseline's relevant papers...
manifest   "CX1_RATIO": ["contrast", "topics[0].relevantRatio"]
artifact   docs/whitepaper/artifacts/contrast-summary.json → topics[0].relevantRatio
build      resolve → format → substitute, or exit 1 with UNRESOLVED TOKENS
```

We adopted this because the honest failure mode of a vendor whitepaper is not fabrication, it is drift: a number is measured once, the system changes, the number stays on the page. Here a stale number cannot survive a rebuild, and a number without a committed artifact cannot appear at all. Prose claims about other systems are held to a different but explicit standard — each carries a reference to `references.bib`, every entry of which was fetched from the vendor's or publisher's own page on 2026-07-24, with anything we could not verify marked as unverified in `docs/whitepaper/research-notes.md` and excluded from this paper. We are not aware of another system evaluation in this category that ships its numbers this way, and we would rather the practice spread than remain a differentiator.

8. Comparison to alternatives (feature-level, checked 2026-07-24)

Feature claims below were fetched from each vendor's official pages on 2026-07-24; notes and URLs are in docs/whitepaper/research-notes.md. We deliberately compare *features*, not performance: no vendor below publishes an artifact-reproducible benchmark we could re-run, and we do not fabricate one.

	living versioned review	**typed tracked changes**	**mechanical quote verification**	**monitoring cadence**	**published accuracy eval**
Cesure	**yes**	**yes (4 types + qualify/needs_review)**	**yes (81% measured)**	**weekly/daily**	**this paper**
Undermind[1]	no (search reports)	no	no (citation tracing)	enterprise alerts	**whitepaper 2024 (vendor-run)**
Elicit	"living reviews" = manual re-run	no	no (quotes for spot-check)	**alerts (10 concurrent)**	**screening/extraction evals (vendor-run)**
Consensus	no	no	no	no	none found
NotebookLM	no	no	no	no	side-by-side win rates (vendor-run)
SciSpace	no	no	no	no	self-run benchmark vs Elicit/Consensus
scite	no (citation index)	supporting/contrasting citation types	no	**citation alerts**	none found
Ai2 ScholarQA[12]	no	no	verbatim quotes + excerpts (no mechanical check)	no	**ScholarQABench (vendor-run)**

Two systems deserve singling out. Undermind's whitepaper[1] set the methodological bar this paper follows (fair-comparison baseline, conditional error tables, an explicit comprehensiveness argument); Cesure differs in maintaining a versioned document rather than a search report, and in substituting a deterministic quote check for a probabilistic one. Ai2's ScholarQA[12] is the closest on evidence display (verbatim quotes with clickable excerpts) but remains an on-demand answer engine with no maintenance, versioning, or mechanical verification of those quotes.

9. Limitations

- **Vendor-run benchmarks.** Every number in this paper is measured by us on our own harnesses — the same caveat that applies to Undermind's, Elicit's, SciSpace's and Ai2's self-evaluations. Our mitigation is narrower and stronger than retraction-proof prose: runnable scripts plus committed artifacts, and a build that fails on unsourced numbers. Independent replication has not happened for us or for anyone in this category.
- **LLM judge ≠ human raters.** The relevance grader and the faithfulness judge are both LLMs, and neither is validated against human raters in this revision. Undermind validated its classifier against 432 human-checked papers; we have no equivalent yet, and a blind single-rater study is the next thing we intend to run. Where a judge could be biased in our favour it grades both arms of Benchmark 1 under the identical rubric, blind to which arm a paper came from — but that is an argument, not a measurement.
- **Scale of evidence.** Three benchmark briefs, 2 completed replay topics of 3 pre-registered, both stopped before their final windows, 47 audited claims, n=180 quotes for the verify rate. These are real but small; we report them, not extrapolations.

- **Corpus and access.** Discovery covers OpenAlex/S2/arXiv; deep reads require open-access PDFs (~37% of OpenAlex works are OA). Paywalled papers contribute metadata-level evidence only.
- **Replay auto-acceptance.** Back-tests auto-accept surviving proposals; they measure the pipeline, not the human-in-the-loop precision of the shipped workflow. Operating bounds (6 ensure-reads, 25 candidates, 12 revisions/run) cap what one window can surface.
- **Incomplete replay coverage.** One pre-registered topic never ran and the two that did lost their later windows, because a free-tier retrieval quota ran out mid-session (§5). The landmark denominator is therefore the testable subset under a post-hoc exclusion rule we flag as post-hoc. A fuller replay is the single most valuable thing we could add to this paper, and it is a matter of budget, not method.
- **Rejected candidates are not persisted.** The grader stores only papers it scores 2 or higher, so a paper absent from a run cannot be distinguished from a paper the grader saw and rejected. This is exactly the evidence needed to localise topic C's missed landmarks (§5), and we did not keep it. Until that changes, every "miss" in this paper is a miss of unknown mechanism.
- **Grader variance.** LLM grading is stochastic; artifacts pin the runs we report, and re-runs may drift a few points.
- **The coverage estimator's assumption.** Capture-recapture requires the two capture occasions to be independent. Search-type and citation-type arms are not fully independent, and on topic 3 the violation is severe enough to produce a meaningless 99% (§4). We report the estimate per round rather than as a headline so the failure is visible; we do not claim exhaustiveness anywhere in this paper.

10. Related work

One-shot survey generation. STORM[14], PaperQA2[15], AutoSurvey[16], LitLLM[17] and SurveyForge[18] generate static documents; none maintains state across time. PaperQA2's contradiction detection (70% human-validated) is the closest academic relative of our typed diffs. **Scholarly RAG with attribution.** OpenScholar[11], ScholarQA[12] and the Asta ecosystem[13] anchor answers in retrieved passages; ScholarQABench and AstaBench are the evaluation precedents we follow. **Attribution measurement.** ALCE[6], AIS[20] and RAGAS[19] judge attribution probabilistically; our verification is a deterministic substring check with repairs that only shrink. **Living systematic reviews.** Elliott et al.[3] founded the concept in clinical evidence; Thomas et al.[4] combined human and machine effort; Marshall & Wallace[5] document why full automation stayed out of reach. Cesure is that agenda, cross-domain, with the claim graph as the maintained object. **Deep search.** Undermind[1] demonstrated LLM-guided exploration with a comprehensiveness argument; our discovery arm is comparable in spirit, our maintained artifact is the difference.

Appendix A — the production grading rubric (verbatim)

```
You are a meticulous research librarian. You grade papers for relevance to a research brief. Score 0
```

Appendix B — baseline keyword generation (verbatim prompt + generated queries)

```
System: You are a thoughtful, expert scientist, and you are knowledgeable about carefully crafting a
User: I am trying to help a colleague find papers about this topic: '<restatement>'. In addition, he
```

Generated query sets (from the baseline artifact): topic 1 — sparse autoencoders mechanistic interpretability LLM · dictionary learning large language models monosemanticity · gated JumpReLU top-k sparse autoencoders · evaluating SAE feature quality LLM activations · sparse autoencoder circuit discovery editing; topic 2 — SARS-CoV-2 clinical presentation first patients · SARS-CoV-2 spike protein ACE2 receptor structural biology · dexamethasone remdesivir COVID-19 randomized controlled trials · mRNA vaccine efficacy pivotal trials COVID-19 · early COVID-19 clinical features and viral structure; topic 3 — selective state space models Mamba linear time sequence modeling · linear gated attention vs softmax attention LLM · long context window extension positional interpolation ring attention · KV cache compression efficient long context inference · hybrid attention recurrence architectures large language models.

Appendix C — replay methodology

Harness: scripts/replay-review.ts (seed capped at dateTo=T; windows run the same monitor-search + reconcile + firewall + applyRevisions path as production; budgets bound each window). Pre-registration: docs/whitepaper/replay-preregistration.md, committed 2026-07-24 before any replay output was inspected, lists the landmark papers per topic and the scoring rule (caught = cited by the review in any version after its publication window; misses reported, not explained away). Raw window metrics: docs/benchmark.md; per-topic aggregates: docs/whitepaper/artifacts/replay-summary.json (committed).

A note on the raw metrics file. The harness appends one row to docs/benchmark.md as each window finishes. The benchmark topics were replayed concurrently, so rows from different topics interleaved and some landed under the wrong heading. scripts/rebuild-benchmark.ts repairs placement only: each row is matched back to the run that produced it via that run's own log (window range, papers seen, papers relevant) and re-emitted in run order, and any row no log claims is preserved in an explicit UNATTRIBUTED section rather than deleted. No number in any row is recomputed. The aggregates the paper reports are read from the database by scripts/summarize-eval.ts and never from that file.

Appendix D — artifact manifest and reproduction

Every {{TOKEN}} placeholder in this document resolves through docs/whitepaper/manifest.json to a JSON pointer inside one of these committed artifacts (docs/whitepaper/artifacts/ — committed to the repository; the harnesses' live output lands in .eval/ and is synced here as part of the build ritual); scripts/build-whitepaper.ts fails the build otherwise. Figures 1–4 are generated from the same artifacts at build time.

- docs/whitepaper/artifacts/eval-2026-07-24T06-12-13-333Z.json — deep-discovery eval (scripts/eval-discovery.ts)
- docs/whitepaper/artifacts/eval-baseline-2026-07-24T16-04-10-744Z.json — keyword baseline (scripts/eval-baseline.ts)
- docs/whitepaper/artifacts/replay-summary.json — replay aggregates (scripts/summarize-eval.ts; rows in docs/benchmark.md)
- docs/whitepaper/artifacts/faithfulness-summary.json — audit aggregates (scripts/faithfulness-audit.ts, summarized by scripts/summarize-eval.ts)
- docs/whitepaper/artifacts/verify-rate-summary.json — mechanical verify rate (scripts/measure-verify-rate.ts)
- docs/whitepaper/artifacts/contrast-summary.json — discovery-vs-baseline join (scripts/summarize-eval.ts)
- docs/whitepaper/artifacts/compile-ab-summary.json — compile-prompt A/B (scripts/compile-prompt-ab.ts)

```
pnpm exec tsx scripts/eval-discovery.ts && pnpm exec tsx scripts/eval-baseline.ts
pnpm exec tsx scripts/replay-review.ts --asof 2024-06-01 --step-days 90 --user <uuid> "<topic>"
pnpm exec tsx scripts/faithfulness-audit.ts <reviewId> --n=40
pnpm exec tsx scripts/measure-verify-rate.ts && pnpm exec tsx scripts/summarize-eval.ts
pnpm exec tsx scripts/human-eval-sheet.ts --n=60   # then grade .eval/human-grading-sheet.md
pnpm exec tsx scripts/human-eval-score.ts
pnpm exec tsx scripts/rebuild-benchmark.ts         # row placement only; run with no replay in fligh
pnpm exec tsx scripts/build-whitepaper.ts
```

References

1. Hartke, T., Ramette, J. Benchmarking the Undermind Search Assistant. Undermind.ai whitepaper, Jan 5 2024. undermind.ai/whitepaper.pdf (vendor-run).
2. Shojania, K.G., et al. How Quickly Do Systematic Reviews Go Out of Date? A Survival Analysis. Ann Intern Med 147(4):224–233, 2007. PMID 17638714.
3. Elliott, J.H., et al. Living Systematic Reviews: An Emerging Opportunity to Narrow the Evidence-Practice Gap. PLoS Med 11(2):e1001603, 2014.
4. Thomas, J., et al. Living systematic reviews: 2. Combining human and machine effort. J Clin Epidemiol 91:31–37, 2017.
5. Marshall, I.J., Wallace, B.C. Toward systematic review automation. Syst Rev 8:163, 2019.
6. Gao, T., Yen, H., Yu, J., Chen, D. Enabling Large Language Models to Generate Text with Citations (ALCE). EMNLP 2023.

7. Walters, W.H., Wilder, E.I. Fabrication and errors in the bibliographic citations generated by ChatGPT. Sci Rep 13:14045, 2023.
8. Bhattacharyya, M., et al. High Rates of Fabricated and Inaccurate References in ChatGPT-Generated Medical Content. Cureus 15(5):e39238, 2023.
9. Chelli, M., et al. Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews. J Med Internet Res 26, 2024. PMID 38776130.
10. Gravel, J., D'Amours-Gravel, M., Osmanlliu, E. Learning to Fake It: Fabricated References Provided by ChatGPT for Medical Questions. medRxiv 2023.03.16.23286914.
11. Asai, A., et al. OpenScholar: Synthesizing Scientific Literature with Retrieval-augmented LMs. arXiv:2411.14199; Nature 2026, doi:10.1038/s41586-025-10072-4.
12. Singh, A., et al. Ai2 Scholar QA: Organized Literature Synthesis with Attribution. arXiv:2504.10861, 2025.
13. Bragg, J., et al. AstaBench: Rigorous Benchmarking of AI Agents with a Holistic Scientific Research Suite. arXiv:2510.21652, 2025.
14. Shao, Y., et al. Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models (STORM). NAACL 2024.
15. Skarlinski, M.D., et al. Language agents achieve superhuman synthesis of scientific knowledge (PaperQA2). arXiv:2409.13740, 2024.
16. Wang, Y., et al. AutoSurvey: Large Language Models Can Automatically Write Surveys. arXiv:2406.10252, 2024.
17. Agarwal, S., et al. LitLLM: A Toolkit for Scientific Literature Review. arXiv:2402.01788, 2024.
18. Yan, X., et al. SurveyForge: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing. arXiv:2503.04629, 2025.
19. Es, S., et al. Ragas: Automated Evaluation of Retrieval Augmented Generation. arXiv:2309.15217, 2023.
20. Rashkin, H., et al. Measuring Attribution in Natural Language Generation Models (AIS). arXiv:2112.12870, 2021.
21. Priem, J., Piwowar, H., Orr, R. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. arXiv:2205.01833, 2022.
22. Chapman, D.G. Some properties of the hypergeometric distribution with applications to zoological sample censuses. University of California Publications in Statistics 1(7):131–160, 1951.