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

Supplementary retrospective audit

Xuan-Bach Le Phat T. Tran-Truong Son Ha Xuan Duy Nguyen

WSDM 2027, Paper 2148

Correspondence: lexuanbach@hcmut.edu.vn
Assigned paper DOI: https://doi.org/10.1145/3862322.3863670

Scope and annotation convention

The companion archive checks the exact-covariance identities with `revision/numerical-check.py` and records errors in `revision/numerical-check.json`.

The threshold grids and historical covariance audit use archived outputs with the original annotators, fits, and samples. Separate sections report the six-model copy-count experiment and the covariance, downstream-threshold, and held-out calibration analyses on recovered inputs. These analyses have distinct protocols and records. The historical claims remain withdrawn. We report every threshold in the declared grids and keep audit labels out of operating-point selection. The main manuscript retains the lineage threshold $q \geq 0.5$ and text-reuse threshold 0.80.

Pair evaluation retains the original agreement convention: a pair enters the analysis only if all available ratings agree on SAME_ORIGIN or INDEPENDENT. Unsure or disagreeing pairs are excluded. Partition evaluation retains the original primary-A convention (A when available, otherwise the available annotator), preserving the original annotation-selection rule. The full audit contains 640 unique pair items and 48 partition tasks. Natural pair metrics use 469 decided pairs and natural partition metrics use 36 episodes. Constructed and natural strata are reported separately. The metrics characterize these audited samples. Population prevalence requires a different sampling design.

Threshold results

A false link is a predicted positive pair with an independent human label, while a missed link is a predicted negative pair with a same-origin label. Clustering joins edges above threshold and takes connected components. Its false-merge and false-split counts range over all pairs inside the partition tasks, including pairs connected indirectly. They therefore have a different denominator from the directly annotated pair decisions. ARI is averaged over partition tasks rather than computed by pooling unrelated episodes.

Pair decisions at fixed lineage thresholds

Scope	Threshold	Precision	Recall	F1	False links	Missed links
Natural	0.3	0.454	0.885	0.600	166	18
Natural	0.4	0.468	0.808	0.593	143	30
Natural	0.5	0.472	0.712	0.568	124	45
Natural	0.6	0.479	0.583	0.526	99	65
Natural	0.7	0.506	0.506	0.506	77	77
Constructed	0.3	0.962	1.000	0.981	3	0
Constructed	0.4	0.962	1.000	0.981	3	0
Constructed	0.5	0.961	0.974	0.967	3	2
Constructed	0.6	0.986	0.961	0.973	1	3
Constructed	0.7	0.985	0.868	0.923	1	10

Partitions after transitive closure

Scope	Threshold	Mean ARI	False merged pairs	False split pairs
Natural	0.3	0.383	137	53
Natural	0.4	0.401	82	53
Natural	0.5	0.401	81	53
Natural	0.6	0.384	74	54
Natural	0.7	0.389	63	54
Constructed	0.3	1.000	0	0
Constructed	0.4	1.000	0	0
Constructed	0.5	1.000	0	0
Constructed	0.6	1.000	0	0
Constructed	0.7	1.000	0	0

Pair decisions at fixed text-reuse thresholds

Scope	Threshold	Precision	Recall	F1	False links	Missed links
Natural	0.7	0.867	0.628	0.729	15	58
Natural	0.8	0.907	0.250	0.392	4	117
Natural	0.9	1.000	0.173	0.295	0	129
Constructed	0.7	0.985	0.842	0.908	1	12
Constructed	0.8	0.983	0.750	0.851	1	19
Constructed	0.9	0.980	0.658	0.787	1	26

Interpretation and reproduction boundary

At $q = 0.5$, natural pair F1 is 0.568 and mean ARI is 0.401, reproducing the main paper after rounding. Lower thresholds trade additional false links for fewer missed links. Pair F1 and partition ARI respond differently because a single link can merge whole components. The grid quantifies the pair-versus-cluster tradeoff at each threshold. Threshold selection and downstream evaluation are separate steps from this retrospective comparison.

The audit package contains numeric consensus labels and frozen predictions without passage text, original item identifiers, or annotation blinding keys. Its script recomputes all grid entries. Seven original table outputs (including stratum and natural-recall fragments) can be regenerated from stored summaries; later camera-ready tables and editorial labels are separate. The updated manifest verifies 19 source/table bindings. The original manifest is retained separately because three bindings referred to an older annotation-report hash. The current bindings identify the reproducible report, while the retained manifest documents the earlier version. The artifact distinguishes current saved-output checks from legacy construction and endpoint tests. Its test-support inventory records missing inputs, retired claim checks, and optional dependencies. The saved-output suite validates these summaries. Full construction validation requires the underlying inputs and weights.

The retained episodes, passage-level NLI outputs, full lineage predictions, and frozen lineage weights were later

recovered from an offline backup, with every recorded sha256 matching. The section on recovered-input analyses reports the downstream threshold audit, the covariance refit, and held-out answer calibration that these inputs allow. The public package still supports saved-output reproduction only. Later analysis records are supplied in `camera-ready-analysis.zip`, available in the public repository. All `revision/` paths below refer to files inside that archive. Origin probability and verifier-error correlation require separate estimation. The historical residual-correlation number lacks a recoverable source summary and is not used as verified evidence.

Historical covariance audit

Three recovered prediction files cover the same 359 development and 371 test claims, with no claim overlap between splits. Their hashes and label/corruption joins were verified against the frozen controller outputs. We recompute metrics with the archived fits held fixed. Development log loss selected covariance ridge 10 and precision ridge 10,000. The test results below assess those historical selections and play no role in fitting or configuration choice.

Arm	Split	Acc.	Log loss	Brier	ECE	Mean n_{eff}
Covariance ridge	dev	0.9331	0.2264	0.0630	0.0541	1.67
Covariance ridge	test	0.9353	0.1982	0.0543	0.0471	1.70
Precision ridge	dev	0.9248	0.3757	0.0706	0.0657	3561930.00
Precision ridge	test	0.9380	0.2133	0.0583	0.0805	191599.27
Uniform	dev	0.9276	0.2280	0.0633	0.0574	18.91
Uniform	test	0.9380	0.2005	0.0542	0.0585	19.15

Effective-evidence values retain the archived internal scale. The large precision-ridge value reflects that scale, rather than a count of independent observations or acquired evidence.

For each comparison we draw 1,000 paired bootstrap samples of claims with replacement (seed 217), using identical claim indices for both arms. The table gives ridge minus uniform and percentile 95% intervals. These exploratory intervals are unadjusted for multiple comparisons. They condition on the archived predictions and do not include refitting or selection uncertainty. Accuracy, Brier, log loss and ten-bin ECE use the original metric conventions.

Arm / test subset	Metric	Difference	95% interval
Covariance ridge / All, $n = 371$	Accuracy	-0.0027	$[-0.0108, +0.0000]$
Covariance ridge / All, $n = 371$	Log loss	-0.0023	$[-0.0061, +0.0012]$
Covariance ridge / All, $n = 371$	Brier	$+0.0001$	$[-0.0009, +0.0013]$
Covariance ridge / All, $n = 371$	ECE	-0.0114	$[-0.0173, +0.0003]$
Covariance ridge / Mixed, $n = 105$	Accuracy	$+0.0000$	$[+0.0000, +0.0000]$
Covariance ridge / Mixed, $n = 105$	Log loss	$+0.0063$	$[-0.0001, +0.0115]$
Covariance ridge / Mixed, $n = 105$	Brier	$+0.0028$	$[+0.0014, +0.0042]$
Covariance ridge / Mixed, $n = 105$	ECE	$+0.0020$	$[-0.0080, +0.0150]$
Precision ridge / All, $n = 371$	Accuracy	$+0.0000$	$[+0.0000, +0.0000]$
Precision ridge / All, $n = 371$	Log loss	$+0.0129$	$[-0.0038, +0.0269]$
Precision ridge / All, $n = 371$	Brier	$+0.0042$	$[+0.0018, +0.0066]$
Precision ridge / All, $n = 371$	ECE	$+0.0220$	$[+0.0069, +0.0384]$
Precision ridge / Mixed, $n = 105$	Accuracy	$+0.0000$	$[+0.0000, +0.0000]$
Precision ridge / Mixed, $n = 105$	Log loss	-0.0017	$[-0.0507, +0.0316]$
Precision ridge / Mixed, $n = 105$	Brier	$+0.0026$	$[-0.0037, +0.0085]$
Precision ridge / Mixed, $n = 105$	ECE	$+0.0390$	$[-0.0074, +0.0598]$

All four overall covariance-ridge contrasts with uniform include zero. Precision ridge has higher overall Brier and ECE with intervals above zero, at unchanged accuracy. On mixed claims, covariance ridge also has higher Brier with an interval above zero. Uniform aggregation is therefore the critical control: gains over weaker baselines leave the lineage-to-covariance mapping unvalidated.

The archived summary reports mean L_1 distance of normalized weights from uniform of 0.08433 for covariance ridge and 0.0003233 for precision ridge. These weight statistics, the shared-scorer grid, independence comparisons and dependence-removal diagnostics remain *summary-only evidence*. Recomputing them requires raw episodes

and passage-level predictions. The public numeric package separates these historical summaries from recomputed metrics and bootstrap results.

Natural probability-bin audit

Table S1 groups the audit by predicted lineage probability. Share denotes each bin's fraction of all natural pairs, while Same origin counts audited positives. Sampling concentrates on high scores: the lowest bin contains 98.1% of natural pairs but yields two audited same-origin pairs. The design therefore supports rates for the audited sample, while leaving population recall unidentified. Calibration error is pronounced in the high-score region that covariance weighting treats as strong dependence. Among 105 natural pairs assigned mean probability 0.983, annotators mark only 61.9% as same origin. Overall natural Brier score is 0.241, compared with 0.021 on the constructed sample. These calibration errors motivate validating a downstream intervention before using lineage probabilities as covariance entries. At the partition level, the human reference contains 18.31 groups per natural episode and our estimator predicts 18.39. Their similar counts coexist with ARI 0.401 because incorrect merges and splits occur on different passages. Matching the number of sources is therefore not sufficient to recover their lineage structure.

Table S1: Natural-lineage audit by predicted-probability bin. Share is the fraction of all natural pairs. Same origin counts audited positives.

Score bin	Pop. share	Audited pairs	Same-origin pairs
$[0.0, 0.2)$	98.1%	116	2
$[0.2, 0.5)$	0.5%	118	43
$[0.5, 0.8)$	0.3%	119	43
$[0.8, 1.0]$	1.1%	116	68

All 284,981 natural pairs. The lowest bin holds most of the population but yielded two audited same-origin pairs, so population-level recall is not identified by this design, and a population-level rate would require design weights.

Copy-count experiment across model scales

This camera-ready experiment has its own frozen protocol and outputs. The historical copy-count and residual-correlation claims remain withdrawn. We rebuilt the 20,000-claim selection from the official FEVER 1.0 training file with the recovered, unmodified selection code (seed 103). All 730 claim identifiers in our saved predictions matched the rebuild in split, label, and corruption type, with no mismatch. The recovered duplication script then fixes the sample as the first 40 refuted test claims under a seeded hash order. We froze the protocol, sample, and code hashes before any model output.

Each claim appears in a context of 20 slots. The repeated condition places 1, 2, 5, 10, or 20 verbatim copies of the false claim at seeded positions and fills the other slots with neutral text of matched length. The representative condition shows one copy with an instruction that sources may repeat, and its input is identical at every copy count. We ran Qwen2.5 models with 1.5B, 3B, 7B, 14B, and 72B parameters on Ollama 0.30.10 (`https://ollama.com`) at temperature 0 and seed 103, using a forced TRUE or FALSE answer token and its log probability. Claude Sonnet 4.6 received the identical binary prompt through the Anthropic API and returns no log probabilities. All 2,400 binary calls returned a valid answer. The saved compute ledger estimates API spend at USD 1.35.

Table S2 reports acceptance of the false claim. Every model accepted more claims with twenty copies than with one, and every claim-clustered trend per unit of log copy count is positive after Holm adjustment. Claude Sonnet 4.6 is not monotone: it accepted 16, 12, 13, 12, and 29 claims across the five copy counts. For the Qwen models, log-probability confidence in the false claim rose in step. The representative condition lowered twenty-copy acceptance for four models, but not for the 3B and 14B models, where it matched the repeated condition.

A JSON prompt that asks for a verdict and a self-reported probability is a separate measure. Under it, acceptance rose for the 3B model (8 to 19 claims) and the 72B model (9 to 18, Holm-adjusted $p = 0.023$) and stayed flat otherwise. Self-reported probability did not rise with copy count for any model. Schema failures affect 271 of 400 responses for the 1.5B model and 87 of 400 for Claude Sonnet 4.6, limiting inference from these self-report runs. The design shows no refuting evidence, and source identifiers name the repeated and neutral roles, as in the original script. A positive trend across six models does not establish a law in parameter count. Protocol, raw responses, spend ledger, and analysis are in `revision/eg13`.

Table S2: False-claim acceptance among 40 refuted FEVER claims by copy count. Rep. is the representative condition at twenty copies, and Rep. gain is twenty-copy acceptance minus Rep. Slope is the change in acceptance per unit of log copy count with claim-clustered 95% CIs. Conf. is the mean log-probability confidence in the false claim at one and twenty copies.

Model	1 copy	20 copies	Rep.	Rep. gain [95% CI]	Slope [95% CI]	Holm p	Conf. 1 to 20
Qwen2.5 1.5B	27	40	35	0.125 [0.025, 0.225]	0.099 [0.051, 0.147]	<0.001	0.65 to 0.99
Qwen2.5 3B	1	13	13	0.000 [−0.075, 0.075]	0.106 [0.056, 0.156]	<0.001	0.02 to 0.35
Qwen2.5 7B	6	12	7	0.125 [0.025, 0.225]	0.053 [0.016, 0.090]	0.012	0.17 to 0.30
Qwen2.5 14B	1	7	7	0.000 [−0.075, 0.075]	0.046 [0.009, 0.083]	0.015	0.04 to 0.18
Qwen2.5 72B	11	20	8	0.300 [0.175, 0.450]	0.092 [0.041, 0.144]	0.003	0.31 to 0.50
Claude Sonnet 4.6	16	29	7	0.550 [0.400, 0.700]	0.082 [0.034, 0.131]	0.001	n/a

Recovered-input analyses

After acceptance we recovered the original validated subset, lineage predictions, NLI scores, and lineage fits from an offline backup. All five files with a recorded sha256 match it exactly. Rescoring all 142,003 development and test pairs with the recovered fits reproduces the stored predictions exactly, and the threshold audit gives natural F1 0.568 and ARI 0.401 at 0.5. The saved table-regeneration audit reproduces the aggregation, selection, and endpoint summaries without difference. The three analyses below are new. Paired claim bootstrap intervals use 2,000 resamples.

Downstream threshold sensitivity. After the reported results were seen, we varied only the hard lineage threshold t in $h_{ij} = \mathbb{1}[q_{ij} \geq t]$ from 0.1 to 0.9. Claims, splits, NLI scores, lineage fits, the controller, the near-duplicate threshold 0.80, and the operating point 0.50 stayed fixed. Each threshold received its own two-parameter development calibrator. On the construction, lineage probabilities are almost binary, with 0.05% of pairs between 0.05 and 0.95. From 0.3 to 0.7, learned hard lineage keeps 86.0% test accuracy and the hard-lineage controller keeps 85.4% with 6.66 passages, and no test decision changes. Brier and log loss move by at most 0.0002 and 0.0009, ECE by at most 0.005, and every interval contains zero. Across the full grid at most one of 371 test decisions changes. In the natural endpoint pools the threshold changes the answer context. Moving from 0.50 to 0.30 or 0.70 changes the representative set for 9.2% or 5.8% of FEVER cases, 7.2% or 6.6% of HoVer cases, and 1.0% of AVeriTeC cases, and shifts mean context size by at most 0.09 passages. We did not generate new answers for these contexts, so the effect on natural answers remains unmeasured. The audit does not select a threshold.

Covariance refit. We refitted covariance models on the 359 development claims and evaluated them once on the 371 test claims after freezing every choice (`revision/eg11-refit-protocol.json`). Independence, the deployed heuristic $r_{ij} = q_{ij}$, a shared-scorer term, and covariance ridge reach 91.9%, 85.7%, 86.5%, and 93.5% test accuracy, and uniform averaging reaches 93.8%. Ridge applied to the independence covariance makes the same 371 decisions as covariance ridge. In log loss, covariance ridge 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 by -0.002 $[-0.006, 0.001]$. Its benefit is therefore shrinkage toward uniform weights rather than lineage content. Without ridge, the correctly aligned heuristic is worse than independence GLS (the original Table 3 "Count arm) by 6.2 points $[3.5, 8.9]$. This refit uses 2,000 bootstrap resamples, versus 1,000 in Table 3. Its 3.0-point loss relative to the permuted heuristic has interval $[0.0, 5.9]$ and is unresolved. A direct test of the covariance explains this negative result. On test pairs, edge errors of same-origin passages correlate at 0.995 $[0.990, 0.998]$, consistent with their near-unit lineage scores, but errors of origin-distinct passages correlate at 0.354 $[0.277, 0.422]$ despite their near-zero lineage scores. The mapping uses predicted q_{ij}, not a gold same-origin indicator. On held-out errors the heuristic covariance has lower Gaussian likelihood than independence by 12.9 nats per passage $[2.6, 34.3]$, and a shared within-episode term raises it by 14.1 $[3.9, 35.6]$. These likelihood intervals are wide because same-origin blocks are nearly singular. In the construction, lineage identifies which passages copy one another, but it does not supply the scoring-error covariance.

Held-out answer calibration. The backup holds the answer model's own stated label probabilities for closed-book answering, fixed top-5 retrieval, and the EvidenceGraph context. These are probabilities stated in the model's structured answer, not token likelihoods. Graph probabilities from the override arm were not used. A fixed hash of the case identifier split each complete-case endpoint population into a calibration half and an evaluation half, and one temperature per benchmark and arm was fitted on the calibration half (`revision/a7-calibration`). Deterministic no-evidence abstentions stayed unchanged and every label stayed the model's argmax, so accuracy, macro-F1, false support, and abstention rates do not change. On the evaluation halves, calibration lowered log loss in all nine benchmark and arm cells and Brier in seven, with intervals excluding zero. For the EvidenceGraph context, Brier fell by 0.049 $[0.010, 0.088]$ on FEVER, 0.041 $[-0.003, 0.086]$ on HoVer, and 0.036 $[0.018, 0.054]$ on AVeriTeC, and ECE fell from 0.195 to 0.081, 0.172 to 0.105, and 0.207 to 0.105. Fitted temperatures above one show that the stated probabilities are overconfident. On HoVer fixed retrieval the temperature reached its upper bound, which leaves near-uniform confidence. Swapped halves and two-fold cross-fitting give the same direction in every cell. A temperature fitted on the other two benchmarks did not transfer reliably and raised Brier by 0.190 $[0.103, 0.268]$ on AVeriTeC fixed retrieval. Part of the log-loss gain comes from exact zero probabilities floored at 10^{-6}, while Brier and ECE do not depend on that floor.