FIXED RANKINGS, MOVING LEADERS: A RELEVANCE-LABEL AND SOURCE-IDENTITY AUDIT OF SCIENTIFIC RETRIEVAL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We show that a scientific retrieval benchmark can change its conclusions while every retrieval system stays fixed. In CER-BENCH, a historical collection of 304 synthetic biomedical retrieval tasks over 4,936 PubMed records, replacing seed relevance labels with pooled automated judgments reverses nine of 45 pairwise system orderings on identical saved rankings (Kendall's $\tau_b = 0.584$) and moves the mean-recall leader from a three-round search agent to a dense retriever. Label incompleteness is a known problem in retrieval evaluation, but it is rarely audited together with *source identity*: whether an identifier in the corpus refers to the article text attached to it. We audit both. An authoritative re-fetch of all 4,936 records traces a parser error that let cited-paper identifiers replace an article's own identifiers, changing 247 PMC identifiers, 226 of which occur in reference lists. A controlled Monte Carlo experiment with 1,000 paired, nested disclosures of the existing judgment pool locates the change of leader between six and nine revealed judgments per query, with paired empirical intervals overlapping zero at both points. We release a source-identity-checked retrieval dataset (4,936 documents, 10,313 chunks, 525 documents with conservatively restored own-article text), a hash-bound BM25 substrate, all saved rankings, judgments, disclosure arrays, and audit code. We deliberately do not release a corrected leaderboard: without human relevance validation or new model runs, the defensible output is an explicit boundary between reproducible arithmetic, verified source identity, and still-unvalidated relevance.

1 INTRODUCTION

Scientific search often asks for an evidence set rather than an isolated topically similar paper. A comparison requires both sides; a multi-paper argument requires a valid chain; and a constrained question may require one study satisfying several conditions jointly. Evaluating these requests requires more than a ranking function. It requires trustworthy source identities, an appropriate relevance relation, and a metric whose success condition matches the scientific question.

Incomplete judgments and assessor variation are established problems in information retrieval (IR) (Buckley & Voorhees, 2004; Voorhees, 2000). We do not claim that ranking sensitivity to labels is new. Instead, we examine what remains defensible when label sensitivity coexists with a source-identity failure in a scientific retrieval collection. A score can be exactly reproducible while its biomedical interpretation is unsupported: matching two identifier strings does not establish that either string was attached to the intended article text.

CER-BENCH originally paired synthetic questions with a biomedical corpus and saved outputs from lexical, dense, sparse, fusion, reranking, and agentic retrieval configurations. Its construction made two dependencies particularly important. Generating source papers were retained as positive seed labels, whereas additional candidates received automated judgments. Separately, source identifiers linked PubMed metadata to locally stored article XML. This audit finds problems at both boundaries. The labels change the ordering of fixed saved systems, and an overly broad XML traversal could substitute identifiers from references for the current article's identifiers.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our question is therefore not which retriever wins a repaired benchmark. It is *which conclusions survive an explicit separation of historical arithmetic, authoritative source identity, and validated relevance?* We answer it with three empirical components and a bounded software contribution:

- **A three-way validation framing.** We separate arithmetic validity, source identity and ownership, and scientific evaluation validity (Section 3). The separation blocks a tempting shortcut: rewriting historical labels onto repaired identifiers and presenting the rescored table as a corrected benchmark.
- **An all-record identity audit and conservative text restoration.** Re-fetching all 4,936 PubMed records traces cited-identifier contamination to a descendant XML traversal (247 PMC identifiers change; 226 of the changed historical identifiers occur in reference lists). A restoration that accepts only own-article JATS identities yields a 4,936-document, 10,313-chunk retrieval dataset with byte-reproducible builds (Section 4).
- **Exact label-sensitivity reconstruction and a controlled disclosure experiment.** On fixed saved rankings, seed and expanded labels disagree on nine of 45 pairwise system orders ($\tau_b = 0.584$), and a stored-qrel omission is worth exactly $-1/2430$ in one system mean. A 1,000-replicate paired, nested disclosure of the existing pool brackets the leader change between fractions 0.2 and 0.3 (Section 5).
- **A verified substrate and explicit non-claims.** A hash-bound BM25 index, tested contracts for structural success, selective answering, and matched candidate budgets, and a complete release of saved rankings, judgments, disclosure arrays, traces, and audit code (Section 6, Appendices C–N), together with a list of conclusions the evidence does not support.

This framing matters for interpreting negative findings. We neither replace the historical winner with a new biomedical winner nor treat successful parser tests as a substitute for relevance assessment. Human validation was not performed. No new model experiments were run, and the corrected corpus has neither newly generated benchmark tasks nor an independently validated leaderboard. These restrictions are part of the result: they delimit which claims the retained evidence can carry.

2 Related Work

Relevance evaluation. Buckley & Voorhees (2004) study retrieval evaluation with incomplete information, and Voorhees (2000) examine variations in relevance judgments and measured effectiveness. Our disclosure experiment is a conditional perturbation of available labels, not an estimator of exhaustive relevance. Automated judges introduce additional uncertainty; documented LLM-judge biases motivate caution rather than a correction factor applicable to this corpus (Zheng et al., 2023). In particular, a model-generated label is not equivalent to independent human adjudication.

Scientific retrieval and evidence. DORIS-MAE and LitSearch target scientific document retrieval through multi-aspect and literature-search queries, respectively (Wang et al., 2023; Ajith et al., 2024). BioASQ and TREC Precision Medicine provide biomedical retrieval and question-answering evaluation settings (Tsatsaronis et al., 2015; Roberts et al., 2017); SciFact explicitly links scientific claims and evidence (Wadden et al., 2020). BEIR examines retrieval across heterogeneous datasets, while SciRepEval studies multiple formats for scientific representations (Thakur et al., 2021; Singh et al., 2023). These works motivate careful task and relevance definitions, not a claim that the historical CER-BENCH families are uniquely comprehensive.

Retrieval infrastructure and synthesis. BM25 provides a transparent lexical reference (Robertson & Zaragoza, 2009); reciprocal rank fusion combines rankings without equating their score scales (Cormack et al., 2009). MedCPT illustrates domain-specific retrieval learned from PubMed search logs (Jin et al., 2023). Scientific retrieval-augmented synthesis further increases the importance of preserving the relationship between article identity and evidence text (Asai et al., 2026). Our contribution is not a new retriever or synthesis model. It is a case-specific, joint audit of label dependence and source provenance, together with a limited reconstruction of the retrieval substrate.

Table 1: Distinct evidence regimes. A common count of 4,936 documents does not imply interchangeable text, labels, or identities. "None" denotes no labels for evaluation in that regime.

Regime	Documents	Chunks	Label status
Historical corpus	4,936	17,902	Synthetic; identity-conflicted
Authoritative metadata	4,936	4,914	None; abstract chunks only
Authoritative restored text	4,936	10,313	None; unscored retrieval

3 Audit Design and Evidence Boundaries

3.1 Historical collection and units of analysis

The historical collection contains 304 synthetic tasks in eight intended families: constraint, comparative, contradiction, abstention, multi-hop, temporal, aggregation, and negative-result retrieval (Appendix C). Its original train/development/test counts are 119/60/125. The test set has 108 tasks with nonempty seed labels and 17 with empty seed labels. We call the former *supported* only as an operational name for the nonempty-label stratum. Empty labels on the other 17 tasks are unsupported-query proxies, not verified evidence of corpus-wide absence.

The historical corpus has 4,936 records and 17,902 chunks; 652 records carried a full-text flag. Those are historical counts, not validated ownership statements. The forensic audit identified ten colliding corpus identifier strings, representing 13 excess records under those strings. Consequently, our historical calculations treat document IDs as opaque tokens. They do not dereference them to recover biomedical meaning, choose among colliding records, or transfer their labels to corrected records.

A prior post hoc identity-repair candidate retained 238 of the 304 tasks and assigned 94/49/95 train/development/test tasks (Appendix F). This is not the new authoritative retrieval dataset. The tasks had already been exposed, so rearranging them does not create a fresh held-out benchmark. A new source partition and new task generation remain absent. We retain these distinctions rather than combining historical tasks, repaired identities, and saved rankings into a nominally corrected experiment.

3.2 Three different validation questions

Arithmetic validity. Given fixed token rankings and a specified label set, are counts, denominators, ties, and aggregate scores computed correctly? The forensic rescoring and disclosure experiment answer this question. Their success cannot repair the identity-conflicted inputs.

Source identity and ownership. Does a document's canonical identifier match its own authoritative metadata, and does restored text come from the same article rather than references or nested articles? PubMed re-fetching, JATS front-identity matching, and source-offset checks address this question. They do not establish whether a study's finding is true or relevant to a query.

Scientific evaluation validity. Are tasks meaningful, relevance judgments appropriate, evidence relationships valid, and evaluation splits independent of development? This requires evidence beyond identifier matching or executable tests. We do not claim to have completed it. Automated historical labels remain proxies; no humans supplied biomedical relevance judgments in this audit.

4 Authoritative Identity Audit and Text Restoration

4.1 All-record metadata verification

All 4,936 intended historical PMIDs were re-fetched from PubMed EFetch XML on September 9, 2026. The retained snapshot comprises 26 successful downloads and 140,943,352 bytes, excluding a separate three-record pilot. After recovery of three request-transcription errors, every intended PMID was returned exactly once in the final corpus union; no intended record remained missing or

Table 2: All-record comparison with the authoritative PubMed snapshot. Field counts overlap. Metadata changes exclude canonical document-ID renaming and are not all attributed to the parser bug.

Quantity	Count
Compared records	4,936
Records with metadata changes, excluding canonical renaming	815
PMC identifiers changed	247
Removals / replacements / additions	219 / 8 / 20
Changed historical PMC identifiers found in cited references	226
DOI changes / changed old DOIs found in references	3 / 3
Title changes / abstract changes	118 / 22
Current records with own PMC identifier	2,641

unsupported. Actual request logs retain the mistakes and recovery rather than replacing them with an idealized request plan.

The root cause was an identifier lookup under `PubmedData` using `.//ArticleId`. Descendant traversal includes `ReferenceList/Reference/ArticleIdList`; these are identifiers of cited papers, not necessarily the current article. Both PMC extraction and a DOI fallback could therefore inherit a cited identifier. In the pilot, PMIDs 38308006, 30610625, and 40828286 have no own PMC identifier, yet the old traversal's first PMC value is 6286148, belonging to a cited article. A collision is not necessary for this error: a single incorrect association already violates ownership.

The corrected parser uses direct article-level paths, supports namespaces, and preserves mixed-inline title and abstract text. Own DOI values take priority; a direct own-article electronic-location field is the fallback, never a reference identifier. Conflicting own IDs, duplicate records, missing identity, and unsupported record types are explicit diagnostics. Appendix B gives the scope rule in pseudocode.

The full comparison finds 247 PMC-identifier changes; 226 changed historical PMC identifiers occur in reference lists of the re-fetched XML. All three changed historical DOIs also occur in cited references. This is direct evidence consistent with cited-ID contamination. The broader 815-record metadata discrepancy includes current metadata updates and changed parsing rules, so it is not a count of 815 proven bug-induced corruptions. Nor can the re-fetch alone identify which differences occurred after the historical snapshot. Exact current source agreement and causal attribution to a historical bug are different claims.

4.2 Conservative own-article restoration

Metadata verification first produced an abstract-only version. A separate restoration then rechecked its sources and inspected all 645 locally stored XML files. A candidate body required a unique match between the own JATS front PMID and PMCID and the authoritative PubMed identity. Filenames, cited IDs, and subarticle IDs were never identity evidence. There were 590 unique identity-matched raw candidates.

Titles were checked using normalized equality or both character similarity at least 0.90 and token-set Jaccard at least 0.75. Of 590 candidates, 581 titles matched after normalization and seven more passed both thresholds. Two major mismatches were quarantined, including a local title carrying a retraction prefix. Of the 588 title-accepted candidates, 525 supplied extractable own-body text and 63 did not. These thresholds are conservative string rules, not human semantic validation of titles or retraction adjudication. Appendix G tabulates the funnel and the dataset schema.

Extraction walks selected own front/body structures instead of flattening the entire XML tree. It preserves source order, paragraphs, sections, captions, and explicit table-cell coordinates, while excluding references, back matter, other articles, images, and supplementary payloads. Formula rendering can lose structure and is explicitly marked as not semantically validated. We therefore describe the result as partial source-owned text restoration, not complete recovery of every visual or scientific element.

The restored dataset uses the same canonical string, `PMID:`*pmid*, for each document's `doc_id` and `article_id`. Its 4,936 documents have 10,313 chunks. Titles are actual indexed text, not hidden metadata boosts. Of 22 documents without authoritative abstracts, one gains local text and 21 remain title-only. No document has zero chunks. Chunking uses a hard maximum of 4,096 Unicode codepoints and at most 256 codepoints of within-document overlap. Source intersections bind rendered text offsets to raw-file hashes and XML locators.

Independent rebuilding reproduced all eight dataset files byte-for-byte. Post-build checks covered 673 source hashes and 68,828 chunk/source intersections. These checks establish deterministic source binding and structural ownership. They do not validate biomedical relevance, mathematical equivalence of formula renderings, or redistribution rights. The new corpus is a retrieval dataset without benchmark gold; historical questions and labels were not promoted to new gold.

5 Historical Label Sensitivity

5.1 Fixed-ranking forensic reconstruction

The diagnostic universe consists of the same 108 nonempty-seed test queries for every method. All 125 saved test rows remain accounted for, with null retrieval metrics on the 17 seed-empty rows rather than zeros that would alter the mean. We use the ten original pooling-system identifiers: BM25, dense, BGE, E5-large, MedCPT, SPLADE, hybrid, hybrid-reranked, BGE-reranker, and agent (exact historical configurations in Appendix C). These names identify saved configurations, not new executions. Supplemental RM3 and single-step-agent runs are excluded from rank correlations because they were outside the original pool.

There are 264 seed query–token pairs and 3,240 raw judgment records, exactly 30 per supported query. The stored expanded labels contain 1,369 pairs, while reconstruction from raw judgments with seeds retained contains 1,370. The missing pair is query `multihop_0189` and token `37307965`. On that query, for example, the agent's token recall changes from $4/9$ to $4/10$, an exact macro-mean change of $-1/2430$ over 108 queries. This storage discrepancy is separate from both pooling bias and the corpus-identity error.

For a query q, fixed top-20 token set $T_{s,q}$ from system s, and current positive token set G_q, we compute

$$R_{s,q}(G_q) = \frac{|T_{s,q} \cap G_q|}{|G_q|}, \qquad M_s(G) = \frac{1}{108} \sum_{q=1}^{108} R_{s,q}(G_q). \tag{1}$$

The seed-label leader is the agent, with $M = 44/81 = 0.543209877$. Under reconstructed expansion, BGE has $M = 0.416792975$ and the agent has $M = 0.351057634$, ranking fifth. These values are token-arithmetic endpoints, not an endorsed biomedical leaderboard. We do not interpret their difference across label regimes as an absolute loss of retrieval ability: the rankings are fixed, while both the numerator and relevance denominator can change.

Across the ten systems there are nine strict pairwise order reversals out of 45 pairs between seed and either expanded regime. Kendall's $\tau_b = 0.584306547$ and Spearman's $\rho = 0.735565708$ describe the resulting orders. Stored and reconstructed expansion have identical system order despite different means. We emphasize order rather than a large historical performance table because the scientific finding concerns dependence of the conclusion on label construction, not a corrected estimate of each retriever's effectiveness. Appendix D lists every endpoint mean, every reversed pair, and the exact effect of the missing pair.

5.2 Controlled nested disclosure

The endpoint audit does not describe how much of the available pool must be disclosed before the ordering changes. We therefore ran a local controlled Monte Carlo experiment on the saved rankings and existing raw judgments. This is new computation on historical evidence, not new relevance annotation or model execution.

For each replicate $r \in \{1, \ldots, 1000\}$ and supported query q, we independently uniformly permute all 30 raw judgment records, including negatives. The generator is PCG64 with replicate seeds

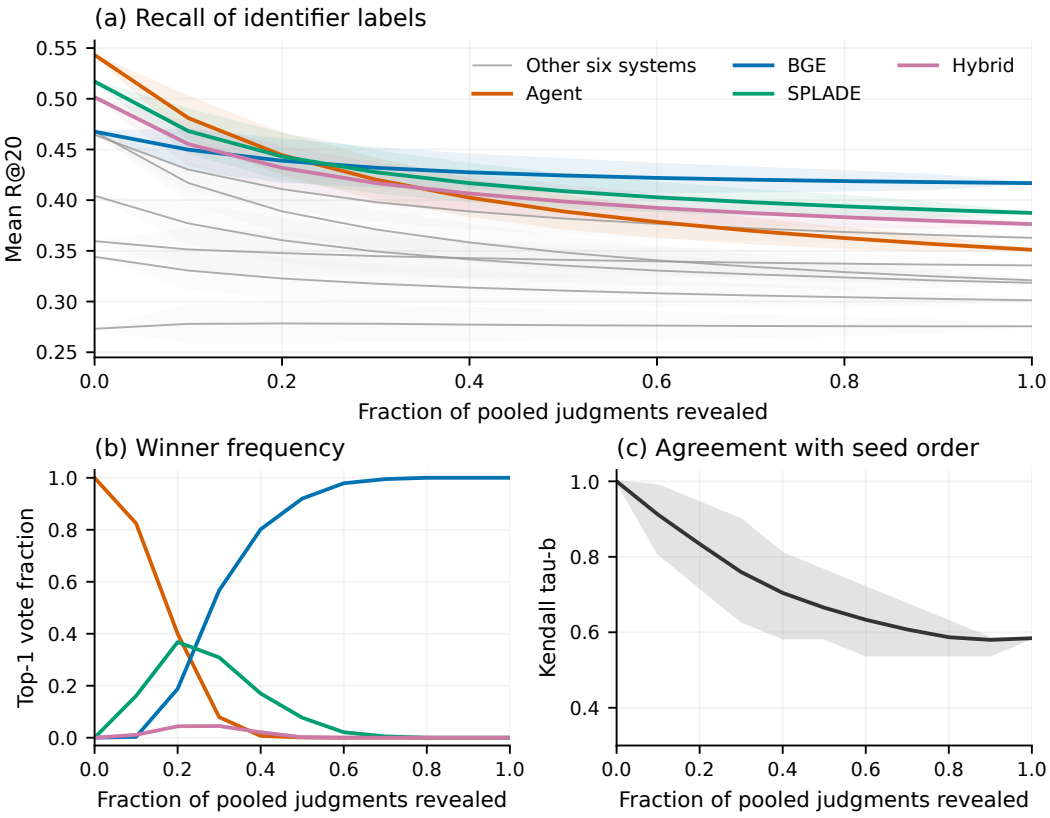

Figure 1: Controlled nested disclosure of the existing historical judgment pool (1,000 paired replicates; ten systems; 108 fixed supported queries). Panel (a) shows all ten fixed systems, highlighting the four with nonzero observed top-1 votes; six others are gray. Panel (b) shows tie-adjusted top-1 frequencies and (c) tie-aware rank agreement. Bands in (a) and (c) are central empirical disclosure intervals, not population confidence intervals; panel (b) shows point frequencies only. These are conditional identifier-level diagnostics, not biomedical retrieval accuracy. Full disclosure reconstructs 1,370 positive pairs. The complete numerical grid is in Appendix E.

20260909 through 20261908. At each fraction $f \in \{0, 0.1, \ldots, 1\}$, reveal exactly $m_f = \lfloor 30f \rfloor$ records from the beginning of that permutation. Every seed positive remains present. Only newly revealed positive judgments add labels:

$$G_{q,r}(f) = G_q^{\text{seed}} \cup \{d_{q,\pi_{q,r}(j)} : j \le m_f,\ y_{q,\pi_{q,r}(j)} = 1\}. \tag{2}$$

All systems share these sets, and each replicate uses the same query permutations across fractions. Thus comparisons are paired across systems and nested across disclosure levels. Revealing six records is not the same as adding six positives; at $f = 0.2$, 648 records are revealed overall but the mean total positive-pair count is 485.286, including seeds. At $f = 0.3$, 972 records are revealed and the mean positive-pair count is 596.013.

Label normalization recognizes negative verdicts before positive verdicts, so the substring in `NOT_RELEVANT` cannot become a positive. Unknown labels, duplicate pairs, and contradictions between a negative raw label and a retained seed fail closed. Unrevealed candidates are operationally nonrelevant for Equation 1, not established scientific negatives. Exact integer score units based on a common denominator avoid artificial tie changes from floating-point summation. All 20 endpoint system/regime means agree with the exact forensic rational references.

5.3 Results and Statistical Interpretation

The leader by mean token R@20 across replicates is the agent at $f = 0.2$ and BGE at $f = 0.3$. This brackets a change on the tested grid between six and nine revealed judgments per query. It is not an

Table 3: Paired conditional disclosure results. $\Delta = M_{\mathrm{BGE}} - M_{\mathrm{agent}}$. Intervals are the empirical 2.5th–97.5th percentiles of 1,000 paired differences, not confidence intervals for a population mean. Top-1 votes are split among exact ties across all ten systems.

f	Revealed/query	Mean Δ	Empirical 95% interval	BGE top-1
0.2	6	-0.005432	$[-0.030669,\ 0.020421]$	18.8%
0.3	9	0.011653	$[-0.013106,\ 0.035547]$	56.7%
0.4	12	0.024971	$[0.001644,\ 0.047598]$	80.2%
0.5	15	0.035463	$[0.014555,\ 0.056464]$	92.0%
1.0	30	0.065735	$[0.065735,\ 0.065735]$	100%

estimated continuous crossover threshold, evidence of monotonicity, or a universal judgment-budget prescription.

At both $f = 0.2$ and $f = 0.3$, paired empirical intervals include zero (Table 3). BGE strictly exceeds the agent in 34.9% and 83.1% of replicates, respectively, but wins against all ten systems in only 18.8% and 56.7%. These are distinct statistics: beating one comparator does not imply winning overall. At $f = 0.5$, BGE receives 92.0% of top-1 votes. At full disclosure it receives every vote, deterministically conditional on these fixed inputs. Intermediate observed rates of zero or one in 1,000 replicates do not prove impossibility or certainty under the disclosure distribution.

Kendall's τ_b complements the leader plot because it measures pairwise order agreement throughout the system list and corrects for ties. The seed ordering ties BM25 and BGE, so a tie-ignorant coefficient would conflate breaking that tie with reversing a strict preference. Average ranks similarly preserve ties; strict reversals require opposite nonzero signs. Spearman correlation provides a rank-displacement summary of the endpoints, not an independent significance test. Definitions are in Appendix A.

The randomness here concerns which already judged records are revealed, conditional on the selected pool, automated verdicts, seeds, queries, and saved rankings. The intervals are not uncertainty about a biomedical population, true relevance, or a model's future performance. Randomly disclosing records from a biased pool does not model missing-not-at-random unknown relevance outside it. Candidate-origin system fields were not retained in the raw judgments, so a provenance-grounded leave-one-system-out pooling analysis cannot be recovered. Ranking overlap is not a substitute for those missing origins.

6 What the Repaired Infrastructure Establishes

6.1 Actual local retrieval, without accuracy claims

The restored dataset was indexed using a hash-bound inverted BM25 implementation. It contains 4,936 documents, 10,313 chunks, 63,109 lexical vocabulary terms, and 2,010,185 postings. Document selection uses the original query's maximum chunk score, with stable index-order ties. This transparent implementation supports equation checks rather than relying on an unverified legacy serialized index. Appendix H states the scoring contract and the eight probe queries.

Eight predeclared, unscored probes were run under three non-LLM conditions: original-query top-24, repeated original query with three batches of eight new documents, and deterministic token-window reformulation. All 24 local runs completed with 24 unique candidates. For every probe, repeated-original and original-top-24 produced exactly the same candidates and primary selection, as expected for stable ranking with adequate candidates. Independent naive-formula comparisons numbered 290, with maximum absolute error 7.1054×10^{-15}. These checks support implementation correctness and integration, not retrieval accuracy or the probes' intended scientific capabilities.

Recorded build and verified-load times were 3.16048 and 2.28054 seconds. Single loaded-index prefix-search measurements over the eight probes ranged from 0.923 to 1.524 milliseconds, with median 1.182 milliseconds. These are bounded local systems measurements, not a latency guarantee; they do not include a live model workflow. Separate native-adapter checks yielded 56 dry

condition plans and 24 offline retrieval checks, with zero provider/model calls. These are not additional model experiments.

6.2 Structural success and selective evaluation contracts

Document recall does not measure whether the retrieved set satisfies a comparison, contradiction relation, or multi-hop chain. A tested evaluator therefore accepts explicit evidence units and relations rather than inferring them from seed labels. Let S be the retrieved document prefix, U the required units, and $E(d) \subseteq U$ the units annotated for document d. Common diagnostics are

$$C(S) = \frac{|\bigcup_{d \in S} E(d)|}{|U|}, \qquad A(S) = \mathbf{1}\left[U \subseteq \bigcup_{d \in S} E(d)\right]. \tag{3}$$

Family success is more specific. Comparison requires both annotated sides. Contradiction requires some explicitly valid pair p with $p \subseteq S$, and multi-hop requires some annotated valid path $p \subseteq S$; path membership, not retrieval order, determines completion. Joint constraint success requires $\exists d \in S : U \subseteq E(d)$ unless an explicit set-level policy permits union completion. Temporal and aggregation success require all annotated time bins or study-value units, respectively. Negative-result success requires evidence explicitly annotated with the negative finding. Arbitrarily combining halves of different paths must not pass merely because their union covers every unit. Appendix J tabulates these rules.

These are implemented contracts tested on software fixtures, not measured structural scores for the historical or restored corpus. Missing annotations yield null metrics and denominator zero, not seed-derived success. Unvalidated annotations can only enter a separately labeled descriptive-proxy mode. The evaluator's status fields cannot authenticate a claimed adjudication; no such real human adjudication was performed here.

Selective evaluation likewise distinguishes `ANSWER`, `ABSTAIN`, and `UNCERTAIN`. Coverage uses all expected tasks; risk requires external evidence-set failure labels for the answered sets. Empty qrels do not establish appropriate abstention, and uncertain decisions are not relabeled as verified corpus absence. Consequently, this paper reports neither scientific answer accuracy nor calibrated selective risk.

6.3 Candidate budgets do not isolate an iteration mechanism

Seven implemented protocols share a cap of 24 unique candidate documents and an original-query final selector: three non-LLM controls, one-shot rewrite, upfront three queries, evidence feedback, and feedback without accumulated evidence (Appendix I). Fixture tests and real-corpus dry checks exercise the interfaces; no live LLM condition has been evaluated. The heuristic is token-window reformulation, not RM3 or learned relevance feedback.

Equal candidate count does not imply equal compute, prompt tokens, observed evidence, or cost. In the planned feedback schedules, accumulated feedback would expose eight then sixteen documents, whereas the no-accumulation condition would expose eight then eight; both retain query history. This exposure confound prevents an isolated interpretation of memory or iteration. We therefore withdraw any claim that historical differences establish an "iteration mechanism." A causal comparison would require executed, provenance-bound controls and an explicit choice of which resources and observations are matched.

7 Discussion and Limitations

The combined audit separates two mechanisms that should not be conflated. Changing the supplied relevance labels changes system order while saved rankings remain fixed. Correcting source identity changes which scientific text can legitimately be associated with an identifier. The first is measured by conditional token arithmetic; the second is supported by authoritative metadata and own-article XML checks. We have not measured a corrected-corpus ranking change, and do not attribute the observed nine reversals causally to the parser bug.

The case nevertheless reveals an important failure mode for benchmark interpretation. Recomputing a table exactly is insufficient when source identity is wrong; repairing source identity is insufficient

when relevance labels are biased or invalid. A valid reconstruction must preserve separate evidence for the two operations. In this study, that separation prevents a tempting but unsupported repair: simply rewriting historical labels onto new canonical IDs and presenting the rescored outputs as a new benchmark.

Several limits remain substantial. The collection is synthetic and historically exposed; no new source-disjoint task partition exists. Seed positives are forced by construction, and the expanded pool contains automated judgments rather than independently adjudicated relevance. Historical generation and judging used a proprietary model, and this audit does not recreate its original service state or establish cross-model robustness. The disclosure design quantifies variation within a finite biased pool, not its completeness. The source snapshot reflects current metadata, so not every discrepancy can be attributed to the old parser. Restoration is partial, with title-only documents and lossy formula/table rendering. Tests validate specified structural behavior, not scientific semantics. Finally, source ownership and retained license statements do not authorize blanket redistribution of article text. A human validation protocol was designed and sized but not executed (Appendix K).

8 Conclusion

A joint provenance and label audit changes what can responsibly be concluded from CER-Bench. The authoritative re-fetch identifies cited-identifier contamination, while fixed historical rankings show nine pairwise reversals under alternative label sets. Nested disclosure locates a change of the mean-score leader between tested fractions 0.2 and 0.3, with overlapping-zero paired empirical intervals at both points. The new source-owned retrieval dataset and local BM25 checks establish a reproducible substrate, not a corrected biomedical leaderboard. The defensible outcome is an audited boundary between reproducible arithmetic, verified source identity, and still-unvalidated relevance.

Reproducibility Statement

The retained artifacts include raw PubMed request logs and XML hashes, all-record comparisons, source-owned text manifests and offsets, exact historical token arithmetic, disclosure permutations and score arrays, and local retrieval traces. The disclosure experiment fixes its random generator, seeds, query universe, and tie policy; all endpoint means match exact rational references. A companion claim ledger maps every quantitative claim to repository-relative evidence. No new model calls are required to reproduce the disclosed token calculations or local source/index checks. Provider-dependent historical generation is not independently reproduced. The final integrated software suite passes all 311 tests with no failures or skips; this establishes software verification, not independent biomedical validation. Code, task files, saved rankings, judgments, disclosure arrays, audit outputs, and identity manifests accompany this submission as anonymized supplementary material and will be released publicly upon publication.

Ethics and AI-Use Statement

This work analyzes published scientific-source records and synthetic tasks; it reports no participant study or human relevance-annotation study. Human validation was not performed. AI assistance was used for code development, audit analysis, and manuscript preparation; this does not constitute independent relevance assessment. Historical task generation and judging relied on a proprietary language model. Neither those judgments nor this audit should be used as clinical advice or evidence of biomedical correctness. Retained provenance supports local research inspection; it does not establish that all article text may be redistributed. Article text is therefore not redistributed: the release contains identifiers, hashes, offsets, task files, rankings, judgments, and code from which the corpus is rebuilt from public PubMed endpoints.

References

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. LitSearch: A retrieval benchmark for scientific literature search. In *Proceedings of the 2024 Con-*

ference on Empirical Methods in Natural Language Processing, pp. 15068–15083, 2024. doi: 10.18653/v1/2024.emnlp-main.840.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'Arcy, David Wadden, et al. Synthesizing scientific literature with retrieval-augmented language models. *Nature*, 650(8103):857–863, 2026. doi: 10.1038/s41586-025-10072-4.

Chris Buckley and Ellen M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 25–32. ACM, 2004. doi: 10.1145/1008992.1009000.

Gordon V. Cormack, Charles L. A. Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 758–759, 2009. doi: 10.1145/1571941.1572114.

Qiao Jin, Won Kim, Qingyu Chen, Donald C. Comeau, Lana Yeganova, W. John Wilbur, and Zhiyong Lu. MedCPT: Contrastive pre-trained transformers with large-scale pubmed search logs for zero-shot biomedical information retrieval. *Bioinformatics*, 39(11):btad651, 2023. doi: 10.1093/bioinformatics/btad651.

Kirk Roberts, Dina Demner-Fushman, Ellen M. Voorhees, William R. Hersh, Steven Bedrick, Alexander J. Lazar, and Shubham Pant. Overview of the TREC 2017 precision medicine track. In *Proceedings of The Twenty-Sixth Text REtrieval Conference (TREC 2017)*, 2017. URL https://trec.nist.gov/pubs/trec26/papers/Overview-PM.pdf.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 4(1–2):1–174, 2009. doi: 10.1561/1500000019.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. SciRepEval: A multi-format benchmark for scientific document representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5548–5566, 2023. doi: 10.18653/v1/2023.emnlp-main.338. URL https://aclanthology.org/2023.emnlp-main.338/.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R. Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 16:138, 2015. doi: 10.1186/s12859-015-0564-6.

Ellen M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000. doi: 10.1016/S0306-4573(00)00010-8.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609.

Jianyou (Andre) Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. Scientific document retrieval using multi-level aspect-based queries. In *Advances in Neural Information Processing Systems 36*, pp. 38404–38419, 2023. doi: 10.52202/075280-1668.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information*

Processing Systems 36, Datasets and Benchmarks Track, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://arxiv.org/abs/2306.05685.

A Statistical Definitions and Scope

Fixed universe and exact means. The historical test universe contains 125 rows per system. Metrics condition on the same 108 tasks with nonempty seed labels at all disclosure fractions. Since seeds are retained, their denominators never become zero. The remaining 17 rows have null retrieval metrics and remain explicitly excluded, not dropped from coverage accounting. Each supported query has equal weight $1/108$, regardless of its positive-label count. Exact common-denominator integer arithmetic determines ordering before conversion to floating-point display values.

Pairing and empirical intervals. For a pair of methods a, b, let $D_r(f) = M_{a,r}(f) - M_{b,r}(f)$. Reported differences average these 1,000 paired values; interval endpoints are their empirical 0.025 and 0.975 quantiles. They describe how scores vary when a fixed finite pool is partially disclosed. They are not bootstrap confidence intervals, hypothesis tests, or confidence bounds on a win probability. Queries and their underlying judgments are not resampled. At $f = 0$ and $f = 1$, label sets are fixed across replicates, so their conditional intervals collapse to points by design.

Top-1 votes and strict exceedance. Let $W_r(f)$ be the set of methods attaining the exact largest mean in replicate r. Method s receives vote $v_{s,r}(f) = \mathbf{1}[s \in W_r(f)]/|W_r(f)|$. Its top-1 rate is $1000^{-1} \sum_r v_{s,r}(f)$; votes sum to one per replicate. Pairwise strict exceedance instead averages $\mathbf{1}[D_r(f) > 0]$. Neither quantity is the probability that a method is best for biomedical users. The manuscript reports rates, not per-replicate vote quantiles as if they were confidence bounds on those rates.

Strict reversals. For systems a, b, a reversal relative to the seed regime occurs if

$$(M_a(0) - M_b(0))(M_{a,r}(f) - M_{b,r}(f)) < 0. \tag{4}$$

A broken seed tie is not a reversal. There are $\binom{10}{2} = 45$ possible pairs, including the seed-tied pair in the reporting denominator. Thus nine endpoint reversals correspond to $9/45$, not a denominator that silently excludes inconvenient pairs.

Rank correlations. For two score vectors over the same ten methods, let C, D count concordant and discordant pairs, and T_x, T_y count pairs tied only in the first or second vector. Kendall's coefficient is

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_x)(C + D + T_y)}}. \tag{5}$$

Average ranks are assigned to exact ties. Spearman's ρ is the Pearson correlation of these rank vectors. The saved disclosure implementation computes tie-aware agreement between fraction pairs; 363 comparisons against an independent SciPy calculation were recorded. The correlations summarize conditional system order, not relevance validity.

Why no cross-regime recall-drop interpretation? Writing $G' = G \cup H$ changes recall to $(|T \cap G| + |T \cap (H \setminus G)|)/(|G| + |H \setminus G|)$. A fixed ranking may score lower even while retrieving additional newly labeled positives, because the denominator grows faster. Consequently, numerical endpoint differences across seed and expanded regimes are not performance deterioration from a model intervention. Within-fraction paired differences compare systems under the same label sets; across-fraction order changes diagnose dependence on those sets.

B Direct-Child Identity Extraction and Audit Root Cause

The essential ownership rule is to select identifiers from the current article's own metadata container. The following pseudocode omits transport and normalization details but makes the traversal boundary explicit:

```
for record in returned_pubmed_records:
    citation = direct_child(record, "MedlineCitation")
    pmid = required_unique_text(direct_child(citation, "PMID"))
    pubmed_data = direct_child(record, "PubmedData")
    own_list = direct_child(pubmed_data, "ArticleIdList")
    own_ids = direct_children(own_list, "ArticleId")
    pmc = unique_or_absent(own_ids, type="pmc")
    doi = unique_or_absent(own_ids, type="doi")
    if doi is absent:
        article = direct_child(citation, "Article")
        doi = own_elocation_doi_or_absent(article)
    emit(pmid, normalize_pmc(pmc), doi)
    # Never fill absent own IDs from ReferenceList descendants.
```

Actual parsing supports namespace-qualified elements and rejects conflicting identities. Repeated identical IDs within one own list are allowed; duplicate returned article records are not. The title and abstract extraction preserves mixed-inline text with `itertext()`, within explicitly selected own fields. Publication-type and other metadata differences may reflect both updated source content and parser policy, which is why Table 2 does not equate every changed row with bug contamination.

The 26 retained batch downloads contain all intended PMIDs after recovery. Two requests each omitted one intended PMID, and another requested an invalid identifier instead of the intended PMID. A supplemental batch recovered the three intended records. This yielded 4,937 unique requested IDs including one invalid non-corpus ID, but 4,936 returned unique corpus PMIDs. An absent optional field in a valid returned record is not a missing article: the snapshot has 2,295 records without PMC identifiers, 20 without DOI, and 22 without abstracts.

The full-text boundary is analogous: use own `front/article-meta/article-id`, not filename inference, arbitrary body IDs, or reference IDs. A single-article wrapper may be deterministically unwrapped; ambiguous multi-article structures are rejected. Major local-title mismatches quarantine local bodies without invalidating the separate authoritative PubMed metadata. These rules avoid turning a valid metadata record into an unverified full-text association.

C Historical Collection: Families, Splits, and Saved Systems

This appendix records the historical construction so that the audited quantities can be interpreted. It is descriptive context from retained construction files and the historical manuscript, not an endorsement of the historical labels or scores.

Task families. Table 4 summarizes the eight intended families. Each family had 38 generated tasks. Non-abstention tasks were generated by prompting a proprietary model (`claude-sonnet-4-20250514`, temperature 0.7) with the titles, abstracts, and metadata of two to three sampled corpus papers; those papers became the seed labels. Abstention tasks combined plausible constraints that the sampled near-miss papers do not jointly satisfy and received an empty label set. The same model, at temperature 0, later judged pooled candidates blind to contributing system and rank. Prefix-normalizing the saved verdict strings yields 1,106 relevant and 2,134 non-relevant judgments (3,240 total, 30 per supported query); $264 + 1{,}106 = 1{,}370$ is the reconstructed positive-pair count used throughout Section 5.

Splits. The historical split is 119/60/125 (train/development/test) with stratification by family. Of the 125 test tasks, 108 have nonempty seed labels and 17 (the abstention family) have empty labels. The identity-repair candidate of Appendix F regroups 238 surviving tasks into 186 evidence components (largest component: 7 tasks) and assigns 94/49/95 tasks. All tasks in every split were exposed during historical development; none is a fresh held-out set.

Saved systems. Table 5 gives the exact historical configuration behind each saved identifier. Retrieval was deterministic; the controller and judge used temperature 0. These are the systems whose saved top-20 token lists are rescored in Section 5. None was re-executed for this paper.

Table 4: Historical task families, what family success requires, and how many of each family's 38 tasks survive the conservative post hoc identity-repair candidate of Appendix F. Retention counts sum to 238.

Family	Code	Success requires	Retained/38
Constraint satisfaction	`constraint`	one document meeting all stated experimental constraints jointly	33
Comparative	`comparative`	evidence for both sides of the stated comparison	26
Contradiction / conditionality	`contradiction`	a valid pair of conflicting findings plus the resolving condition	28
Abstention	`abstention`	recognizing that no corpus document satisfies the query	34
Multi-hop chain	`multihop`	one complete annotated path linking three or more papers	34
Temporal evolution	`temporal`	coverage of every annotated time bin	28
Aggregation	`aggregation`	coverage of every annotated study-value unit	28
Negative / null result	`negative`	evidence explicitly annotated with the null finding	27

Table 5: Saved system identifiers and their historical configurations. The first ten formed the judgment pool; the last two are supplemental and excluded from rank correlations.

Identifier	Class	Historical configuration
`bm25`	lexical	BM25, k_1=1.2, b=0.75, lowercase alphanumeric tokens
`dense`	dense	SPECTER2 (`allenai/specter2_base` with the `allenai/specter2` proximity adapter), exact inner-product search
`bge`	dense	`BAAI/bge-base-en-v1.5`
`e5large`	dense	`intfloat/e5-large-v2`
`medcpt`	dense	MedCPT query/article encoders
`splade`	learned sparse	`naver/splade-cocondenser-ensembledistil`
`hybrid`	fusion	reciprocal rank fusion (k=60) of `bm25` and `dense`
`hybrid_reranked`	rerank	`hybrid` top-50 chunks reranked by `cross-encoder/ms-marco-MiniLM-L-12-v2`
`bge_reranker`	rerank	`bm25` top-50 reranked by `BAAI/bge-reranker-v2-m3`
`agent`	iterative	`claude-sonnet-4-20250514` controller, three search rounds over BM25 tools
`rm3`	lexical PRF	BM25 with RM3 pseudo-relevance feedback (supplemental)
`agent_single_step`	iterative	same controller, one search round (supplemental)

D Complete Historical Endpoint Arithmetic

Table 6 lists the exact endpoint means for all twelve saved methods under the three label regimes. Values are displayed to six decimals; the retained report stores them to twelve decimals and as exact rationals. Table 7 lists the nine strictly reversed pairs among the ten pooled systems, and Table 8 the exact effect of the omitted stored pair. Table 9 reproduces a conditional token-level cluster inference recorded in the same forensic report.

E Complete Disclosure Grid

Tables 10–13 give the full numerical content of Figure 1 at every tested fraction. All quantities are Monte Carlo averages over the same 1,000 paired replicates; intervals are central 2.5th–97.5th

Table 6: Mean token R@20 over the same 108 supported test queries under seed labels (264 pairs), stored expanded labels (1,369 pairs), and reconstructed expanded labels (1,370 pairs). Stored and reconstructed expansion produce identical system order. These are opaque-identifier arithmetic endpoints, not validated biomedical effectiveness.

Method	Seed	Stored expanded	Reconstructed expanded
bm25	0.467593	0.321393	0.320981
dense	0.344136	0.301470	0.301264
bge	0.467593	0.416073	0.416793
e5large	0.464506	0.362193	0.362811
medcpt	0.273148	0.274872	0.275592
splade	0.516975	0.387611	0.387405
hybrid	0.501543	0.376874	0.376359
hybrid_reranked	0.359568	0.336060	0.335648
bge_reranker	0.404321	0.318894	0.318483
agent	0.543210	0.351469	0.351058
rm3 (supplemental)	0.469136	0.317895	0.317483
agent_single_step (supplemental)	0.299383	0.198720	0.198411

Table 7: The nine strict pairwise reversals (of 45 pairs) between seed and reconstructed expanded labels. Each row gives the left-minus-right difference in mean R@20 under seed labels and under reconstructed expansion; opposite nonzero signs define a reversal. The seed tie between bm25 and bge is broken, not reversed, and is not counted.

Left	Right	Seed difference	Expanded difference
bm25	e5large	$+0.003086$	-0.041829
bm25	hybrid_reranked	$+0.108025$	-0.014667
bge	splade	-0.049383	$+0.029388$
bge	hybrid	-0.033951	$+0.040434$
bge	agent	-0.075617	$+0.065735$
e5large	agent	-0.078704	$+0.011753$
splade	agent	-0.026235	$+0.036348$
hybrid	agent	-0.041667	$+0.025302$
hybrid_reranked	bge_reranker	-0.044753	$+0.017165$

empirical percentiles conditional on the fixed pool, queries, and saved rankings. Revealed totals are $108 \times \lfloor 30f \rfloor$ records; positive-pair means count seeds plus revealed positives.

F Post Hoc Identity-Repair Candidate and Component Splits

Before the authoritative re-fetch, a local, network-free repair audit quantified the historical identity defect and produced a deliberately lossy candidate. It is retained as an intermediate diagnostic; it is not the released retrieval dataset and it carries no gold.

Observed defect. The historical corpus has 4,936 records but 4,923 distinct document-ID strings; its chunk file has 17,902 rows but 17,770 distinct chunk IDs. Ten colliding ID strings account for 13 excess records. The most informative collision is the string 6286148, assigned to three corpus PMIDs (38308006, 30610625, and 40828286), whose titles concern CRISPR genome editing, ACT-PCR mutant screening, and a non-model insect genome, respectively. The locally stored file 6286148.xml has own front identifiers PMC6286148 / PMID 22745249 and the title of a fourth paper (a programmable dual RNA-guided DNA endonuclease). This is direct local evidence of a full-text attachment mismatch, later explained by the descendant-traversal bug of Appendix B. Across the corpus, 58 records had a claimed PMCID whose raw article-front PMID disagreed with the corpus PMID, and 124 conflicting alias keys arose after combining local claims and raw front evidence (alias keys are not independent papers). Of 645 raw XML files, 642 supplied a single readable article-front identity and three failed the strict single-article shape check.

Table 8: Exact effect of the single pair omitted from the stored expanded qrels (query `multihop_0189`, token `37307965`; stored 9 positives, reconstructed 10). Task-level R@20 and the exact change in the 108-query mean are shown as rationals.

Method	Stored task R@20	Reconstructed task R@20	Exact mean change
`bm25`	$4/9$	$2/5$	$-1/2430$
`dense`	$2/9$	$1/5$	$-1/4860$
`bge`	$2/9$	$3/10$	$+7/9720$
`e5large`	$1/3$	$2/5$	$+1/1620$
`medcpt`	$2/9$	$3/10$	$+7/9720$
`splade`	$2/9$	$1/5$	$-1/4860$
`hybrid`	$5/9$	$1/2$	$-1/1944$
`hybrid_reranked`	$4/9$	$2/5$	$-1/2430$
`bge_reranker`	$4/9$	$2/5$	$-1/2430$
`agent`	$4/9$	$2/5$	$-1/2430$
`rm3`	$4/9$	$2/5$	$-1/2430$
`agent_single_step`	$1/3$	$3/10$	$-1/3240$

Table 9: Conditional token-level cluster inference recorded in the forensic report: agent minus comparator mean R@20, paired task-weighted cluster bootstrap interval (10,000 resamples), two-sided cluster sign-flip Monte Carlo p with finite-sample correction, and Holm adjustment across the six contrasts. Clusters are historical evidence components over all original train/dev/test tasks. These are arithmetic properties of saved rankings and labels; they assume symmetric independent component differences and are not validated biomedical or causal results.

Labels	Comparator	Difference	95% cluster interval	MC p	Holm p
seed	`bm25`	$+0.0756$	$[+0.0316, +0.1234]$	0.0014	0.0084
seed	`splade`	$+0.0262$	$[-0.0384, +0.0864]$	0.4471	0.4471
seed	`hybrid`	$+0.0417$	$[-0.0063, +0.0912]$	0.1098	0.3114
reconstructed	`bm25`	$+0.0301$	$[+0.0022, +0.0600]$	0.0466	0.2130
reconstructed	`splade`	-0.0363	$[-0.0705, -0.0006]$	0.0426	0.2130
reconstructed	`hybrid`	-0.0253	$[-0.0554, +0.0052]$	0.1038	0.3114

Repair policy and retention. The candidate normalizes identifiers to `PMID:`*n* and `PMC`*n*, requires source-metadata PMID uniqueness and exact title/abstract agreement, quarantines duplicate canonical PMIDs rather than merging them, strips every full-text field (including apparently consistent ones), and retains an old chunk only if its document ID resolves uniquely and its text equals the source abstract prefix. Tasks touching any ambiguous, unknown, or originally full-text-bearing document are excluded. Table 14 gives the resulting counts.

Component-disjoint split. A splitter unions tasks into components using every recursively annotated document, negative, passage and qrel ID, chunk ID, normalized exact question, token-Jaccard question pairs at ≥ 0.85, and explicit cluster IDs, then assigns whole components with a seeded greedy balance toward 15:8:15. The 238 candidate tasks form 186 components (largest: 7 tasks) and split 94/49/95. All checked cross-split overlaps are zero: documents, exact questions, and lexical near-duplicates (seven lexical pairs and two exact-question edges were grouped). Actual tasks carry no explicit cluster annotations or chunk references, so those particular zero-overlap checks are vacuous for this dataset. Semantic or template leakage below the lexical threshold is not ruled out, and the tasks were already exposed. Historical scores were not migrated to this candidate.

G Restoration Funnel, Chunking, and Dataset Schema

Funnel. Table 15 tabulates the own-article restoration described in Section 4. Filenames, cited identifiers, and subarticle identifiers were never identity evidence; a single-article wrapper could be deterministically unwrapped, and ambiguous multi-article structures were rejected.

Table 10: Per-fraction summary. Revealed records per query are $\lfloor 30f \rfloor$ (total $108\lfloor 30f \rfloor$); "Positives" is the mean number of positive pairs including seeds; Leader is the mean-R@20 leader across replicates; "Rev./45" is the mean share of the 45 pairs strictly reversed relative to seed order; τ_b is tie-aware agreement with the seed order; Δ is BGE minus agent; the last column is the strict paired rate of BGE exceeding the agent. Brackets are central 95% empirical intervals.

f	Rev./q	Positives	Leader	Rev./45 [95%]	τ_b vs. seed [95%]	Mean Δ [95%]	BGE > agent
0.0	0	264.0	agent	.000 [.000, .000]	1.000 [1.000, 1.000]	$-.0756$ [$-.0756$, $-.0756$]	.000
0.1	3	374.6	agent	.038 [.000, .089]	.912 [.809, .989]	$-.0312$ [$-.0562$, $-.0054$]	.008
0.2	6	485.3	agent	.077 [.022, .133]	.834 [.719, .944]	$-.0054$ [$-.0307$, $+.0204$]	.349
0.3	9	596.0	bge	.113 [.044, .178]	.760 [.629, .899]	$+.0117$ [$-.0131$, $+.0355$]	.831
0.4	12	706.4	bge	.141 [.089, .200]	.704 [.584, .809]	$+.0250$ [$+.0016$, $+.0476$]	.980
0.5	15	817.0	bge	.160 [.111, .200]	.665 [.584, .764]	$+.0355$ [$+.0146$, $+.0565$]	.999
0.6	18	927.6	bge	.176 [.133, .222]	.633 [.539, .719]	$+.0435$ [$+.0249$, $+.0609$]	1.000
0.7	21	1,038.0	bge	.189 [.156, .222]	.607 [.539, .674]	$+.0504$ [$+.0336$, $+.0658$]	1.000
0.8	24	1,148.6	bge	.199 [.178, .222]	.587 [.539, .629]	$+.0562$ [$+.0436$, $+.0687$]	1.000
0.9	27	1,259.3	bge	.202 [.200, .222]	.580 [.539, .584]	$+.0613$ [$+.0515$, $+.0702$]	1.000
1.0	30	1,370.0	bge	.200 [.200, .200]	.584 [.584, .584]	$+.0657$ [$+.0657$, $+.0657$]	1.000

Table 11: Mean token R@20 of each fixed system at each disclosure fraction (Monte Carlo mean over 1,000 replicates). Columns $f = 0$ and $f = 1$ equal the seed and reconstructed-expanded endpoints of Table 6. Values across columns are not comparable as performance changes because the label denominator grows with f.

System	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
`bm25`	.4676	.4169	.3890	.3711	.3582	.3484	.3407	.3344	.3291	.3247	.3210
`dense`	.3441	.3305	.3227	.3176	.3137	.3108	.3082	.3060	.3043	.3026	.3013
`bge`	.4676	.4498	.4389	.4319	.4274	.4244	.4219	.4202	.4188	.4177	.4168
`e5large`	.4645	.4301	.4108	.3981	.3889	.3819	.3766	.3723	.3686	.3655	.3628
`medcpt`	.2731	.2780	.2784	.2781	.2772	.2766	.2763	.2759	.2757	.2756	.2756
`splade`	.5170	.4682	.4431	.4275	.4168	.4089	.4028	.3980	.3939	.3904	.3874
`hybrid`	.5015	.4553	.4319	.4169	.4066	.3987	.3924	.3873	.3831	.3795	.3764
`hybrid_reranked`	.3596	.3514	.3478	.3448	.3429	.3411	.3396	.3383	.3374	.3364	.3356
`bge_reranker`	.4043	.3771	.3605	.3494	.3415	.3355	.3305	.3269	.3238	.3208	.3185
`agent`	.5432	.4810	.4443	.4203	.4025	.3889	.3785	.3698	.3626	.3564	.3511

Chunking policy. Chunks target and hard-cap 4,096 Unicode codepoints, prefer rendered block boundaries and then whitespace, use a two-newline separator, allow at most 256 codepoints of within-document overlap, and never cross documents. Offsets are half-open codepoint offsets in rendered text, not byte positions in raw XML. Explicit rendering additions are limited to block separators and table-cell coordinate labels.

Files and schema. The dataset comprises eight files: `corpus.jsonl`, `chunks.jsonl`, `source_blocks.jsonl`, `candidate_audit.jsonl`, `local_xml_audit.jsonl`, `excluded_metadata.jsonl` (empty by construction), `summary.json`, and `dataset_manifest.json`. Each corpus row carries a canonical `doc_id`, an identical `article_id`, and the shared `corpus_id` `cerbench-authoritative-fulltext-v1`. Each chunk row adds a unique opaque `chunk_id` and explicit `text`. Each source block records source type, XML locator, raw path and SHA-256, section ownership, and its intersections with chunk-relative and document-relative ranges. The manifest records the SHA-256 of every input and output file and of the producing code, and carries explicit flags: `identity_validated` and `canonical_ids_unique` are true; `human_validation`, `biomedical_relevance_validated`, `historical_tasks_or_qrels_used`, and `public_redistribution_authorized` are false; and `use_scope` is `local_research_only`. Those flags are why the release ships identifiers, offsets, hashes, and rebuild code rather than article text.

Table 12: Tie-adjusted top-1 vote rate at each fraction for the four systems that ever receive a vote; the other six systems receive exactly zero votes at every fraction. Rows sum to one within each column up to display rounding.

System	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
agent	1.000	.825	.400	.079	.007	.001	.000	.000	.000	.000	.000
splade	.000	.161	.368	.309	.170	.077	.021	.005	.000	.000	.000
hybrid	.000	.011	.044	.045	.021	.002	.000	.000	.000	.000	.000
bge	.000	.004	.188	.567	.802	.920	.979	.995	1.000	1.000	1.000

Table 13: Mean average rank (1 = best; exact ties receive average ranks) of each fixed system at each fraction. The seed tie between bm25 and bge at rank 4.5 is visible in the first column.

System	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
agent	1.00	1.19	1.96	2.95	3.64	4.07	4.37	4.61	4.83	4.98	5.00
splade	2.00	2.07	2.12	2.09	2.12	2.12	2.11	2.08	2.03	2.00	2.00
hybrid	3.00	3.18	3.40	3.40	3.17	3.00	2.96	2.95	2.97	3.00	3.00
bm25	4.50	5.86	5.97	6.03	6.12	6.30	6.54	6.83	7.08	7.11	7.00
bge	4.50	3.62	2.58	1.64	1.24	1.08	1.02	1.00	1.00	1.00	1.00
e5large	6.00	5.08	4.99	4.93	4.84	4.74	4.54	4.36	4.16	4.02	4.00
bge_reranker	7.00	7.02	7.15	7.33	7.50	7.64	7.77	7.82	7.85	7.89	8.00
hybrid_reranked	8.00	8.01	7.86	7.64	7.39	7.06	6.69	6.35	6.07	6.00	6.00
dense	9.00	8.97	8.98	8.99	8.99	8.99	9.00	9.00	9.00	9.00	9.00
medcpt	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00	10.00

H VERIFIED BM25: SCORING CONTRACT, PROBES, AND MEASUREMENTS

Tokenizer and scoring. The sole tokenizer lowercases Unicode text and extracts `[a-z0-9]+` matches; there is no stemming, stop list, expansion, title concatenation, metadata boosting, or subword vocabulary, and only each chunk's `text` is indexed. For N chunks, chunk length $L(c)$, mean length $\overline{L}$, term frequency $f(t,c)$, and chunk frequency $\mathrm{df}(t)$,

$$\mathrm{idf}_0(t) = \ln \frac{N - \mathrm{df}(t) + 0.5}{\mathrm{df}(t) + 0.5}, \qquad \mathrm{idf}(t) = \begin{cases} 0.25 \cdot \overline{\mathrm{idf}_0} & \mathrm{idf}_0(t) < 0 \\ \mathrm{idf}_0(t) & \text{otherwise,} \end{cases} \tag{6}$$

$$\mathrm{score}(q,c) = \sum_{t \in q} \mathrm{idf}(t)\, \frac{f(t,c)\,(k_1+1)}{f(t,c) + k_1\big(1 - b + b\,L(c)/\overline{L}\big)}, \qquad k_1 = 1.5,\ b = 0.75, \tag{7}$$

where $\overline{\mathrm{idf}_0}$ is the arithmetic mean of idf_0 over the vocabulary. This is the `rank_bm25.BM25Okapi` convention, verified against an independent direct implementation rather than by importing that package. Repeated query terms contribute repeatedly; out-of-vocabulary terms contribute zero; the replacement IDF is not clipped at zero. All chunk scores participate in descending ranking with stable input-order ties. A document's score is the maximum original-query score over all of its chunks; document ties follow the order of the first winning chunk. The retriever receives no task annotation or label.

Index binding. The build writes numeric postings (`postings.npz`) and an `index.json` holding the exact ordered vocabulary, document and chunk IDs, chunk-to-document relation, configuration, corpus identity and status, input/config/retriever-code hashes, per-array content hashes, and build timing. Loading requires all original inputs again, validates every stored hash, loads NPZ with `allow_pickle=False`, and recomputes postings and IDFs once to verify semantic binding to the current text. The real index has 63,109 vocabulary terms, 2,010,185 term/chunk postings, and 4,243,853 indexed tokens.

Probes. Table 16 lists the eight queries declared before the real index was built. They are capability *intentions*, not verified task properties, and carry no relevance labels; nothing about abstention correctness, contradiction discovery, or relevance is claimed from them.

Table 14: Identity-repair candidate: what survives the conservative policy. Eighty candidate documents have no retained chunk; tasks depending on them are excluded. Whole-record retention is not evidence coverage.

Item	Historical	Candidate	Excluded / quarantined
Corpus records	4,936	4,936 unique canonical PMIDs	0 whole records
Full-text-bearing records	652	0	652 stripped of full text
Chunks	17,902	4,856 abstract chunks	13,046
Tasks	304	238	66

Table 15: Conservative own-article restoration funnel. Title acceptance used normalized equality or (character similarity ≥ 0.90 and token-set Jaccard ≥ 0.75). $525 + 63 + 2 = 590$.

Stage	Count
Locally stored XML files inspected	645
Files with a single readable own article-front identity	642
Unique candidates whose own front PMID/PMCID match authoritative identity	590
Titles equal after normalization	581
Titles accepted by both similarity thresholds	7
Major title mismatches quarantined (incl. one retraction-prefixed title)	2
Title-accepted candidates	588
With extractable own-body text (restored bodies)	525
Without extractable own-body text	63
Documents in released dataset / title-only documents	4,936 / 21
Chunks in released dataset	10,313
Source hashes / chunk–source intersections verified after independent rebuild	673 / 68,828

Measurements. Every probe admitted 24 unique candidate documents; unique-document prefixes required 24–48 chunk rows, and only the first 24 chunk IDs, document IDs, and scores per query were saved, with no article text. Build took 3.16048 s and full verified load 2.28054 s in the first real-corpus run; a later native integration measured 1.711876 s for an explicit verified load and 3.152019 s for a complete adapter invocation including provenance collection and trace writes. Cumulative process high-water memory was 538.53 MiB after build, 668.50 MiB after load, and 1,969.48 MiB after the entire integration including duplicated transport and verification state; these are cumulative `getrusage` values, not per-phase requirements. Prefix-search latency over the eight probes ranged from 0.923 to 1.524 ms (median 1.182 ms; single warm measurements). Independent re-tokenization and naive BM25 equations produced 290 numerical comparisons across 29 sampled chunks and ten queries with maximum absolute error 7.105427×10^{-15}. All 24 non-LLM protocol runs (eight probes, three conditions) completed; repeated-original and original-top-24 agreed on ordered candidates, original-query selection, and RRF selection for every probe.

I Candidate-Budget Protocols

Table 17 lists the seven implemented conditions. All share a cap of 24 unique candidate documents, three rounds of eight new documents where iterative, an original-query primary selector returning at most 20 candidates (descending best-chunk original-query score, stable index-order ties), RRF with $k = 60$ as a diagnostic selector only, and a 420-character snippet cap that is a character clip, not token accounting. The heuristic condition uses the question plus its first and last 24 unique tokens; it is not RM3, a relevance model, or pseudo-relevance feedback. Task annotations never enter a model request. Each admitted document contributes one representative chunk, so duplicate chunks cannot consume budget.

Exposure confound. In the two feedback conditions the model would observe 8 then 16 admitted documents versus 8 then 8; both retain query history. Equal candidate count therefore does not equalize model-visible evidence, prompt tokens, compute, latency, or cost, and no iteration-mechanism conclusion is drawn.

Table 16: Predeclared unscored probe queries (one per family intention).

Intention	Query text
constraint	EGFR inhibitor resistance human lung cancer cell line assay
comparative	CRISPR knockout compared with RNA interference gene silencing efficiency
contradiction	autophagy inhibition increases decreases cancer cell survival conflicting evidence
abstention	quantum entanglement treatment reverses inherited mitochondrial disease clinical trial
multihop	BRCA1 homologous recombination deficiency PARP inhibitor synthetic lethality
temporal	SARS CoV 2 vaccine neutralizing antibodies variants 2020 2024
aggregation	meta analysis immune checkpoint inhibitor melanoma overall survival studies
negative	metformin randomized trial no improvement cancer survival negative results

Table 17: Equal-candidate-budget conditions (protocol v2). "Model calls" counts planned generation calls per task; the primary selector never calls a model. No live condition has been executed.

Identifier	Candidate generation	Model calls
`original_top24`	original question, 24 new documents	0
`repeated_original_3x8`	original question repeated, excluding prior documents	0
`heuristic_keywords_3x8`	question; first 24 unique tokens; last 24 unique tokens	0
`one_shot_rewrite`	one question-only rewrite, 24 new documents	1
`upfront_three_queries`	three question-only queries fixed before any retrieval, 8 new documents each	1
`evidence_feedback`	original question, then two refinements seeing all previously admitted evidence	2
`no_accumulated_evidence`	same iteration, each refinement sees only the latest round's evidence	2

Legacy shards. Historical multi-shard runs of an earlier LLM-query condition exist but are unbound: their rows carry no input hashes, the latest eight-shard group covers only 62 of 125 test queries, and 14 rows have final selections that violate the current selection contract. A structural audit reports exact missing task IDs, duplicates, and conflicts, but cannot recover original scheduling parameters, chunk observations, or selector order. Those shards are retained as unbound exploratory diagnostics and are not merged into any result in this paper.

J Structural and Selective Evaluation Contracts

Table 18 states the per-family success rule implemented and tested in the evaluator. Annotations must supply explicit evidence units and relations; the evaluator never infers them from seed labels or retrieval order. Missing annotations yield null metrics with denominator zero. Annotations without a validated status can enter only a separately labeled descriptive-proxy mode; status fields are declarations that the evaluator cannot authenticate.

Selective decisions. Runs declare `ANSWER`, `ABSTAIN`, or `UNCERTAIN` per task. Coverage divides by all expected tasks. Risk is computed only over answered tasks and only when external evidence-set failure labels exist; otherwise it is null. `ABSTAIN` is the sole positive prediction of corpus absence; `UNCERTAIN` lowers coverage without predicting absence. Empty historical qrels are an unsupported-query proxy and are never treated as verified absence, so no abstention accuracy is reported.

Table 18: Family success rules for a retrieved document prefix S, required units U, per-document annotated units $E(d)$, and explicitly annotated valid pairs/paths.

Family	Success condition
Comparison	both annotated sides are covered by $\bigcup_{d \in S} E(d)$
Contradiction	some explicitly annotated valid pair p satisfies $p \subseteq S$
Multi-hop	some explicitly annotated valid path p satisfies $p \subseteq S$ (membership, not order)
Constraint	$\exists d \in S : U \subseteq E(d)$, unless an explicit set-level policy permits union completion
Temporal	every annotated time bin is covered
Aggregation	every annotated study-value unit is covered
Negative result	retrieved evidence is explicitly annotated with the negative finding

K Human Validation Protocol: Designed and Sized, Not Executed

A blinded two-annotator relevance-validation kit was prepared to make the cost of independent validation explicit. **No human judgment was collected; every label field in the kit is blank.** The kit selected 98 supported test queries before identity filtering, retained 93 after filtering, and added all 17 abstention queries, for 110 queries and 12,657 candidate pairs per annotator (Table 19). Candidates combine retained seed pairs (226), each available major system's top-20 pool, and 550 uniformly sampled out-of-pool control pairs over 3,609 unique documents; 3,986 pairs have parsed full text and 8,671 are abstract-only. Two independent annotators imply 25,314 pair judgments plus 220 task-level reviews. Under a planning assumption of 2–5 minutes per pair, pair review alone is 843.8–2,109.5 person-hours, with a further 105.47–632.85 person-hours if 10–30% of pairs need adjudication at 5–10 minutes each. These are planning assumptions, not measured times.

Table 19: Provisional validation kit size per annotator. All labels are blank.

Family	Retained queries	Candidate pairs
Constraint	14	1,674
Comparative	14	1,543
Contradiction	13	1,366
Multi-hop	14	1,511
Temporal	12	1,306
Aggregation	14	1,809
Negative	12	1,299
Abstention	17	2,149
Total	110	12,657

The kit's validators enforce exact supporting-span matching, provenance hashes, and routing of every uncertain or low-confidence verdict to adjudication; agreement between annotators is never treated as adjudicator sign-off, and final-qrel export is blocked until an independently authorized adjudication import exists. The kit was built with a less conservative identity filter than the repair candidate of Appendix F, so it must be rebuilt against the authoritative corpus, agreed primary systems, and a locked split before any annotation begins. Independently written queries (80 blank collection slots) are likewise planned but absent.

L Trace, Schema, and Data Availability

Historical evidence. Raw pooled records contain `task_id`, `doc_id`, and `judgment`; they do not preserve a candidate-origin system field. The forensic report preserves stored and reconstructed qrels separately, including the exact missing-pair effect. The disclosure artifact retains query/candidate order, all per-query permutations, random seeds, exact score units and divisor, floating means, average ranks, top-1 votes, reversal arrays, and tie-aware rank agreements. Its manifest

binds the inputs, code, configuration, and outputs. No corpus record is dereferenced in that disclosure experiment.

New retrieval data. Each document and chunk carries a canonical `doc_id`, matching `article_id`, and a common versioned `corpus_id`. Each chunk has a unique opaque `chunk_id` and explicit text. Offsets are half-open Unicode-codepoint offsets in rendered text, not byte positions in raw XML. Source blocks retain source type, XML locator, raw path/hash, section ownership, and intersections with chunk-relative and document-relative ranges. Retained nonindexed local abstracts are distinguished from indexed authoritative abstracts to avoid duplicate indexing.

Protocol traces. The implemented budget protocol distinguishes logical searches, cache hits, ranked chunk output, admitted candidates, and model-visible observations. A full retrieval ranking is not automatically a model observation. Candidate admission is document-deduplicated; each admitted document supplies a representative chunk. The intended model trace includes request, response, displayed evidence, usage when available, latency, and typed errors. Character clipping is not token accounting. Dry plans contain no actual API observations or model responses, and cannot be resumed as completed experiments. Historical agent traces record, per round, the action (`SEARCH` or `REFINE`), the issued query, and the controller's stated reasoning.

Release layout. The anonymized supplement, and the public release upon publication, are organized as follows. Code: corpus clients and the direct-child PubMed parser, own-article JATS restoration, verified BM25, budget protocol, structural/selective evaluators, and 15 test modules. Data: historical task files (v1 splits and the v2 collection), the 238-task component-disjoint candidate, task schema, and the identity-checked dataset's manifest, audits, hashes, and offsets. Runs and traces: saved per-task rankings for all twelve historical configurations, the 3,240 raw judgments and expanded qrels, historical agent traces, legacy equal-budget shards, and the equal-budget control runs. Audits: the all-record authoritative comparison, the identity-repair report, the forensic qrel report, the disclosure summary and replicate arrays, and the verified-BM25 receipts. Paper: this manuscript's source and the claim-to-artifact ledger. Article text is not included; a documented script rebuilds the corpus from public PubMed endpoints.

Rights. Public PubMed endpoints identify the metadata sources; local article text remains subject to source-specific permissions. License recognition is not license adjudication, and public redistribution of article text has not been authorized by this audit. The release therefore contains identifiers, hashes, offsets, rankings, judgments, and code, from which the text corpus is rebuilt.

M Software Verification Versus Scientific Evidence

The authoritative-parser evidence records 51 targeted/adjacent passing tests at its snapshot. The restoration evidence records 83 passing tests and byte-identical independent builds. The disclosure evidence records 11 passing tests and exact agreement of 20 rational endpoints. The real-corpus retrieval evidence records 68 passing tests and 290 independent formula comparisons. These suites overlap; their counts must not be summed into a count of unique tests or independent scientific replications.

An intermediate adapter verification encountered a stale historical source-hash assertion; its failed and filtered-run receipts are preserved. A subsequent, separately versioned native-backend integration retained strict provenance checks, reused the immutable index, and bound the updated sources. The final complete configured suite passes 311 tests with no failures, errors, skips, or deselection, in 6.36 s, across 15 modules: budget protocol, component splits and identity, evaluation integration, human-validation kit, identity repair, live-provider clients (network forbidden), authoritative PubMed parsing, qrel disclosure, qrel sensitivity, selective metrics, strict metrics, structural metrics, submission readiness, verified BM25, and verified JATS. Post-suite verification checked 79 frozen source bindings, 213 native runtime bindings, all diagnostic and index hashes, and historical file preservation. Twenty-four actual offline retrieval checks and 24 separate dry plans ran in that integration, with zero model calls. Old build timings and formula comparisons are explicitly identified as historical measurements rather than presented as rerun measurements.

Neither these tests nor programmatically fictional structural fixtures are human experiments. Simulated adjudication statuses in fixtures test software branches; they must not be interpreted as annotations for real papers. Future scientific evaluation would need new tasks tied to the corrected source version, independently justified evidence relations, and explicit split/exposure control. Any future automated-only results would still require the absence of human validation to remain visible. The present manuscript does not fill this gap with historical gold migration, fabricated structural scores, or unexecuted model comparisons.

N Reproduction Commands

All commands run from the repository root with Python 3.11, NumPy, SciPy, and pytest; no provider credentials, network access (except the optional PubMed re-fetch), or model dependencies are required for the audits below. Output directories are write-once: generated artifacts refuse to overwrite existing paths, so choose fresh `-output` locations to reproduce rather than replace the retained evidence.

```
# Full software suite (expected: 311 passed, 0 failed/skipped)
PYTHONDONTWRITEBYTECODE=1 python -m pytest -q -p no:cacheprovider

# Fail-closed submission gate (expected: exit 2, BLOCKED)
python scripts/45_submission_readiness.py --require-ready

# Historical label sensitivity: strict mode fails on identity defects
# by design; the explicit forensic mode treats IDs as opaque strings
python scripts/44_analyze_historical_qrel_sensitivity.py --help

# Controlled nested disclosure (1,000 replicates; PCG64 seeds
# 20260909-20261908); choose a fresh output directory
python scripts/51_qrel_disclosure_experiment.py \
    --output results/readiness/qrel_disclosure/rerun

# Authoritative PubMed re-fetch and direct-child comparison (network)
python scripts/47_build_authoritative_corpus.py --help

# Own-article restoration and verified BM25 index/run
python scripts/50_restore_verified_fulltext.py --help
python scripts/49_index_and_run_verified_bm25.py --help

# Post hoc identity-repair candidate and component-disjoint split
python scripts/39_audit_repair_identity.py --help
python scripts/38_make_component_disjoint_splits.py --help
```

Every retained result directory contains a manifest with SHA-256 hashes of its inputs, code, and outputs; the companion claim ledger maps each manuscript number to the file that produced it.