EvidenceGraph: Measuring and Controlling Repeated Evidence in Retrieval-Augmented Verification

Xuan-Bach Le*†
Ho Chi Minh City University of Technology (HCMUT), VNU-HCM
Ho Chi Minh City, Vietnam
lexuanbach@hcmut.edu.vn

Phat T. Tran-Truong*
Ho Chi Minh City University of Technology (HCMUT), VNU-HCM
Ho Chi Minh City, Vietnam
phatttt@hcmut.edu.vn

Son Ha Xuan
The Business School
RMIT University
Ho Chi Minh City, Vietnam
ha.son@rmit.edu.vn

Duy Nguyen
University of Economics Ho Chi Minh City
Ho Chi Minh City, Vietnam
duynguyen.726202321440@st.ueh.edu.vn

Abstract

Ten Web pages may repeat one report rather than provide ten independent pieces of evidence. With EvidenceGraph, we test whether tracing these shared origins helps an agent verify a claim. We distinguish common origin from shared errors: unrelated reports can still mislead the same verifier. We combine mathematical analysis, a blinded human audit, and controlled comparisons on fixed retrieval results from FEVER, HoVer, and AVeriTeC. Pair F1, which measures how well shared-origin pairs are identified, falls from 0.967 on constructed evidence to 0.568 [0.506, 0.625] on naturally retrieved evidence. Agreement with human passage groups, measured by the adjusted Rand index, is 0.401 [0.269, 0.551] on natural evidence. On 371 held-out constructed cases, uniformly averaging passage scores (support minus refutation) reaches 93.8% accuracy, versus 85.7% for the fixed rule that weights scores by estimated origin. Active selection with lineage gains 5.7 points overall on constructed cases, but loses 15.2 on support-refute mixtures. On natural retrieval results, selecting and grouping passages by estimated origin shows no statistically resolved accuracy gain. Replacing the answer model's verdict with the graph's decision can reduce accuracy. Our results separate origin recovery, evidence selection, and verdict quality. We show why provenance supports source inspection while its value for verification requires matched decision tests.

CCS Concepts

• **Information systems** → **Retrieval models and ranking**; *Web mining*; • **Computing methodologies** → **Intelligent agents**.

Keywords

web research agents, evidence aggregation, source dependence, fact verification, calibration, misinformation

*Both authors contributed equally to this research.
†Corresponding author.

This work is licensed under a Creative Commons Attribution 4.0 International License.
WSDM '27, Hong Kong, Hong Kong
© 2027 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3117-4/2027/02
https://doi.org/10.1145/3862322.3863670

ACM Reference Format:
Xuan-Bach Le, Phat T. Tran-Truong, Son Ha Xuan, and Duy Nguyen. 2027. EvidenceGraph: Measuring and Controlling Repeated Evidence in Retrieval-Augmented Verification. In *Proceedings of the Twentieth ACM International Conference on Web Search and Data Mining (WSDM '27), February 15–19, 2027, Hong Kong, Hong Kong*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3862322.3863670

1 Introduction

Suppose a Web research agent retrieves ten pages supporting a claim. If nine copy the tenth, their agreement traces to one report. The agent has found more text, but not more independent evidence. This problem arises when agents reason over retrieved passages [5, 24, 27, 31, 47]: republication, mirrored pages, citation chains, and paraphrasing can make one report look like many agreeing sources.

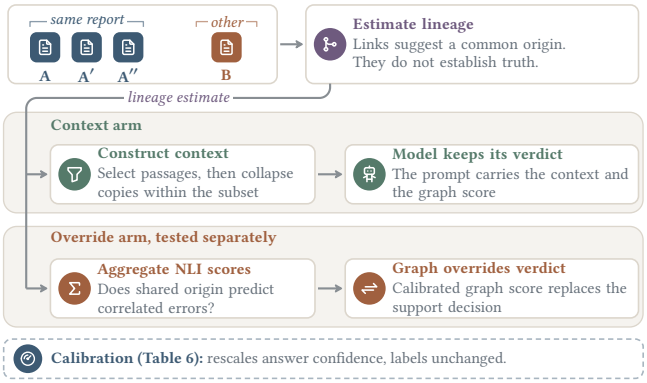

Figure 1: Lineage feeds two separately tested interventions: constructing the answer context and overriding the verdict. Label-preserving confidence calibration is reported in Table 6.

We call the relationships between passages and their origins *source lineage*. Figure 1 illustrates the idea: A, A′, and A″ share one report, while B comes from another. These links reveal repeated reports. Source reliability poses a different question: a separate origin can still be false, and unrelated reports can cause similar

errors in the same verifier. *Error dependence* describes this tendency to be wrong together. Source lineage and error dependence must therefore be tested separately.

With EvidenceGraph, we test two uses of the same estimated lineage. The figure's upper arm changes what the answer model reads. It selects passages, keeps one representative from each predicted origin, and supplies this context together with a graph-derived score. The model keeps its verdict, so this tests the context and prompt score together.

The lower arm aggregates natural-language inference (NLI) passage scores into a calibrated graph probability that replaces the support decision. Applying both rules to the same generated answers isolates replacement. The dashed box tests a third intervention: calibrating reported confidence while keeping every verdict unchanged.

Lineage-based weighting requires both accurate origin recovery and a valid model of shared scoring errors. Our design tests each requirement separately. Classical copying models account for shared sources. Generalized least squares (GLS) combines scores using their error covariance, which describes how errors vary together. Its statistical justification requires that covariance to be correct. We test the mapping from common origins to this covariance, including cases where an independent counter-report is false.

What this study adds. We test origin recovery and its consequences using constructed evidence with known origins and natural evidence from fixed retrieval pools:

(1) **Combining evidence scores.** We prove that exact duplicate errors add no GLS information under a correct covariance model, then test practical lineage-based aggregation against matched alternatives. Uniformly averaging passage scores (support minus refutation) reaches 93.8% accuracy on the constructed test cases.
(2) **Recovering shared origins.** We audit passage pairs and document groups against blinded human judgments. Pair F1, which balances precision and recall for shared-origin pairs, falls from 0.967 on constructed evidence to 0.568 $[0.506, 0.625]$ on natural evidence. Grouping agreement, measured by the chance-adjusted Rand index (ARI), is 0.401 $[0.269, 0.551]$ on natural evidence. A false link can merge separate groups, so both tests matter.
(3) **Changing answers.** Our selection policy with lineage gains 5.7 points overall on constructed cases, but loses 15.2 on support-refute mixtures. On natural evidence, accuracy gains from lineage-based contexts remain statistically unresolved, and letting the graph replace the model's verdict can harm decisions.

We evaluate 2,298 validated constructed cases, fixed retrieval results from FEVER, HoVer, and AVeriTeC, and one hosted answer model.[1] After related work (Section 2), Section 3 defines how we select evidence and change decisions. Section 4 explains their behavior when covariance is fixed. Sections 5 and 6 test origin recovery and answer quality separately. Section 7 turns these findings into guidance for deployment and further validation.

[1]Saved-output artifact: https://github.com/lexuanbach/wsdm2027-evidencegraph-audit

2 Related Work

Retrieval and verification. Retrieval-augmented generation (RAG) couples a retriever with a generator [18, 24, 27], while agentic variants interleave generation with actions, retrieval, or verification [5, 11, 47]. Fact-verification systems retrieve and classify evidence in FEVER [42], multi-hop HoVer [25], and AVeriTeC [39]. Their shared tasks evaluate evidence and verdicts jointly [2, 38]. Recent graph-structured systems organize claims and evidence [16, 29]. With EvidenceGraph, we ask whether agreeing passages come from a shared origin and whether that relationship should change how their scores are combined.

Conflicting and polluted evidence. Knowledge-conflict benchmarks study disagreement among retrieved sources [9, 33, 46]. Other work shows that retrieval-augmented language models follow repeated majorities [26], measures corpus poisoning [41, 48, 49], or tests which source properties persuade models [43]. Astute RAG reconciles model knowledge with retrieval [44]. MIRAGE and EvidentialRAG retain cross-document structure or unresolved conflict [14, 21]. We focus on repeated agreement, where two supporting passages express one underlying report. Paraphrases of a single document can influence decisions more than distinct documents [32], which further motivates an explicit lineage audit.

We isolate classification, grouping, selection, and GLS aggregation through controlled tests on constructed and natural evidence. Matched comparisons identify when each intervention helps or harms verification.

Reliability, copying, and calibration. Source reliability estimates how often a source is correct [23]. Source dependence concerns whether two reports provide distinct evidence. A reliable source may be copied, and an independent source may be wrong. Truth-discovery methods estimate source accuracy and copying from repeated source, object, and value observations [12, 13, 28]. Within each retrieval pool, we compare learned lineage with domain, near-duplicate, and hard-origin GLS rules. Semantic deduplication [1] offers a further comparison for future evaluation. Attribution and credibility research asks whether a claim is supported [6, 8, 15, 36]. We ask whether its supporting passages provide independent evidence. Because neural-network probabilities are often miscalibrated [17], we separately test confidence calibration and covariance weighting. SURE-RAG provides a relevant graph-free comparison by aggregating verifier signals without a covariance model [34].

3 EvidenceGraph

We use EvidenceGraph to select evidence from a fixed retrieval pool. We freeze the answer model and NLI scorer, which classifies each passage as supporting, refuting, or neutral toward the claim. On constructed evidence, we fit three components: a model of shared origin, a controller that selects passages, and a two-parameter map from evidence scores to probabilities. We freeze these components before testing their effect on verdicts.

Passages and pairwise relations. For passage s_i, the signed stance $x_i = p_i(\text{support}) - p_i(\text{refute})$ is positive for support and negative for refutation. Similar wording can indicate copying, but shared origin and shared scoring errors are separate quantities. The deterministic near-duplicate indicator u_{ij} is one when the mean of

token-count cosine and character 4-gram Jaccard similarity reaches 0.80. A logistic lineage estimator uses nine text and metadata features to produce

$$q_{ij} = \Pr(s_i, s_j \text{ share an origin} \mid \text{text, metadata}). \tag{1}$$

The features are token cosine, token and character 4-gram Jaccard, shared rare-token rate, same domain, citation, author, temporal order, and log time gap. We average the stored fits for seeds 103 to 105 before thresholding. The fits are identical, so averaging leaves the probabilities unchanged. The third quantity, r_{ij}, is the assumed correlation between scoring errors. Our implementation uses the heuristic $r_{ij} = q_{ij}$. Section 4 distinguishes this assumption from the conditions under which GLS information has a statistical interpretation.

Clusters and representatives. For context construction, the hard relation $h_{ij} = \mathbf{1}[q_{ij} \geq 0.5]$ defines an undirected graph. Passages linked directly or through other selected passages form a predicted origin group, or *cluster*. Within each cluster, we keep the passage with the highest support or refute probability. Earlier retrieval rank breaks ties. We order representatives by the same score. Near-duplicate, citation, and domain relations are comparison rules rather than additional forced merges.

We freeze thresholds 0.50 and 0.80 and the stored fits. One false link can merge entire groups, so we test sensitivity at both pair and cluster level. Human labels evaluate these retrospective checks without changing the deployed threshold. Construction test decisions remain unchanged across thresholds 0.3 to 0.7 in the retrospective audit (supplement).

Budgeted selection. We retrieve and score all top-20 candidates before selection. The controller then chooses which enter the context, without further Web search. Each policy starts with ranks 1 to 5 and may add three candidates. The controller state contains source count, clipped unit-variance GLS information, mean q_{ij}, support, refute, and neutral mass, conflict, stance uncertainty, and remaining budget. A logistic classifier scores each action against the others. It learns to imitate a teacher with access to the construction's known origins and stances. The action FIND_INDEP chooses the unselected passage predicted to add the most unit-variance information J (Section 4). Passive selection takes the next-ranked candidate. We use construction labels, origin identifiers, and stances to generate teacher targets, keeping them out of deployed features.

Aggregation and endpoint policies. GLS combines selected stances into a weighted mean $\hat{\theta}$. We scale it by its estimated precision w_{eff} to obtain $z = \hat{\theta}\sqrt{w_{\text{eff}}}$ (Section 4), then use a development-fitted logistic map to convert this score to a support probability. Two endpoint policies use this probability. *Provenance-aware context* supplies one representative per predicted cluster and the graph support score in the prompt, then keeps the answer model's verdict. *EvidenceGraph with probability override* replaces its support decision with the calibrated graph probability and retains the model's highest-scoring non-support label otherwise. We store both outputs and rescore the same generations to isolate the effect of override. This paired design separates changing the evidence shown to the answer model from changing its verdict after generation. Selection and aggregation can therefore be evaluated as distinct interventions, even when they use the same lineage estimates.

Execution order and scope. Selection precedes clustering: after the controller adds up to three passages, we group the selected passages and retain their representatives. Unselected passages cannot connect groups. Removing redundant passages can leave fewer than five in the context. On natural evidence, the answer model reads these representatives. On constructed evidence, aggregation uses all selected NLI scores before representatives are chosen.

Graph calibration and answer calibration. Graph calibration can change the verdict through override. In contrast, positive-temperature scaling softens or sharpens the answer model's probabilities without changing their ranking or its chosen label. We fit a temperature on half of each endpoint population and evaluate it on the other half (Table 6). Withholding low-confidence answers requires a separately validated abstention rule.

4 Aggregation under Dependence

Ten independent measurements can provide more information than one, but ten copies of the same measurement cannot. GLS formalizes this distinction through error covariance. EvidenceGraph estimates common origin, so applying GLS requires a mapping between lineage and correlated errors. We derive information under known covariance, then test our approximation.

GLS information. For n passages, model each signed stance as $x_i = \theta + \varepsilon_i$, where θ is their common underlying stance and ε_i is scoring error. Assume $\mathbb{E}[\varepsilon_i] = 0$ and positive-definite error covariance Σ. Generalized least squares (GLS) forms a weighted mean, accounting for both individual uncertainty and shared errors. With $\mathbf{1}$ denoting a vector of ones, it gives

$$\hat{\theta} = \frac{\mathbf{1}^\top \Sigma^{-1} x}{\mathbf{1}^\top \Sigma^{-1} \mathbf{1}}, \qquad w_{\text{eff}} = \mathbf{1}^\top \Sigma^{-1} \mathbf{1}. \tag{2}$$

For fixed, correct Σ, w_{eff} is the inverse variance of $\hat{\theta}$: larger values mean a more precise estimate. If R is a correlation matrix, then $J = \mathbf{1}^\top R^{-1} \mathbf{1}$ measures information after setting each passage's error variance to one. It equals n under independence, but negative GLS weights can make $J > n$ even when all pairwise correlations are nonnegative. The controller uses $\min\{n, \max(0, J)\}$ as a bounded feature. Clipping applies only to the controller feature, leaving Eq. (2) unchanged.

PROPOSITION 1 (AN EXACT DUPLICATE ADDS NO INFORMATION). *Let $x \in \mathbb{R}^n$ have nonsingular covariance Σ, and let $A \in \mathbb{R}^{(n+1)\times n}$ append an exact copy of coordinate j. Thus $x_+ = Ax$, $\Sigma_+ = A\Sigma A^\top$, and $\mathbf{1}_{n+1} = A\mathbf{1}_n$. Using the Moore–Penrose pseudoinverse for the singular augmented covariance,*

$$\mathbf{1}_{n+1}^\top \Sigma_+^+ \mathbf{1}_{n+1} = \mathbf{1}_n^\top \Sigma^{-1} \mathbf{1}_n.$$

PROOF. A has full column rank and $\Sigma A^\top$ has full row rank, so the reverse-order law gives $\Sigma_+^+ = (A^\top)^+ \Sigma^{-1} A^+$. Since $A^+ A = I_n$ and $A^\top (A^\top)^+ = I_n$, substituting $\mathbf{1}_{n+1} = A\mathbf{1}_n$ returns $\mathbf{1}_n^\top \Sigma^{-1} \mathbf{1}_n$. □

Duplicating a measurement repeats its error, so it adds no information. The proposition requires the errors themselves to be

identical. Two equal stance scores or a predicted shared-origin probability $q_{ij} = 1$ do not establish that condition.

Proposition 2 (Saturation under positive equicorrelation). *For unit-variance observations with common correlation* $0 < \rho < 1$,

$$J = w_{\text{eff}} = \frac{n}{1 + (n-1)\rho} \longrightarrow \frac{1}{\rho} \quad \text{as } n \to \infty.$$

At $\rho = 0$, *information grows linearly as* $w_{\text{eff}} = n$.

Proof. $R\mathbf{1} = (1+(n-1)\rho)\mathbf{1}$, so $w_{\text{eff}} = \mathbf{1}^\top R^{-1}\mathbf{1} = n/(1+(n-1)\rho)$. □

For ten unit-variance observations, independence gives information 10, whereas correlation 0.9 gives $10/9.1 \approx 1.10$. Thus ten strongly correlated observations carry only slightly more information than one independent observation.

Our numerical check covers 30 (n, ρ) settings and single-coordinate duplication of positive-definite covariances. It reproduces both identities to floating-point precision. We archive the script and errors with the analysis records.

The deployed covariance. We use the stance $x_i \in [-1, 1]$ and proxy variance $v_i = \max(0.05, 1 - |x_i|)$. We set $r_{ij} = q_{ij}$ and blend the raw matrix R_0 (unit diagonal, off-diagonal q_{ij}) with the identity matrix using $\lambda = 0.05$. This shrinks the assumed correlations toward independence. We then make the matrix positive semidefinite (PSD) by flooring its eigenvalues at 10^{-9}, obtaining $\widetilde{R}$. The diagonal is not rescaled after this correction. We form $\Sigma_{ij} = \sqrt{v_i v_j}\,\widetilde{R}_{ij}$ and use a pseudoinverse cutoff of 10^{-10}. These operations define our covariance proxy. We test how well it describes scoring errors.

The evidence statistic

$$z = \hat{\theta}\sqrt{w_{\text{eff}}} = \hat{\theta}/\widehat{\text{SE}}(\hat{\theta}) \tag{3}$$

is the GLS mean divided by its estimated standard error. A logistic map $\ell(az + b)$ converts this score to a probability between zero and one. We fit $a \geq 0$ and b on construction development data, so larger scores cannot yield smaller probabilities. Probability calibration depends on this fitted map and its target regime, especially when the label space changes.

Why origin and error correlation differ. Setting $r_{ij} = q_{ij}$ treats shared-origin probability as error correlation. Lineage and error correlation encode different relationships. Distinct origins can share a claim, retrieval process, or NLI passage scorer, and thus share error even when q_{ij} is small. Conversely, common origin does not establish the accuracy of a report. Testing that dependence requires measuring scoring errors against reference labels. Because the proxy variances depend on observed stances, the deployed standard error is a heuristic score rather than the exact sampling uncertainty of a fixed-weight estimator. Our propositions establish the fixed-covariance guarantees. We test the deployed estimate empirically.

The covariance estimate also determines the controller's acquisition score. For an unopened passage s', our FIND_INDEP action maximizes the predicted change in J over a bounded pool. It chooses the largest immediate gain under the estimated covariance, so it can favor a different origin over a more relevant passage. Section 6 tests the resulting selection behavior and compares aggregation rules.

5 Experimental Setup

Studies and benchmarks. We combine an analytical duplication diagnostic with three empirical studies on fixed inputs: recovering origins and combining scores, selecting passages within a budget, and measuring final answers. We use development data from FEVER [42], HoVer [25], and AVeriTeC [39]. FEVER and HoVer retrieve from the 2017 Wikipedia dump using BM25 [37], hyperlink expansion, fusion with E5 text embeddings [45], and a MiniLM cross-encoder [40], then retain 20 passages. AVeriTeC reranks 100 of 1,000 claim-scoped blocks with Mixedbread embeddings [30] and maximal marginal relevance (MMR) [7], then retains 20. We hold retrieval fixed across policies.

We evaluate deterministic $N = 500$ subsets. A stable-hash sample with seed 103 selects FEVER and HoVer cases with complete retrieval, while AVeriTeC uses all 500 development claims. Complete annotated evidence is present for 259 of 348 FEVER pools with annotated gold evidence and at least six retrieved passages, 93 of 500 HoVer cases, and 31 of 500 AVeriTeC cases. This counts complete evidence, not just any annotated item. Overall results use retained endpoint cases. The observed-origin proxy is Wikipedia page identity for FEVER and HoVer, and registered domain for AVeriTeC after archive unwrapping. Page and domain identities alone do not establish where a report originated.

Controlled construction. EvidenceGraph-20K pairs 20,000 balanced FEVER claims with 5 to 30 candidates from four recorded families. They are independent evidence or distractors, verbatim mirrors under synthetic domains, citation copies, and locally generated paraphrases. Paraphrases must preserve anchors, pass bidirectional entailment at 0.8, and satisfy a structural gate. We vary exact copying, repeated paraphrases, mixtures of support and refutation, and useful evidence hidden behind eight same-origin distractors. Topic-disjoint splits allocate 14,000, 3,000, and 3,000 candidate claims to train, development, and test.

Of 20,000 candidate claims, 2,298 pass the complete construction-validation process. These yield 1,568 training, 359 development, and 371 test episodes, 44,064 sources, and 453,986 within-episode pairs. We report construction results on this validated subset. The public artifact (Ethical Considerations, Availability) rebuilds the original saved-output tables and Figure 3. Later analyses are supplied in the companion analysis archive. The full construction release remains pending its release gates.

Human lineage audit. Two independent annotators each labeled 384 pair items and partitioned 30 episodes into common-origin groups. Corpus, sampling stratum, provenance condition, and model scores were blinded. The audit contains 640 pair items over 491 episodes, including 128 double-annotated pairs, and 48 partitioned episodes, including 12 double-annotated episodes. Cohen's kappa [10] measures pair-label agreement beyond chance. Adjusted Rand index (ARI) [22] measures agreement between passage groupings, also corrected for chance. Intervals resample episodes 2,000 times. The natural subset contains 469 decided pairs over 363 episodes and 36 partitioned episodes. The constructed subset contains 159 decided pairs over 123 episodes and 12 partitions. We exclude UNSURE from primary precision and recall. Its 1.6% rate is covered by sensitivity bounds. Stratified sampling measures

performance on the audited categories. Population-wide prevalence and recall require representative sampling. The supplement reports score-bin coverage.

Selection protocols. Each policy begins with ranks 1 to 5, then adds at most three candidates from the fully scored top-20 pool. We compare passive next-rank, MMR with $\lambda = 0.75$, near-duplicate diversity, observed-origin diversity, hard lineage diversity, soft FIND_INDEP, and a reference rule that knows the annotated evidence and gives a non-deployable performance ceiling. Natural pools contain 17.1 to 17.8 distinct observed origins on average. Deeper pools and other retrievers require separate tests.

Models and comparison rules. NLI uses MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli [19] at frozen revision `b3546ea6`. End-to-end evaluation uses one fixed answer backbone, us.anthropic.claude-haiku-4-5-20251001-v1:0 [3] with temperature 0 requested. Hosted decoding provides no seed guarantee.

We isolate verification after retrieval. Live search, adaptive queries, and other model scales require further tests. Aggregation compares independence GLS (inverse-variance weighting), domain deduplication, near-duplicate clustering, learned hard lineage, learned soft GLS, and references using recorded construction origins. We use archived predictions to compare two regularized weighting rules (ridge variants) with uniform stance averaging on the same claims. Controller ablations remove active seeking, lineage features, contradiction handling, and effective information scaling. End-to-end arms compare closed-book answering, fixed top-5 retrieval, EvidenceGraph context selection, and EvidenceGraph with probability override.

Metrics and uncertainty. We report accuracy, class-averaged F1 (macro-F1), and expected calibration error (ECE), the mismatch between confidence and accuracy across ten bins. We also report false acceptance on false construction claims, true-claim retention, and false-support rate (FSR) at the endpoint: among gold non-Supported claims, the fraction predicted Supported. For lineage pairs, precision measures how often predicted links are correct, while recall measures how many true links are found. We also report F1 and Brier score, the mean squared probability error. Confidence intervals (CIs) use 1,000 paired claim or claim-cluster bootstrap resamples unless noted. Pairing keeps methods' outputs for the same claim together. These intervals quantify individual comparisons without joint correction across policies, metrics, or datasets.

Protocol controls. We fit learned components and calibrators on construction training and development data. The archived covariance selection records choose ridge parameters by development log loss. We retain those choices when recomputing test metrics. Soft GLS versus hard lineage has no fixed equivalence margin, so zero-containing intervals leave direction unresolved rather than establishing equivalence. Human-audit task files are generated and blinded before either annotator sees an item.

For the end-to-end table, we retain complete structured outputs from every arm, yielding 498 FEVER, 494 HoVer, and 447 AVeriTeC cases. Override contrasts instead use all paired generations (500, 494, and 481, respectively). The artifact reproduces summaries and prediction audits. Full reruns require raw construction data and weights absent from the public release. Private inputs support the covariance, threshold, and held-out calibration analyses reported here. These analyses have their own protocols and records, separate from the withdrawn historical claims.

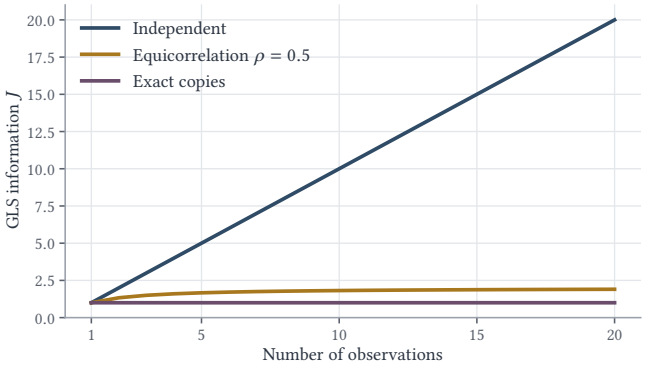

Figure 2: Analytical GLS information under unit variance. Independence gives linear growth, positive equicorrelation gives saturation, and exact copies contribute one unit. These curves assume correctly specified covariance.

6 Results

We trace how repeated evidence shapes the final verdict. We begin with the information carried by copies (Section 6.1), then test origin recovery (Section 6.2), aggregation and selection on constructed evidence (Section 6.3), and transfer to natural retrieval pools (Section 6.4).

6.1 RQ1: What does exact copying add?

Figure 2 illustrates Propositions 1 and 2. With unit variance and equal correlation 0.5, twenty observations contain $20/10.5 \approx 1.90$ units of GLS information. Twenty exact copies contain one, while twenty independent observations contain twenty. These curves concern statistical information. To test whether models resist repetition in practice, we run a separate experiment (Supp. Table S2). We rebuilt 40 refuted FEVER claims with the original selection rule and filled 20 context slots with 1 to 20 verbatim copies of each claim. Twenty copies raised false-claim acceptance relative to one for Qwen2.5 models (1.5B to 72B) [35] and Claude Sonnet 4.6 [4], with positive trends against log copy count after accounting for shared claims (Holm-adjusted $p < 0.02$ [20]). A single representative copy with a dependence-aware instruction lowered twenty-copy acceptance for four of the six models, but not for the 3B and 14B models.

Key finding. Copies add no information when their shared errors are known. Yet all six tested models accept more false claims with twenty copies than with one; intermediate counts need not increase monotonically. Detecting repetition and resisting its influence are separate problems.

6.2 RQ2: How well does lineage transfer?

We evaluate lineage recovery against an independent, blinded human reference. The two raters obtain three-way $\kappa = 0.951$ on 128 shared pair items and 0.977 observed agreement. On 12 shared

Table 1: Lineage estimation against human labels. Intervals use episode-level bootstrap resampling.

	Natural pairs (n = 469)			Constructed	Natural clusters
System	**Precision**	**Recall**	**F1 [95% CI]**	**F1 [95% CI]**	**ARI [95% CI]**
EvidenceGraph lineage	0.472	0.712	0.568 [0.506,0.625]	0.967 [0.919,1.000]	0.401 [0.269,0.551]
Text-reuse clustering	0.907	0.250	0.392 [0.295,0.483]	0.851 [0.763,0.924]	0.477 [0.320,0.629]
Observed-origin identity	1.000	0.160	0.276 [0.188,0.365]	0.000 [0.000,0.000]	0.393 [0.253,0.542]

Two-rater agreement: pair κ = 0.951 (n = 128), partition ARI = 0.983 (n = 12). EvidenceGraph constructed-cluster ARI = 1.000. Frozen NLI stance macro-F1 = 79.8 and ECE = 0.121 on 7,106 construction edges.

Table 2: Natural-lineage recovery by corpus. Edge F1 decomposes the pooled natural result in Table 1. Same origin gives the audited positive-pair rate.

			Edge F1		
Corpus	n	**Same origin**	**EvidenceGraph**	**Reuse**	**Origin ID**
AVeriTeC (domain)	159	0.75	0.804	0.347	0.312
FEVER (page id)	156	0.12	0.309	0.462	0.273
HoVer (page id)	154	0.12	0.337	0.552	0.000

Two-rater agreement by corpus: AVeriTeC κ = 1.000 (n = 26), FEVER 0.877 (n = 35), HoVer 0.839 (n = 31)

partition episodes, their mean ARI is 0.983. Against this reference, Table 1 compares the lineage rules. Our frozen lineage estimator reaches pair F1 0.967 with a 95% interval of [0.919, 1.000] on the human-audited constructed sample, and its clusters match the human partitions with ARI 1.000. On natural evidence, pair F1 falls to 0.568 [0.506, 0.625] and cluster ARI to 0.401 [0.269, 0.551].

Natural recall is 0.712, but precision is 0.472: fewer than half of predicted links are correct. False links are the principal precision failure. Text-reuse clustering misses more related pairs but groups passages more accurately: it reaches pair F1 0.392 and a higher cluster ARI of 0.477. Observed-origin identity reaches pair F1 0.276 and cluster ARI 0.393. Natural ECE is 0.180, so lineage probabilities also need target-domain validation.

Natural recovery varies with the available provenance signals (Table 2). The *Same origin* column gives the audited positive-pair rate, and κ measures rater agreement. Recovery is strongest on AVeriTeC, whose sources carry a registered-domain proxy and whose audited pairs include many cross-domain reproductions with strong textual signals, and weakest on FEVER and HoVer, where the origin proxy is a Wikipedia page id. AVeriTeC also has the highest rater agreement (κ = 1.000); FEVER and HoVer have 0.877 and 0.839, respectively. Text reuse outperforms our estimator on the two page-identity corpora. These results characterize the sampled pairs. Explaining cross-corpus differences or estimating origin prevalence requires a separate sampling design.

Our probability-bin audit exposes overconfidence at high lineage scores (supplement): among 105 natural pairs assigned mean probability 0.983, annotators mark only 61.9% as same origin. Similar predicted and human group counts also coexist with ARI 0.401: getting the number of groups right does not mean putting passages in the right groups.

Threshold sensitivity. We vary the threshold retrospectively over $q \in \{0.3, 0.4, 0.5, 0.6, 0.7\}$. Natural pair F1 is 0.600, 0.593, 0.568, 0.526, and 0.506 and cluster ARI 0.383, 0.401, 0.401, 0.384, 0.389, respectively. At the reported threshold 0.5, 124 natural pair false positives coexist with 45 false negatives. On the partition task, 81 falsely merged and 53 falsely split passage pairs remain across 36 episodes. We use the same stored scores and annotation-consensus convention throughout. Moving the threshold alone does not repair natural clustering, so we retain 0.5. The supplement separates direct link errors from grouping errors. Its text-reuse grid reaches natural pair F1 0.729 at threshold 0.70, versus 0.392 at the frozen 0.80; this retrospective result does not select a new operating point.

Table 3: Aggregation on 371 held-out construction episodes. Differences compare frozen soft GLS with each original rule. The lower panel uses recovered predictions on the same claims. Rates are percentages.

Rule	**Acc.**	**Ours minus rule [CI]**	**ECE**	**False acc.**	**True**	**Mixed**
Independence GLS	91.9	−6.2 [−8.9, −3.8]	0.038	10.2	93.8	81.9
Domain dedup GLS	91.1	−5.4 [−8.1, −2.7]	0.047	10.7	92.8	79.0
Near-dup cluster GLS	87.3	−1.6 [−4.0, +0.8]	0.038	11.9	86.6	69.5
Learned hard lineage	86.0	−0.3 [−0.8, +0.0]	0.040	15.3	87.1	65.7
EvidenceGraph soft GLS	85.7	n/a	0.037	15.3	86.6	64.8
Recorded-origin hard GLS	83.8	+1.9 [+0.5, +3.5]	0.058	16.4	84.0	60.0
Recorded-origin soft GLS	86.0	−0.3 [−0.8, +0.0]	0.045	15.3	87.1	65.7
Historical prediction controls (same test claims)						
Covariance ridge	93.5	n/a	0.047	n/a	n/a	87.6
Precision ridge	93.8	n/a	0.081	n/a	n/a	87.6
Uniform mean stance	93.8	n/a	0.059	n/a	n/a	87.6

We trace threshold changes through aggregation and selection. We hold claims, splits, NLI scores, lineage fits, and the selection policy fixed, and refit only the two-parameter development calibrator for each threshold. In the construction, only 0.05% of pair probabilities lie between 0.05 and 0.95. From thresholds 0.3 to 0.7, hard-lineage aggregation stays at 86.0% accuracy and its controller at 85.4% with 6.66 passages. No test verdict changes. Across the wider 0.1 to 0.9 grid, at most one of 371 verdicts changes. This stability reflects nearly binary construction scores, not reliable natural clustering.

Natural contexts are more sensitive. Relative to 0.50, thresholds 0.30 and 0.70 change the representative set for 9.2% and 5.8% of FEVER cases, 7.2% and 6.6% of HoVer cases, and 1.0% of AVeriTeC cases. Mean context size changes by at most 0.09 passages: similar context lengths can conceal different selected evidence. Testing how these changed contexts affect answers requires new generations.

Key finding. A false link can merge passages from different origins. On natural evidence, text reuse forms groups closer to human judgments than the lineage model (ARI 0.477 versus 0.401). Changing the threshold does not repair the grouping errors.

6.3 RQ3: When does lineage help?

Uniform averaging is a critical control. Uniform averaging gives every stance equal weight. We compare it with two ridge variants,

which stabilize covariance-based weighting by adding a diagonal term to the covariance or its inverse (precision). All three methods use the same 359 development and 371 test claims, with disjoint splits. We recompute their test metrics and obtain 93.5% accuracy for covariance ridge, and 93.8% for both precision ridge and uniform mean-stance aggregation. The covariance-ridge arm retains the lineage-derived dependence matrix, unlike uniform averaging. Its ECE is 0.0471, compared with 0.0805 for precision ridge and 0.0585 for uniform. The archived development log-loss choices are covariance ridge 10 and precision ridge 10,000.

Against uniform averaging, all four overall covariance-ridge differences have paired bootstrap intervals containing zero. Precision ridge has higher Brier by 0.0042 $[0.0018, 0.0066]$ and ECE by 0.0220 $[0.0069, 0.0384]$, with unchanged accuracy. Covariance ridge also has higher Brier on mixed claims by 0.0028 $[0.0014, 0.0042]$. These unadjusted intervals condition on the saved predictions. We test whether dependence modeling improves on equal passage weights with a frozen refit (supplement). Ridge on the independence covariance makes the same 371 test decisions as covariance ridge, supporting shrinkage toward uniform weights as the explanation for the gain. Scoring errors of origin-distinct test passages still correlate at 0.354 $[0.277, 0.422]$, despite their near-zero lineage scores, versus 0.995 for same-origin pairs. The heuristic's held-out Gaussian likelihood, measured in natural-log units (nats), is 12.9 nats per passage below independence. Thus lineage alone misses substantial error dependence in this construction.

We choose all tunable quantities on the 359 development claims, freeze them, and evaluate once on the 371 test claims. Adding a shared-scorer covariance term reaches 86.5% accuracy, compared with 91.9% under independence, 85.7% for the lineage heuristic, 93.5% for covariance ridge, and 93.8% for uniform averaging. The tested shared-error term leaves an accuracy gap. Covariance ridge's log loss differs from lineage-free ridge by +0.002 $[-0.003, 0.008]$, from ridge on within-episode permuted lineage by +0.005 $[-0.003, 0.015]$, and from uniform averaging by −0.002 $[-0.006, 0.001]$. All three intervals contain zero. Our controls support shrinkage as the source of the gain. The incremental contribution of lineage remains unresolved.

Likelihood measures how well the covariance describes held-out scoring errors. Accuracy measures whether the resulting score yields the right verdict. Improving one need not improve the other. The likelihood loss of 12.9 nats per passage has a wide interval $[2.6, 34.3]$ because same-origin passages have almost perfectly correlated errors, making their covariance difficult to invert. A shared within-episode term improves likelihood by 14.1 $[3.9, 35.6]$ nats per passage without restoring the independence baseline's accuracy. Natural-evidence evaluation must likewise test covariance fit and verdict quality separately.

Table 3 reports the original lineage-based rules on matched NLI scores. Acc., False acc., True, and Mixed denote overall accuracy, false acceptance, true-claim accuracy, and mixture accuracy, all in percentages. Differences subtract each original rule from frozen soft GLS. The lower panel uses recovered historical predictions, with uncomputed contrasts and class-specific rates marked n/a. Soft GLS reaches 85.7% accuracy, while hard thresholding of the same lineage probabilities reaches 86.0%. Their paired difference is −0.3 points with a 95% interval of $[-0.8, 0.0]$. Independence GLS reaches 91.9%, making soft GLS 6.2 points worse $[-8.9, -3.8]$. It also lowers true claim retention from 93.8% to 86.6% and mixture accuracy from 81.9% to 64.8%.

Lineage weighting can suppress reliable support. In the duplication condition ($n = 129$), independence GLS reaches 93.8% accuracy and soft GLS reaches 91.5%, with ECE 0.055 and 0.071. Mixture ECE is 0.122 for independence GLS and 0.162 for soft GLS. These retained-condition summaries assess the original aggregation rules, separately from the historical ridge audit.

In the support-refute construction, several genuine supporting passages share one origin, while a fabricated counter-source is independent. Soft weighting reduces the support cluster toward one effective observation, allowing the counter-source to dominate. Recorded-origin soft weighting also reaches 86.0%, so known lineage alone cannot supply source reliability.

Key finding. A different origin does not make a source trustworthy. Equal-weight averaging reaches 93.8% accuracy versus 85.7% for soft GLS. Discounting repeated support can give a false independent source too much influence.

Active selection finds hidden evidence. Table 4 compares policies starting with five passages. The full system gains 5.7 accuracy points overall over fixed top-5, changing both selection and aggregation. At fixed independence aggregation, active selection gains 3.5 points overall and 9.7 on hidden-evidence cases. Adding lineage yields another 2.2 overall and 19.4 on hidden cases, but loses 15.2 on support-refute mixtures. Thus the full system gains 29.1 on hidden cases and loses 15.2 on mixtures relative to fixed top-5. These condition contrasts use the bounded panel. Contradiction handling recovers 18.1 mixture points relative to its own ablation, without beating fixed top-5. Effective-information scaling adds no detected gain.

Table 4 separates the frozen policy from a labeled sensitivity analysis that clips the GLS mean to $[-1, 1]$ and caps information at $n/0.05$. Primary results and natural graph scores use unbounded GLS. The bounds limit extreme scores in a retrospective check. They improve the no-active-seek ablation's accuracy and reduce calibration error in the no-contradiction ablation. Every other row's accuracy is unchanged. These numerical bounds do not establish valid covariance or calibration.

Key finding. Selection plus lineage gains 5.7 accuracy points overall, driven by hidden evidence (+29.1), but loses 15.2 on mixtures. Selection alone at fixed independence aggregation gains 3.5 overall. Covariance scaling adds no detected gain.

6.4 RQ4: What transfers to natural pools?

Relevance remains essential. Our soft policy increases origin diversity, adding 0.09 observed origins on AVeriTeC and 0.35 on HoVer (0.43 on FEVER, a legacy diagnostic) relative to passive rank. Natural top-20 pools already contain 17.1 to 17.8 distinct observed source identifiers, unlike the construction with eight planted decoys. A rule that knows the annotated evidence can gain 4.9 AVeriTeC, 2.6 FEVER, and 6.2 HoVer recall points, confirming that useful evidence remains available.

Table 4: Matched controller ablations on 371 test episodes. Fixed top-5 uses independence GLS; active policies may add three passages and use lineage-weighted GLS unless noted. The no-lineage arm also uses independence GLS. Accuracy and false acceptance (False acc.) are percentages. Condition columns omit the three paraphrase-flood episodes.

	Overall (frozen)			Overall (bounded)			Accuracy by condition (bounded)			
Method	**Accuracy**	**ECE**	**False acc.**	**Accuracy [95% CI]**	**ECE**	**False acc.**	**Copies**	**Hidden**	**Mixed**	**Passages**
Fixed top-5 (indep.)	79.5	0.083	35.6	79.5 [75.5, 83.3]	0.083	35.6	93.8	64.9	80.0	5.0
Active, no lineage	83.0	0.066	27.7	83.0 [79.2, 86.5]	0.066	27.7	93.8	74.6	80.0	7.7
EvidenceGraph active	85.2	0.032	14.7	85.2 [81.7, 88.9]	0.047	14.7	92.2	94.0	64.8	6.7
hard lineage	85.4	0.035	14.7	85.4 [81.9, 88.9]	0.036	14.7	92.2	94.0	65.7	6.7
no $\sqrt{w_{\text{eff}}}$ scaling	84.9	0.049	14.7	84.9 [81.4, 88.4]	0.048	14.7	92.2	94.0	63.8	6.7
no contradiction	80.1	0.167	26.6	80.1 [75.7, 84.1]	0.059	26.6	94.6	91.8	46.7	6.4
no active seek	52.3	0.002	100.0	75.2 [70.6, 79.2]	0.101	42.9	95.3	64.9	62.9	7.4

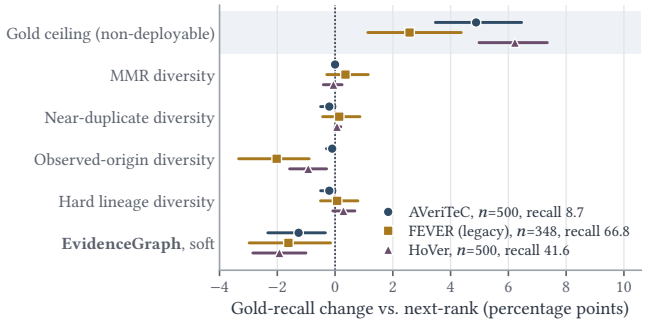

Figure 3: Gold-evidence recall change from passive next-rank selection in frozen top-20 pools. Bars: paired 95% bootstrap intervals. The legend gives rank-order recall. FEVER is a legacy logging diagnostic.

Figure 3 tests whether added diversity recovers more gold evidence. The horizontal axis gives recall changes in percentage points, with zero marking passive selection. MMR, near-duplicate diversity, and hard lineage diversity have intervals containing zero. Observed-origin diversity loses 2.0 recall points on FEVER and 0.9 on HoVer. Our soft FIND_INDEP policy loses 1.3 recall points on AVeriTeC and 1.9 on HoVer. The archived FEVER loss is 1.6 points, but its same-domain feature was logged as a constant. We retain that estimate as a legacy implementation diagnostic, not a validated FEVER transfer result. Revised page-identity outputs are unavailable for verification. On the other two benchmarks, added origins do not recover more annotated evidence, and soft selection recovers less. Passive rank provides the reference for these pools. Other retrievers and redundancy levels require separate comparisons.

Key finding. A new origin helps only if it supplies useful evidence. Natural pools already have about 17 observed origins among 20 passages. The soft selection rule finds fewer annotated evidence items on HoVer and AVeriTeC despite increasing diversity.

Probability override can harm decisions. Table 5 first compares the provenance-aware context with fixed top-5 retrieval. Accuracy changes by +1.2, −1.6, and −2.5 points on FEVER, HoVer, and AVeriTeC, with every interval containing zero. The provenance-aware context reduces FSR by 2.6 points on FEVER and 3.3 on AVeriTeC, and changes ECE by −0.059, −0.202, and +0.010, respectively. These descriptive contrasts measure the combined changes to evidence and prompt information. We next test whether confidence can improve without changing decisions.

To isolate override, we hold selected evidence and generation fixed and replace the model's support decision with the graph probability. Table 5 reports losses of 7.8 FEVER and 16.2 AVeriTeC accuracy points when using all paired generations. These comparisons include up to 34 more cases (FEVER 500 versus 498; AVeriTeC 481 versus 447; HoVer unchanged). The construction calibrator was fitted on balanced episodes with evidence present, then transferred to retrieval pools where annotated evidence can be absent. It also turns a choice among several labels into a binary support decision, retaining the best non-support label otherwise. This transfer changes the semantics of *not enough information* as well as the source of the probability.

Neutral NLI scores reveal the calibrator's default: without directional evidence, its output is determined by the fitted intercept, $\ell(0.598) = 0.645$, which favors *Supported*. Thus the binary calibrator can favor support even without supporting evidence. Quantifying how often this default causes multiclass errors requires a separate audit.

To improve confidence without changing decisions, we fit a temperature on half of each endpoint population. On the other half, log loss falls in all nine benchmark-policy combinations and Brier score in seven with intervals excluding zero (supplement). Both measure probability error, with lower values better. Fitting on other benchmarks does not transfer reliably. Context-selection ECE changes measure a different intervention.

Table 6 reports the decision-preserving result for the EvidenceGraph context. A fixed hash of the case identifier splits each complete-case population into calibration and evaluation halves. Each benchmark and arm receives its own fitted temperature. The inputs are label probabilities stated in the model's structured answer, not token likelihoods or graph-override probabilities. Deterministic no-evidence abstentions remain unchanged, so accuracy, macro-F1, false support, and abstention rates are identical before and after calibration.

The intervals use 2,000 paired claim bootstrap resamples. These evaluation halves differ from the full complete-case populations in Table 5, so their calibration values describe a different set of answers.

HoVer's Brier interval contains zero despite its favorable estimate. Swapping halves and two-fold cross-fitting preserve the direction of

Table 5: End-to-end results on frozen retrieval. Context deltas use fixed RAG, while override deltas use context and pair all available generations (FEVER 500, HoVer 494, AVeriTeC 481). F1 is macro-F1, and false support is a percentage.

Method	Accuracy [95% CI]	F1	ECE	False support
FEVER (n=498)				
Closed-book	56.2 [52.2, 60.6]	54.6	0.254	13.8
Fixed top-5 RAG	67.1 [62.9, 70.9]	64.9	0.275	14.7
EvidenceGraph context	68.3 [64.1, 72.3]	65.7	0.215	12.2
Δ context (n=498)	+1.2 [−1.2,+3.8]	+0.8	−0.059 [−0.087,−0.036]	−2.6 [−4.7,−0.6]
EvidenceGraph with override	60.4 [56.4, 64.5]	52.4	0.199	50.3
Δ override (n=500)	−7.8 [−11.6,−4.2]	−13.2	−0.017 [−0.054,+0.018]	+38.2 [+32.9,+43.3]
HoVer (n=494)				
Closed-book	55.3 [50.6, 59.7]	50.5	0.279	9.7
Fixed top-5 RAG	55.9 [51.2, 60.3]	52.7	0.380	14.4
EvidenceGraph context	54.3 [49.4, 58.9]	50.3	0.178	13.6
Δ context (n=494)	−1.6 [−5.3,+1.6]	−2.4	−0.202 [−0.234,−0.160]	−0.8 [−4.8,+3.3]
EvidenceGraph with override	54.3 [50.0, 58.7]	50.4	0.149	72.5
Δ override (n=494)	+0.0 [−6.5,+6.9]	+0.1	−0.029 [−0.097,+0.035]	+58.9 [+52.3,+65.1]
AVeriTeC (n=447)				
Closed-book	45.6 [40.9, 50.1]	35.1	0.250	2.4
Fixed top-5 RAG	69.4 [65.1, 73.6]	47.9	0.180	13.5
EvidenceGraph context	66.9 [62.6, 71.1]	49.7	0.190	10.2
Δ context (n=447)	−2.5 [−5.6,+0.7]	+1.8	+0.010 [−0.021,+0.040]	−3.3 [−5.6,−1.1]
EvidenceGraph with override	50.8 [46.3, 55.7]	33.3	0.338	45.9
Δ override (n=481)	−16.2 [−20.2,−12.5]	−16.0	+0.154 [+0.118,+0.194]	+36.2 [+31.1,+41.5]

Table 6: Held-out temperature calibration of the EvidenceGraph context. Positive Brier reduction is better; every verdict is unchanged.

	ECE		Brier reduction
Benchmark	Before	After	[95% CI]
FEVER	0.195	0.081	0.049 [0.010, 0.088]
HoVer	0.172	0.105	0.041 [−0.003, 0.086]
AVeriTeC	0.207	0.105	0.036 [0.018, 0.054]

the calibration changes. Fitting on the other two benchmarks raises Brier by 0.190 $[0.103, 0.268]$ on AVeriTeC fixed retrieval. Flooring exact zero probabilities at 10^{-6} explains part of the log-loss improvement, without affecting Brier or ECE. We establish calibration gains within each benchmark and test their transfer separately.

Key finding. Temperature calibration improves held-out log loss within each benchmark without changing verdicts. Graph-based verdict replacement can hurt accuracy, and calibration learned on other benchmarks transfers unreliably.

7 Discussion and Scope

Our results separate shared origin from trustworthiness. In the support-refute construction, discounting genuine support gives a fabricated independent source greater influence. Correct lineage leaves this reliability problem intact. We therefore recommend validating links and groups against blinded target-domain judgments before discarding evidence. Error dependence requires residual measurements because unrelated sources share the claim, retriever, and verifier.

Diversity helps when eight same-origin decoys hide useful evidence. Natural pools already contain about 17 observed origins among 20 passages, so increasing that count can leave the evidence gap untouched. A controller must find information that completes or corrects the evidence set. Diversity is one cue for that decision.

Provenance can help readers inspect likely copies while preserving the answer model's verdict. Replacement needs target-domain validation. Positive temperature calibration preserves labels, while held-out tests assess confidence reliability. Transfer must preserve label meanings and abstention behavior.

Our stratified audit uses one hosted answer model and frozen pools. Threshold changes from 0.5 to 0.3 or 0.7 alter 1 to 9% of natural contexts, with answer effects still untested. Broader tests need natural-lineage annotations and more answer models. By fixing retrieval, scoring, and budgets, we locate failures in grouping, selection, verdict replacement on fixed generations, and held-out calibration. Report retained cases for each comparison.

8 Conclusion

With EvidenceGraph, we distinguish source lineage, shared scoring errors, and verdict quality through separate tests. Copies raise false-claim acceptance in the copy-count experiment, while lineage weighting can suppress reliable support. Selection improves accuracy when repeated distractors hide useful evidence. On natural retrieval pools, its accuracy gains remain unresolved. Our findings support provenance for evidence inspection and acquisition. Matched tests identify when aggregation and verdict changes help or harm verification.

Acknowledgments

We acknowledge Ho Chi Minh City University of Technology (HCMUT), VNU-HCM, for supporting this study.

Ethical Considerations

Information integrity and dual use. Repetition can inflate confidence through routine syndication or deliberate source flooding [41, 49]. We conduct the intervention within a frozen corpus. Shared origin does not establish truth. Discounting similar reports can suppress evidence even when the reports are independent. Soft discounting reduces constructed true-claim retention by 7.2 points. Soft independence selection reduces gold-evidence recall on AVeriTeC and HoVer. The retained FEVER estimate has a logging limitation. Constructed false acceptance across lineage-certainty quartiles, whose top two share one certainty range, is 40.5%, 17.5%, 0.0%, and 3.9% (93/93/93/92 total episodes; each rate uses only false claims). These harms motivate showing where evidence comes from and explicitly leaving a claim unresolved when evidence is insufficient. Repetition control addresses one vulnerability: a single convincing fabrication can still determine the answer.

Generated content and licensing. Synthetic paraphrases and fabricated counter-sources carry a `synthetic` flag, reserved `.invalid` domain, and origin identifier. The full construction remains unreleased. The public artifact covers code and audits. Future redistribution must retain these markers and avoid indexable false content. FEVER, AVeriTeC, HoVer, and the 2017 English Wikipedia snapshot retain their source licenses. FEVER-derived text retains attribution and share-alike obligations. The AVeriTeC Web knowledge store is not redistributed. A future EvidenceGraph-20K dataset release requires its own licensing review.

Human lineage audit. Two final-year computer-science undergraduates at HCMUT agreed to participate voluntarily and without compensation. HCMUT did not require formal ethics review for this annotation activity. Each labeled 384 pairs and partitioned 30 episodes. Task files hid corpus, sampling stratum, provenance condition, and model scores. The union contains 640 pair items and 48 episodes. On 128 shared pairs, three-way kappa is 0.951 (agreement 0.977). On 126 jointly decided pairs, kappa is 0.966 (agreement 0.984). Mean inter-rater ARI is 0.983 over 12 shared partitions. Annotators knew the evaluation purpose and that human labels would take precedence on disagreement. No crowdworkers or sensitive personal data are involved.

On 469 decided natural pairs, estimator precision/recall/F1 are 0.472/0.712/0.568. ARI is 0.401 on 36 natural partitions. F1 and ARI miss the artifact's internal release gates of 0.85 and 0.80. These are internal quality gates for release, separate from conference requirements.

Availability and compute. The public artifact at https://github.com/lexuanbach/wsdm2027-evidencegraph-audit contains code for every pipeline stage, checksum-bound summaries, deidentified historical predictions, and retrospective audit inputs. One script rebuilds the original tables and Figure 3. The repository also provides the supplement and companion `camera-ready-analysis.zip`, which holds later analysis records and outputs. Full pipeline reruns require raw construction and frozen lineage weights that were recovered privately but are not in the public artifact. Recovery enabled the new supplement analyses, but the full EvidenceGraph-20K dataset remains unreleased. Local experiments used an Apple-silicon workstation, and endpoint answering used a hosted model. Complete runtime and token accounting is absent from the retained summaries. Energy use, carbon emissions, and per-method efficiency remain separate measurement tasks.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. 2023. SemDeDup: Data-efficient learning at web-scale through semantic deduplication. arXiv:2303.09540 [cs.LG] https://arxiv.org/abs/2303.09540

[2] Mubashara Akhtar, Rami Aly, Yulong Chen, Zhenyun Deng, Michael Schlichtkrull, Chenxi Whitehouse, and Andreas Vlachos. 2025. The 2nd Automated Verification of Textual Claims (AVeriTeC) Shared Task: Open-weights, Reproducible and Efficient Systems. In *Proceedings of the Eighth Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics, Vienna, Austria, 201–223. doi:10.18653/v1/2025.fever-1.15

[3] Anthropic. 2025. Introducing Claude Haiku 4.5. https://www.anthropic.com/news/claude-haiku-4-5

[4] Anthropic. 2026. Introducing Sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6 Published February 17, 2026.

[5] Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. In *International Conference on Learning Representations*. https://openreview.net/forum?id=hSyW5go0v8

[6] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Massimiliano Ciaramita, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, Tom Kwiatkowski, Ji Ma, Jianmo Ni, Lierni Sestorain Saralegui, Tal Schuster, William W. Cohen, Michael Collins, Dipanjan Das, Donald Metzler, Slav Petrov, and Kellie Webster. 2022. Attributed Question Answering: Evaluation and Modeling for Attributed Large Language Models. arXiv:2212.08037 [cs.CL] https://arxiv.org/abs/2212.08037

[7] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 335–336. doi:10.1145/290941.291025

[8] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information Credibility on Twitter. In *Proceedings of the 20th International Conference on World Wide Web*. Association for Computing Machinery, Hyderabad, India, 675–684. doi:10.1145/1963405.1963500

[9] Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. Rich Knowledge Sources Bring Complex Knowledge Conflicts: Recalibrating Models to Reflect Conflicting Evidence. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2292–2307. doi:10.18653/v1/2022.emnlp-main.146

[10] Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement* 20, 1 (1960), 37–46. doi:10.1177/001316446002000104

[11] Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-Verification Reduces Hallucination in Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 3563–3578. doi:10.18653/v1/2024.findings-acl.212

[12] Xin Luna Dong, Laure Berti-Équille, and Divesh Srivastava. 2009. Integrating Conflicting Data: The Role of Source Dependence. *Proceedings of the VLDB Endowment* 2, 1 (2009), 550–561. doi:10.14778/1687627.1687690

[13] Xin Luna Dong, Evgeniy Gabrilovich, Kevin Murphy, Van Dang, Wilko Horn, Camillo Lugaresi, Shaohua Sun, and Wei Zhang. 2015. Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources. *Proceedings of the VLDB Endowment* 8, 9 (2015), 938–949. doi:10.14778/2777598.2777603

[14] Saadeldine Eletter, Ruihong Zeng, Yuxia Wang, Maxim Panov, Aleksandr Rubashevskii, and Preslav Nakov. 2026. MIRAGE: Defending Long-Form RAG Against Misinformation Pollution. arXiv:2607.05069 [cs.CL] https://arxiv.org/abs/2607.05069

[15] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling Large Language Models to Generate Text with Citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Singapore, 6465–6488. doi:10.18653/v1/2023.emnlp-main.398

[16] Shuzhi Gong, Richard Sinnott, Jianzhong Qi, Cecile Paris, Preslav Nakov, and Zhuohan Xie. 2026. Multi-Sourced, Multi-Agent Evidence Retrieval for Fact-Checking. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*. ACM, Melbourne, VIC, Australia, 437–447. doi:10.1145/3805712.3809686

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*. PMLR, Sydney, Australia, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

[18] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*. PMLR, Virtual, 3929–3938. https://proceedings.mlr.press/v119/guu20a.html

[19] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing. In *International Conference on Learning Representations*. https://openreview.net/forum?id=sE7-XhLxHA

[20] Sture Holm. 1979. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics* 6, 2 (1979), 65–70.

[21] S M Asif Hossain, Ruksat Khan Shayoni, and M. F. Mridha. 2026. Evidential-RAG: Quantifying and Mitigating Information Conflict in Multi-Source Retrieval-Augmented Generation via Evidential Deep Learning. arXiv:2607.10491 [cs.LG] https://arxiv.org/abs/2607.10491

[22] Lawrence Hubert and Phipps Arabie. 1985. Comparing Partitions. *Journal of Classification* 2, 1 (1985), 193–218. doi:10.1007/BF01908075

[23] Jeongyeon Hwang, Junyoung Park, Hyejin Park, Dongwoo Kim, Sangdon Park, and Jungseul Ok. 2025. Retrieval-Augmented Generation with Estimation of Source Reliability. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China, 34279–34303. doi:10.18653/v1/2025.emnlp-main.1738

[24] Gautier Izacard and Edouard Grave. 2021. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Association for Computational Linguistics, Online, 874–880. doi:10.18653/v1/2021.eacl-main.74

[25] Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A Dataset for Many-Hop Fact Extraction And Claim Verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 3441–3460. doi:10.18653/v1/2020.findings-emnlp.309

[26] Zhuoran Jin, Pengfei Cao, Yubo Chen, Kang Liu, Xiaojian Jiang, Jiexin Xu, Qiuxia Li, and Jun Zhao. 2024. Tug-of-War between Knowledge: Exploring and Resolving Knowledge Conflicts in Retrieval-Augmented Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. ELRA and ICCL, Turin, Italy, 16867–16878. https://aclanthology.org/2024.lrec-main.1466/

[27] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., Virtual, 9459–9474. https://proceedings.neurips.cc/paper_files/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

[28] Yaliang Li, Jing Gao, Chuishi Meng, Qi Li, Lu Su, Bo Zhao, Wei Fan, and Jiawei Han. 2016. A Survey on Truth Discovery. *ACM SIGKDD Explorations Newsletter* 17, 2 (2016), 1–16. doi:10.1145/2897350.2897352

[29] Junli Liang, Pengfei Zhou, Wangqiu Zhou, Wenjie Qing, Qi Zhao, Ziwen Wang, Qi Song, and Xiangyang Li. 2026. SentGraph: Hierarchical Sentence Graph for Multi-hop Retrieval-Augmented Question Answering. arXiv:2601.03014 [cs.CL] https://arxiv.org/abs/2601.03014

[30] Mixedbread. 2024. mxbai-embed-large-v1 Model Card. https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1

[31] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. WebGPT: Browser-assisted question-answering with human feedback. arXiv:2112.09332 [cs.CL] https://arxiv.org/abs/2112.09332

[32] Atharv Naphade. 2026. Rational Synthesizers or Heuristic Followers? Analyzing LLMs in RAG-based Question-Answering. In *Findings of the Association for Computational Linguistics: ACL 2026*. Association for Computational Linguistics, San Diego, California, United States, 40293–40311. doi:10.18653/v1/2026.findings-acl.2003

[33] Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Yang Wang. 2023. On the Risk of Misinformation Pollution with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics, Singapore, 1389–1403. doi:10.18653/v1/2023.findings-emnlp.97

[34] Jingxi Qiu, Zeyu Han, and Cheng Huang. 2026. SURE-RAG: Sufficiency and Uncertainty-Aware Evidence Verification for Selective Retrieval-Augmented Generation. arXiv:2605.03534 [cs.CL] https://arxiv.org/abs/2605.03534

[35] Qwen Team. 2024. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

[36] Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring Attribution in Natural Language Generation Models. *Computational Linguistics* 49, 4 (2023), 777–840. doi:10.1162/coli_a_00486

[37] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval* (2009). doi:10.1561/1500000019

[38] Michael Schlichtkrull, Yulong Chen, Chenxi Whitehouse, Zhenyun Deng, Mubashara Akhtar, Rami Aly, Zhijiang Guo, Christos Christodoulopoulos, Oana Cocarascu, Arpit Mittal, James Thorne, and Andreas Vlachos. 2024. The Automated Verification of Textual Claims (AVeriTeC) Shared Task. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*. Association for Computational Linguistics, Miami, Florida, USA, 1–26. doi:10.18653/v1/2024.fever-1.1

[39] Michael Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., New Orleans, Louisiana, USA, 65128–65167. doi:10.52202/075280-2842

[40] Sentence Transformers. 2021. Cross-Encoder for MS MARCO: MiniLM-L6-v2. https://huggingface.co/cross-encoder/ms-marco-MiniLM-L6-v2

[41] Jinyan Su, Jin Peng Zhou, Zhengxin Zhang, Preslav Nakov, and Claire Cardie. 2024. Towards More Robust Retrieval-Augmented Generation: Evaluating RAG Under Adversarial Poisoning Attacks. arXiv:2412.16708 [cs.IR] https://arxiv.org/abs/2412.16708

[42] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 809–819. doi:10.18653/v1/N18-1074

[43] Alexander Wan, Eric Wallace, and Dan Klein. 2024. What Evidence Do Language Models Find Convincing?. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 7468–7484. doi:10.18653/v1/2024.acl-long.403

[44] Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arık. 2025. Astute RAG: Overcoming Imperfect Retrieval Augmentation and Knowledge Conflicts for Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vienna, Austria, 30553–30571. doi:10.18653/v1/2025.acl-long.1476

[45] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. arXiv:2212.03533 [cs.CL] https://arxiv.org/abs/2212.03533

[46] Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts. In *International Conference on Learning Representations*. https://proceedings.iclr.cc/paper_files/paper/2024/hash/99261adc8a6356b38bcf999bba9a26dc-Abstract-Conference.html

[47] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=WE_vluYUL-X

[48] Yixiao Zeng, Tianyu Cao, Danqing Wang, Xinran Zhao, Zimeng Qiu, Morteza Ziyadi, Tongshuang Wu, and Lei Li. 2025. RARE: Retrieval-Aware Robustness Evaluation for Retrieval-Augmented Generation Systems. arXiv:2506.00789 [cs.CL] https://arxiv.org/abs/2506.00789

[49] Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2025. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. In *34th USENIX Security Symposium (USENIX Security 25)*. USENIX Association, Seattle, Washington, 3827–3844. https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag