EVIDENTIAL ENTITLEMENT IN RAG: BEYOND CORRECTNESS AND CITATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-augmented generation (RAG) is meant to make answers more trustworthy by grounding them in external evidence. Yet the final answer can be correct and its citation valid even when the cited evidence was never given to the model. The visible output therefore cannot distinguish a properly grounded run from one with stale evidence, an incomplete search, or retrieval failure. We call this missing guarantee *evidential entitlement* and introduce a verification framework that preserves only the execution information needed to establish it. The framework treats positive and no-match claims differently: a positive claim needs a supporting witness, while a no-match claim also needs evidence that the requested search was completed. To identify the smallest sufficient record, we construct pairs of runs that differ in entitlement and select the fields needed to keep them distinguishable. We prove that once an entitlement-relevant distinction is omitted from the visible output, no downstream evaluator can recover it. We evaluate the framework on official Python, Kubernetes, and Federal Register corpora across four model families. Among 11,520 pairs with different entitlement, 42.26% still produce the same correct answer, showing that answer-level evaluation misses many entitlement failures. Source identity and version resolve many ambiguities, while the remaining cases reveal the need for query binding and search completion. On held-out data, the resulting three-field record preserves these distinctions without exposing the full execution trace. Our code is available at this anonymous repository.

1 INTRODUCTION

Retrieval-augmented generation (RAG) aims to make language-model answers more trustworthy by using external sources (Lewis et al., 2020; Karpukhin et al., 2020). At inference time, it retrieves relevant sources and provides them to the model as context. These sources give the model current or domain-specific information. The system can also show the same sources to users with the answer, so users can check whether the answer is supported.

Yet an output may look trustworthy even when the necessary retrieval was not completed. For example, a supporting source may be added only after the model has answered. The source supports the answer, but the model never saw it. Likewise, an `UNKNOWN` response may come from a timeout rather than a complete search. The final output looks normal, so the user cannot see that the required evidence was missing.

Both examples concern the same property: whether the current run satisfied the evidential conditions required by its answer. We call this property *evidential entitlement*. For a positive answer, the system must retrieve a supporting source and give it to the model for the current request. For a no-match answer, the system must instead complete the requested search over its declared scope and find no match.

Entitlement specifies what must happen during the run; interface adequacy asks whether a verifier can determine entitlement from the information the system exposes. Let $T(w)$ indicate whether an execution w has entitlement, and let $X(w)$ denote the information exposed by that execution. The interface is adequate when two executions that look identical to the verifier also have the same entitlement:

$$X(w) = X(w') \implies T(w) = T(w'). \tag{1}$$

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.

If this condition fails, two executions look identical to the verifier but differ in entitlement; we call the pair a *trust-relevant collision*.

We develop a two-stage method for repairing such interfaces. Offline, residual collisions define a minimum-cost field-selection problem whose solution is a compact interaction record H. Online, an independent gateway binds H to the request and delivered evidence, producing an execution-bound certificate. The collision rate also yields a verifier-independent error bound. We evaluate this progression—from answer and trusted gateway citation to selected record and certificate—in controlled synthetic worlds and in RAG executions over versioned public corpora.

Relation to prior work. Existing RAG evaluations measure retrieval quality, answer correctness, faithfulness to retrieved content, and support from displayed sources (Rashkin et al., 2023; Gao et al., 2023b; Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024; Joren et al., 2025). These evaluations ask whether the retrieved sources are useful, whether the generated answer is correct, and whether the displayed source supports the answer. Other methods retrieve a supporting source after the answer has been produced (Gao et al., 2023a). Related attribution studies likewise show that a reference-correct answer need not be supported by the retrieved passages (Muller et al., 2023). Our question is different: does the interface show that the source was retrieved for the current request and delivered to the generator? A source found later may support the answer, but it cannot establish what happened in the original execution. A broader comparison with provenance, authenticated query completeness, source trust, and interface theory appears in Appendix A.

Contributions. We make three contributions:

- **A new verification target.** We distinguish *evidence availability* from *evidential entitlement*: a source may support an answer without having been retrieved and delivered to the generator for the current query and claim. We further show that positive claims can be supported by a witness, whereas no-match claims require search coverage.
- **A method for synthesizing and certifying verifier interfaces.** We show that any downstream verifier must misclassify at least one run in every opposite-entitlement collision. We cast minimum-cost repair as a weighted set-cover problem over residual collision edges, and provide an execution-bound protocol that authenticates the selected record against the request and delivered evidence.
- **A real-system evaluation and executable prototype.** Across three official versioned corpora and four model families, 42.26% of 11,520 entitlement-flipping pairs still produce the same correct answer. Gateway citations reduce long-form collision from 57.43% to 28.26% but leave query binding and search completion unresolved. Leave-one-domain-out synthesis finds three tied minimum-cardinality sets; each uses three fields and leaves zero observed collision in every held-out domain. Under the stated trust model, signed receipts and authenticated coverage witnesses reject all tested integrity attacks.

2 Evidential Entitlement as an Interface Problem

We first define the judgment that the interface should support. We then separate it from neighboring claims that require different evidence.

2.1 Worlds, Interfaces, and Judgments

A world $w \in \mathcal{W}$ describes one retrieval–generation execution. It contains a query q, a target corpus D, a corpus version v, the scope actually searched, the retrieval status, the returned items, the payload delivered to the generator, the generated answer, and any execution record. The verifier observes only an interface $X(w)$.

A trust predicate $T : \mathcal{W} \to \{0, 1\}$ states the judgment that the verifier wants to make. The interface induces

$$w \sim_X w' \quad \Longleftrightarrow \quad X(w) = X(w'), \tag{2}$$

and the predicate induces $w \sim_T w'$ when $T(w) = T(w')$. The interface is adequate exactly when

$$\sim_X \subseteq \sim_T . \tag{3}$$

Table 1: Four questions about a generated claim c. Truth concerns the outcome; availability and entitlement form an evidential progression, while causal use asks whether the evidence actually affected generation.

Claim	Question	What is needed?
Truth	Is c correct?	Reference truth.
Availability	Can some source support c?	A supporting source.
Entitlement	Did this generation actually have that support?	Query-bound evidence delivered for c.
Causal use	Did the evidence affect c?	An intervention on the evidence.

Equivalently, a decision rule τ exists with $T = \tau \circ X$. We use three interfaces. The answer interface X_{ans} exposes only the normalized answer. The local interface $L = (X_R, X_G)$ exposes retriever and generator outputs. The refined interface $X^+ = (L, H)$ also exposes an interaction record H.

2.2 Four claims that should not be collapsed

Table 1 separates truth, availability, entitlement, and causal use at the answer interface.

Evidence availability asks whether a supporting source exists. Evidential entitlement asks whether supporting evidence was delivered and bound to the current query and claim. Causal use asks whether an admissible evidence intervention changes the generated claim. Availability can hold without entitlement. Entitlement can hold without a large causal effect. Each claim therefore induces a different partition of worlds.

This paper focuses on entitlement. The practical question is: did this execution possess an acceptable evidential basis for this claim? The question appears in citation systems, grounded summarization, medical and legal search, enterprise assistants, and tool-using agents.

2.3 Controlled distinguishability

An adequate interface should not reveal every difference between worlds. Let R_{hide} contain world pairs whose differences should remain hidden. A controlled interface should satisfy

$$R_{\text{hide}} \subseteq \sim_X \subseteq \sim_T . \tag{4}$$

The right inclusion preserves the distinctions needed for verification. The left inclusion suppresses distinctions that the verifier does not need. The design target is therefore a small, claim-bound interface rather than an internal dump.

3 Why local views lose entitlement

The retriever exposes evidence items. The generator exposes claims. Entitlement depends on how the execution relates the two. Negative claims add a second requirement: coverage of the declared search space.

3.1 Evidence is a coupling, not a marginal

Fix one query. Let $\mathcal{E} = \{e_1, \ldots, e_m\}$ denote evidence items and $\mathcal{C} = \{c_1, \ldots, c_n\}$ denote generated claims. Let μ and ν be their local masses. A coupling $\Gamma \in \mathbf{\Pi}(\mu, \nu)$ records the execution-level assignment from evidence to claims. Let S_{ij} measure how well e_i supports c_j. We define

$$J_S(\Gamma) = \sum_{i=1}^{m} \sum_{j=1}^{n} S_{ij} \Gamma_{ij}. \tag{5}$$

The local views determine μ and ν. The interaction determines Γ.

Theorem 1 (When local views determine entitlement)**.** *Assume that μ and ν have full support. The following statements are equivalent.*

1. *$J_S(\Gamma)$ is constant over $\mathbf{\Pi}(\mu, \nu)$.*
2. *Every 2×2 cross-difference vanishes: $S_{ij} + S_{i'j'} - S_{ij'} - S_{i'j} = 0$.*

3. Vectors u and v exist such that $S_{ij} = u_i + v_j$ for every pair.

The theorem gives a sharp boundary. Local views determine entitlement only when support separates into an evidence-only score and a claim-only score. Real support is usually relational. A medical source can support one diagnosis and contradict another. A legal clause can support one interpretation and be irrelevant to another.

The smallest example uses

$$S = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}. \tag{6}$$

One coupling binds e_1 to c_1 and e_2 to c_2. Another swaps the bindings. The evidence and claim multisets stay fixed, but entitlement changes.

Corollary 2 (Local-only collision)**.** *If S has a nonzero cross-difference, two worlds exist with identical evidence and claim marginals but different entitlement scores.*

This result also separates generation-time entitlement from post-hoc support. A post-hoc system can retrieve a source after seeing the claim. It constructs a new support relation and can establish availability. It cannot identify the coupling that governed the original execution.

3.2 Positive evidence needs a witness; negative evidence needs coverage

Let $D = \{d_1, \ldots, d_N\}$ be the target corpus and let $m_q(d)$ indicate an exact match for query q. Define

$$T_{\exists}(D, q) = \mathbf{1}[\exists d \in D : m_q(d) = 1], \tag{7}$$

$$T_{\varnothing}(D, q) = \mathbf{1}[\forall d \in D : m_q(d) = 0]. \tag{8}$$

An authenticated item $d^\star$ with $m_q(d^\star) = 1$ certifies the positive claim. A no-match claim has no analogous local witness.

Theorem 3 (Witness–coverage asymmetry)**.** *Suppose a retrieval trace inspects a strict subset $S \subsetneq D$ and finds no match in S. If the verifier receives no authenticated constraint on $D \setminus S$, then the same trace is consistent with one world where $T_{\varnothing} = 1$ and another where $T_{\varnothing} = 0$.*

A negative certificate must therefore account for the target scope, searched scope, corpus version, and successful completion. The representation can be compact, for example through a trusted index or authenticated nonmembership proof. The represented property remains global. We call this principle *no free UNKNOWN*: a model can emit `UNKNOWN` cheaply, but a system must pay for the coverage that makes the claim warranted. Authenticated-database work likewise treats result correctness and completeness as distinct integrity obligations (Xie et al., 2007).

The same distinction applies to operational states. Empty top-k, partial search, timeout, retrieval error, no signal, and complete no-match should not collapse when the trust predicate distinguishes them. A nearest candidate is also not negative evidence: similarity does not certify either exact match or exact absence.

4 Collision-guided entitlement verification

The previous section identifies the execution relations that entitlement depends on. We now give a constructive method with two phases: residual collisions synthesize a compact verifier-facing record, and an independent gateway binds that record to the current request and delivered evidence to produce an entitlement certificate.

Figure 1 summarizes the synthesis phase: a trust-relevant collision becomes a residual graph edge, and candidate interaction fields are selected to cut every such edge at minimum cost.

4.1 Measuring interface failure

Let γ be a distribution over paired worlds (W^+, W^-) with opposite judgments: $T(W^+) = 1$ and $T(W^-) = 0$. We define

$$\mathrm{TrC}_\gamma(X) = \Pr_{(W^+, W^-) \sim \gamma}[X(W^+) = X(W^-)]. \tag{9}$$

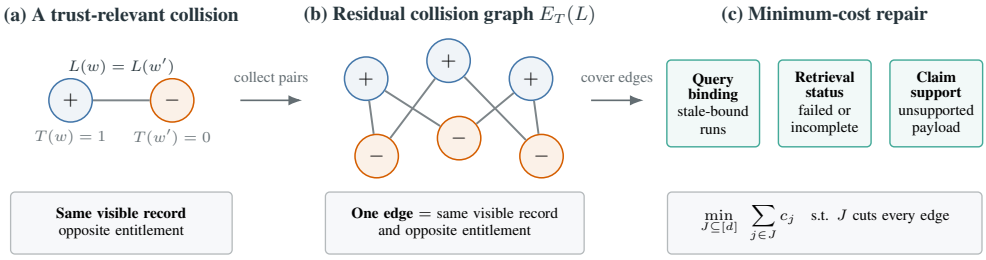


Figure 1: Collision-guided interface repair. Opposite-entitlement runs that remain indistinguishable under the current interface form residual collision edges. Candidate interaction fields cut subsets of these edges, and synthesis selects a lowest-cost field set that separates all residual collisions.

The paired design holds task content fixed and changes the evidential relation.

Theorem 4 (Collision bound)**.** *For every verifier V that observes only X,*

$$\mathrm{Err}_\gamma(V \circ X) \geq \frac{1}{2}\,\mathrm{TrC}_\gamma(X), \qquad \mathrm{Acc}_\gamma(V \circ X) \leq 1 - \frac{1}{2}\,\mathrm{TrC}_\gamma(X). \tag{10}$$

The bound turns an empirical collision rate into a certification ceiling. A collision pair gives the verifier the same message under opposite labels.

Corollary 5 (Post-processing monotonicity)**.** *For every deterministic map g,*

$$\mathrm{TrC}_\gamma(g(X)) \geq \mathrm{TrC}_\gamma(X).$$

A stronger answer-side judge can improve answer evaluation. It cannot infer an execution relation that leaves no trace in its input. The stochastic extension uses total variation and appears in the appendix.

4.2 Collision-guided interface synthesis

Fix a finite design set $\mathcal{W}_0$. Let $L(w) = (X_R(w), X_G(w))$ be the current local interface. Define the residual trust-relevant edges

$$\mathcal{E}_T(L) = \{\{w, w'\} : L(w) = L(w'),\ T(w) \neq T(w')\}. \tag{11}$$

Let $h_1, \ldots, h_M$ be candidate interaction fields. Field h_j cuts an edge when its value differs across the two worlds.

Theorem 6 (Residual collision cover)**.** *For $J \subseteq \{1, \ldots, M\}$, let $H_J(w) = (h_j(w))_{j \in J}$. The refined interface (L, H_J) is adequate for T on $\mathcal{W}_0$ if and only if every edge in $\mathcal{E}_T(L)$ is cut by at least one selected field.*

The minimum-cost repair solves

$$J^\star \in \arg\min_J \sum_{j \in J} c_j \quad \text{subject to every edge in } \mathcal{E}_T(L) \text{ being cut.} \tag{12}$$

Algorithm 1: Collision-guided interface synthesis
Input. Design worlds $\mathcal{W}_0$, judgment T, current interface L, and candidate fields with costs $\{(h_j, c_j)\}_{j=1}^M$.
Build. Construct $\mathcal{E}_T(L)$ and, for each h_j, the residual edges it cuts.
Cover. Solve equation 12; if no cover exists, the candidate field library is insufficient on $\mathcal{W}_0$.
Return. $H_{J^\star}$ and $X^+(w) = (L(w), H_{J^\star}(w))$.

The optimization is weighted set cover and is NP-hard in general.[1] Costs can represent bytes, latency, privacy exposure, or engineering burden. Positive entitlement requires query-bound claim support; no-match entitlement additionally requires status or coverage over the declared corpus scope. The guarantee is exact on $\mathcal{W}_0$; broader threat models may introduce new edges and require additional fields.

[1] In our eight-field study, unit costs reduce the objective to cardinality; we enumerate all 2^8 subsets and freeze every tied minimum before held-out evaluation.

4.3 BINDING THE RECORD INTO A CERTIFICATE

Algorithm 1 determines what the verifier must observe, but an informative record is not yet a certificate: an untrusted application could copy a valid-looking record into an invalid execution.

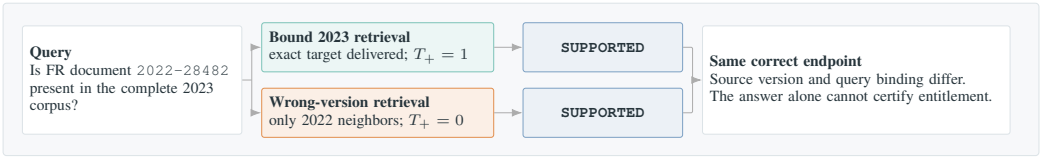
Protocol 1: Execution-bound entitlement certificate
Construct. An independent gateway records $H_{J^\star}$ and signs a receipt binding it to the request commitment, delivered-evidence digest, nonce, and validity interval. A positive certificate carries query-bound delivered support; a no-match certificate additionally carries terminal status and authenticated coverage for the declared corpus version and scope.
Verify. The verifier checks the signature, request and evidence binding, freshness and replay conditions, and any required coverage witness. It then applies the fixed entitlement rule τ to $(L, H_{J^\star})$.

Thus Algorithm 1 selects the semantic content of the certificate, while Protocol 1 binds that content to the execution that produced it. Our implementation uses independently signed receipts and authenticated nonmembership proofs; related systems bind behavior or artifacts to secure action logs, signed process layouts, and transparency-backed signing (Haeberlen et al., 2007; Torres-Arias et al., 2019; Newman et al., 2022). The guarantee assumes the stated trust boundary, including an honest gateway, uncompromised signing keys, and an honestly ingested authenticated snapshot; it does not establish future absence or semantic relevance beyond the declared support predicate.

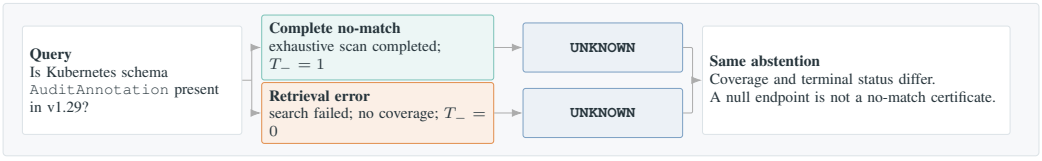
5 EMPIRICAL EVALUATION: FROM ANSWER TO CERTIFICATE

The empirical evaluation is organized around three claims derived from the theory. Answer equality can hide changes in entitlement; trusted citations can recover source identity while leaving execution details unresolved; and residual collisions can reveal the fields required in H. We test these claims in order and then examine whether the resulting interface remains robust and supports executable verification.

We use two complementary settings. The synthetic-world study isolates the evidence relation under known-by-construction labels (Appendix H). The primary study uses versioned Python, Kubernetes, and Federal Register corpora to determine whether the same failures arise with natural documents, real version changes, and realistic retrieval stacks. Figure 2 illustrates the paired comparisons used throughout the public-corpus study.



(a) A correct endpoint survives wrong-version retrieval.



(b) The same UNKNOWN follows complete coverage and retrieval failure.

Figure 2: Two observed trust-relevant collisions from the public-corpus study. The query and visible endpoint remain fixed while version binding or search coverage changes.

5.1 PAIRED EXECUTIONS OVER THREE VERSIONED CORPORA

A collision is meaningful only if entitlement, rather than task content, changes. We therefore hold the query, task truth, model, and corpus family fixed while varying the evidence relation, and test whether the answer, citation, or interaction-record interface distinguishes each pair.

We use official Python 3.10/3.11 Sphinx inventories, Kubernetes v1.29.0/v1.30.0 OpenAPI specifications, and complete Federal Register 2022/2023 snapshots. These domains cover software

documentation, an API schema, and regulatory text; each contributes 64 version-delta entities and 192 queries.

A deterministic BM25 retriever filters by version and logical scope and returns two passages for positive claims. Complete negatives use an exhaustive exact-key scan over the declared scope. Frozen traces isolate the intervention; E5, reciprocal-rank fusion, and ColBERTv2 reruns test dependence on lexical retrieval.

We run Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Granite-3.3-8B-Instruct, and OLMo-2-1124-7B-Instruct with greedy decoding on the frozen traces. Positive reference worlds are paired with stale-cache, wrong-version, partial-scope, timeout, and nearest-only alternatives; complete-no-match worlds with empty top-k, partial scope, retrieval error, no signal, and nearest-only alternatives. The primary graph contains 13,824 generations and 11,520 entitlement-flipping pairs.

We expose progressively richer interfaces. X_{ans} contains the normalized endpoint; a trusted gateway adds every delivered source identifier and version; and H adds query binding, retrieval status, coverage, claim support, and integrity. The generator never receives the held-out entitlement label or verifier receipt. Table 2 summarizes this progression.

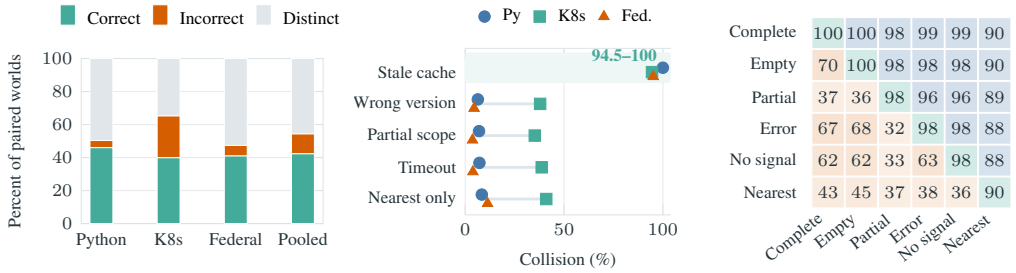
Table 2: Claim-aligned primary results. Positive and negative columns expose the witness–coverage asymmetry; the ceiling converts endpoint collision into an answer-only verification limit.

	Short-answer paired graph (%)				Long-form collision (%)		Record
Domain	Positive	Negative	All	A-only ceiling ↑	Endpoint	+ citation	Full-H residual ↓
Python	25.82	99.30	50.31	74.84	55.00	33.12	0 / 3,840
Kubernetes	49.45	96.64	65.18	67.41	58.33	25.00	0 / 3,840
Federal Register	23.71	94.61	47.34	76.33	58.96	26.67	0 / 3,840
Pooled	**33.00**	**96.85**	**54.28**	**72.86**	**57.43**	**28.26**	**0 / 11,520**

The short-answer graph contains 2,560 positive and 1,280 negative pairs per domain. The separately frozen long-form extension contains 480 pairs per domain. Among all 11,520 short-answer pairs, 42.26% return the same correct answer; 77.85% of answer collisions are paired-correct.

5.2 The endpoint collapses trust-relevant worlds

If the endpoint preserved entitlement, opposite-label pairs would produce distinguishable answers. We test this prediction first; Fig. 3 decomposes the lost distinctions across paired correctness, positive interventions, and negative retrieval states.



(a) Correct collision composition. **(b)** Positive pairs by intervention. **(c)** Negative states merge at the endpoint.

Figure 3: Where the answer interface loses entitlement. (a) The teal mass is the headline result: 42.26% of all real-corpus pairs return the same correct answer. (b) Stale-cache binding is nearly invisible across all three domains, while other positive interventions vary by corpus. (c) In the pooled negative-state matrix, the upper triangle shows normalized-answer collision, the lower triangle shows visible-text collision, and the diagonal shows $P(\texttt{UNKNOWN})$.

Instead, 54.28% of pairs collide, limiting answer-only balanced accuracy to 72.86%. More sharply, 42.26% of all pairs return the same *correct* answer, and 77.85% of the 6,253 collisions are paired-correct. Task success therefore does not identify the evidential relation.

The two sides have different geometry. Positive collision is 33.00%, but stale-cache pairs collide in 94.53–100% across domains because source identity and answer can persist while query binding fails.

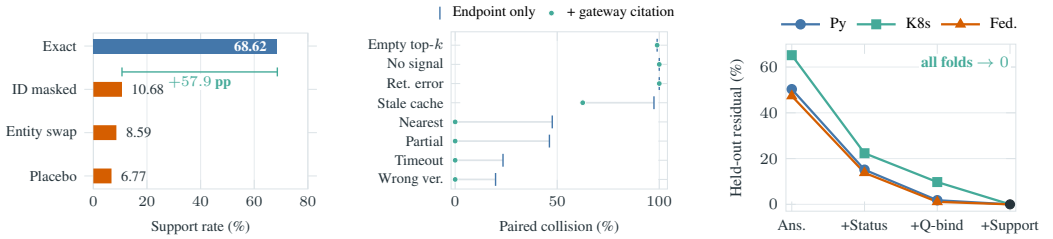
Negative collision is 96.85%: complete no-match, empty output, partial search, retrieval error, no signal, and nearest-only retrieval often share an endpoint.

The abstention token adds almost no coverage evidence: complete coverage rises from a 16.67% prior to only 17.09% given `UNKNOWN`, yielding 51.50% balanced accuracy. Under known-by-construction labels, the synthetic study still finds 93.03% positive-binding and 82.94% negative-state collision (Appendix H), ruling out public-fact memorization as the sole explanation.

5.3 Citation identifies the object; interaction fields recover the relation

We next ask which observations repair the loss. Before treating evidence delivery as a verifier field, a four-arm intervention holds query and decoding fixed while assigning exact evidence, identifier-masked evidence, a length-matched placebo, or valid evidence from another entity. Exact evidence produces `SUPPORTED` in 68.62% of outputs versus 6.77–10.68% for controls, confirming that payload assignment affects the system's behavior.

We generate 1,728 long-form responses and 1,440 paired comparisons. A trusted gateway appends every delivered `SOURCE_ID@VERSION`; the model does not choose identifiers. This favorable interface reduces collision from 57.43% to 28.26%. Figure 4 traces the progression from payload to citation to H.



(a) Delivered payload changes behavior.

(b) Citation resolves payload identity, not status or binding.

(c) Residual collisions synthesize a compact H.

Figure 4: From provenance object to execution relation. (a) Exact evidence changes the output distribution. (b) Gateway citations eliminate collisions when the delivered object changes, but not when provenance is empty or stale. (c) Retrieval status, query binding, and claim support or exact-match status remove held-out residual edges. Of the 407 remaining answer-plus-citation collisions, 93.61% are paired-correct; 70.52% hide retrieval status and 29.48% hide query binding.

Of 407 answer-plus-citation collisions, 381 are paired-correct. Empty top-k, no signal, and retrieval error contribute 70.52% and expose the same null provenance; stale cache contributes 29.48% and preserves source and version under different query binding. Citation names an object, whereas entitlement depends on its execution relation.

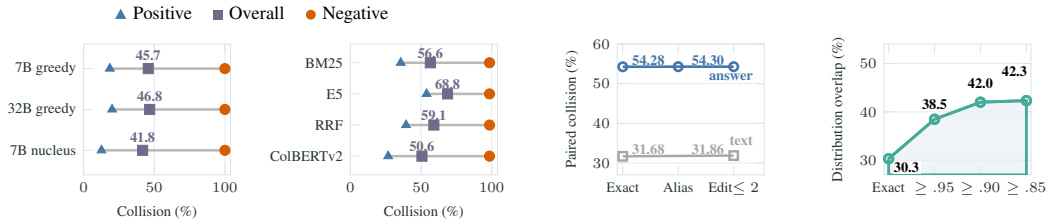
We let the residual graph choose that relation. For each held-out domain, we enumerate all $2^8 = 256$ subsets of eight categorical fields on the other two domains, retain every minimum-cardinality zero-collision set, and evaluate without reselection. All folds recover query binding and claim support plus retrieval status, coverage, or exact-match status. Every selected set and full H has zero observed held-out collision.

5.4 Cross-axis replication and executable verification

We finally test whether the failure survives experimental choices and whether the selected H supports an executable verifier. Figure 5 covers model, decoding, retrieval, and output-equivalence choices.

The failure survives all four axes. Matched 7B greedy, 32B greedy, and 7B nucleus runs retain 41.77–46.77% overall collision; BM25, E5, reciprocal-rank fusion, and ColBERTv2 retain 50.63–68.75%. Negative collision remains 98.44–100%. Conservative alias and edit equivalences change the 54.28% headline by at most 0.02 percentage points; free-form semantic overlap reaches 42.00% at cosine threshold 0.90.

Table 3 reports the resulting decision audit and integrity tests.



(a) Backend and decoding. **(b)** Retriever stack. **(c)** Tolerant equality. **(d)** Semantic overlap.

Figure 5: Robustness across four nuisance axes. Model scale, decoding, retrieval stack, and output equivalence change the measured magnitude without eliminating the interface failure. Exact revisions, seeds, payload-overlap statistics, and confidence intervals appear in the appendix.

Table 3: Interface audit. Panel A evaluates fixed entitlement decisions on 1,728 long-form worlds. Panel B tests integrity and authenticated negative coverage at the operational trust boundary.

Panel A. Entitlement decisions ($N = 1{,}728$: 288 entitled / 1,440 non-entitled)

Verifier input	Balanced acc. (95% CI) ↑	Sens. / spec.	FP / FN	Brier ↓	Residual collision ↓
Answer only	70.45 [69.20, 71.67]	85.42 / 55.49	641 / 42	0.395	57.43
Validated gateway citation	77.29 [76.01, 78.58]	85.42 / 69.17	444 / 42	0.281	28.26
Full interaction record H	**100.00 [100, 100]**	**100.00 / 100.00**	**0 / 0**	**0.000**	**0.00**

Panel B. Record integrity and cost

Mechanism	Test scale	Failure families	Detected / separated	Evidence size	Local latency
Signed receipt	47,808 attacks	field, digest, replay, signature, expiry	**47,808 / 47,808**	∼1.1 KiB	0.23–0.36 ms p95
Coverage witness	512 shard proofs	omission, tamper, replay, insertion	**100%**	31.2 KiB	1.68 / 0.28 ms†

FP/FN are counts. †Coverage latency is median build / verify. These finite tests do not certify a dishonest gateway, a compromised signing key, future corpus states, or semantic non-relevance. C2PA/log transport checks remain in the appendix.

Fixed rules reach 70.45% balanced accuracy from answers and 77.29% with validated gateway citations, which still leave 444 false positives. Full H reaches 100% sensitivity and specificity on the frozen graph, making $T = \tau \circ X$ executable through fixed verifier rules.

Semantic adequacy also requires integrity. An independent gateway signs the record, request commitment, evidence digest, nonce, and validity interval. It accepts 3,456 genuine envelopes and rejects all 47,808 tested attacks; an authenticated-coverage prototype detects every tested omission, tampering, replay, and insertion. Under the stated trust boundary, integrity binds H to the request and delivered evidence.

6 Discussion and Conclusion

Answer correctness asks whether an answer is right; evidential entitlement asks whether the current execution provides enough support for it. Trust-relevant collisions show that runs with different query binding or search coverage can return the same correct answer.

Across the controlled synthetic-world and public-corpus studies, this failure persists across four nuisance axes. Trusted gateway citations separate runs when the delivered evidence changes, while residual collisions identify three-field interaction records with zero observed collision in every held-out domain.

The zero-collision result is established for the tested pairs under an honest gateway and an uncompromised signing key. Within this scope, trusted gateway citations identify the delivered evidence, H preserves the required execution relations, and integrity turns that record into a certificate of evidential entitlement.

THE USE OF LARGE LANGUAGE MODELS

Generative AI tools assisted with language and structural editing, literature discovery and reference formatting, editing analysis and figure-generation code, and improving the presentation of figures and tables. They also provided feedback on the presentation of the experimental design and its connection to the theoretical claims. The authors reviewed and revised all AI-assisted text and code, verified cited works against the original sources, and checked reported values and visualizations against the underlying experimental outputs. The authors take responsibility for the final text, code, figures, claims, and conclusions.

REFERENCES

Aris Anagnostopoulos, Michael T. Goodrich, and Roberto Tamassia. Persistent authenticated dictionaries and their applications. In *Information Security*, pp. 379–393, 2001.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953.

Jan Buchmann, Xiao Liu, and Iryna Gurevych. Attribute or abstain: Large language models as long document assistants. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8113–8140, 2024.

Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Why and where: A characterization of data provenance. In *International Conference on Database Theory*, pp. 316–330, 2001.

Adriane Chapman and H. V. Jagadish. Why not? In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of Data*, pp. 523–534, 2009.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3569–3587, 2024.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16 (1):41–46, 1970.

Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. RAGAs: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, et al. Rarr: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, 2023a.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023b.

Shafi Goldwasser, Silvio Micali, and Charles Rackoff. The knowledge complexity of interactive proof systems. *SIAM Journal on Computing*, 18(1):186–208, 1989.

Todd J. Green, Grigoris Karvounarakis, and Val Tannen. Provenance semirings. In *Proceedings of the 26th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 31–40, 2007.

Andreas Haeberlen, Petr Kouznetsov, and Peter Druschel. PeerReview: Practical accountability for distributed systems. In *Proceedings of the 21st ACM Symposium on Operating Systems Principles*, pp. 175–188, 2007.

Ragib Hasan, Radu Sion, and Marianne Winslett. The case of the fake picasso: Preventing history forgery with secure provenance. In *7th USENIX Conference on File and Storage Technologies*, 2009.

Sophie Henning, Talita Anthonio, Wei Zhou, Heike Adel, Mohsen Mesgar, and Annemarie Friedrich. Is the answer in the text? challenging ChatGPT with evidence retrieval from instructive text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 14229–14241, 2023.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020.

SeungHoo Lee, JuneHyoung Kwon, Jooweon Choi, JungMin Yun, Seunguk Yu, Yoonji Lee, Jinhee Jang, and YoungBin Kim. RefLens: End-to-end evidence-grounded citation verification with LLM agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 41625–41627, 2026. doi: 10.1609/aaai.v40i48.42361.

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. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, 2020.

Yifei Li, Xiang Yue, Zeyi Liao, and Huan Sun. AttributionBench: How hard is automatic attribution evaluation? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 14919–14935, 2024.

Nelson F. Liu, Tianyi Zhang, and Percy Liang. Evaluating verifiability in generative search engines. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7001–7025, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

Benjamin Muller, John Wieting, Jonathan H Clark, Tom Kwiatkowski, Sebastian Ruder, Livio Baldini Soares, Roee Aharoni, Jonathan Herzig, and Xinyi Wang. Evaluating and modeling attribution for cross-lingual question answering. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 144–157, 2023.

Zachary Newman, John Speed Meyers, and Santiago Torres-Arias. Sigstore: Software signing for everybody. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2353–2367, 2022.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pp. 10862–10878, 2024.

David L. Parnas. On the criteria to be used in decomposing systems into modules. *Communications of the ACM*, 15(12):1053–1058, 1972.

Anirudh Phukan, Shwetha Somasundaram, Apoorv Saxena, Koustava Goswami, and Balaji Vasan Srinivasan. Peering into the mind of language models: An approach for attribution in contextual question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11481–11495, 2024.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6037–6053, 2024.

Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840, 2023.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999–22027, 2024.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354, 2024.

Tobias Schreieder, Tim Schopf, and Michael Färber. Attribution, citation, and quotation: A survey of evidence-based text generation with large language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 30956–31000, 2026.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

Qi Shi, Ye Yang, Jiaqi Di, Lin Li, Jie Liu, Youmin Hu, Jinhua Xiao, and Zhongxu Hu. Trust-aware KG-RAG: A framework for trustworthy question answering for engineering equipment. *Advanced Engineering Informatics*, 74:104801, 2026. doi: 10.1016/j.aei.2026.104801.

Shiqi Sun, Kun Zhang, Jingyuan Li, Min Yu, Kun Hou, Yuanzhuo Wang, and Xueqi Cheng. Retriever-generator-verification: A novel approach to enhancing factual coherence in open-domain question answering. *Information Processing & Management*, 62(4):104147, 2025. doi: 10.1016/j.ipm.2025.104147.

Santiago Torres-Arias, Hammad Afzali, Trishank Karthik Kuppusamy, Reza Curtmola, and Justin Cappos. in-toto: Providing farm-to-table guarantees for bits and bytes. In *28th USENIX Security Symposium*, pp. 1393–1410, 2019.

Abraham Wald. *Statistical Decision Functions*. Wiley, 1950.

Min Xie, Haixun Wang, Jian Yin, and Xiaofeng Meng. Integrity auditing of outsourced data. In *Proceedings of the 33rd International Conference on Very Large Data Bases*, pp. 782–793, 2007.

Hua Ye, Siyuan Chen, Ziqi Zhong, Canran Xiao, Haoliang Zhang, Yuhan Wu, and Fei Shen. Seeing through the conflict: Transparent knowledge conflict handling in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34423–34431, 2026. doi: 10.1609/aaai.v40i40.40740.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, 2024.

Xiang Yue, Boshi Wang, Ziru Chen, Kai Zhang, Yu Su, and Huan Sun. Automatic evaluation of attribution by large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 4615–4635, 2023.

APPENDIX CONTENTS AND ROADMAP

The appendix follows the same progression as the main paper. We first position the work, then supply the formal definitions and proofs, and finally document the two empirical layers. The real-corpus study is the primary deployment-shaped evaluation. The private-world study is a mechanism-isolation complement that removes public-fact memorization.

	Appendix	Page
A	Extended related work. RAG verification, citations, provenance, knowledge conflict, and information interfaces.	14
B	Formal setup and factorization. Complete definitions of adequacy, entitlement, and claim-specific interfaces.	15
C	Evidence-coupling proofs. Proof of the coupling theorem and its collision corollaries.	16
D	Witness–coverage proofs. Nonexistence, inspection cost, and nearest-candidate limits.	17
E	Collision bounds. Deterministic, post-processing, and distributional verifier limits.	17
F	Interface synthesis, integrity, and extensions. Set-cover proofs, hiding constraints, self-certification, and adaptive probing.	18
G	Real-corpus protocol and supplementary analyses. Sources, retrievers, receipt schema, robustness, citation analysis, held-out interface selection, and integrity tests.	19
H	Private-world mechanism-isolation study. Controlled positive binding, negative coverage, abstention, nearest candidates, and interface ablations.	27

Claim-to-evidence map.

Main claim	Formal support	Empirical support
Entitlement is relational	Coupling characterization in Appendix C	Bound, stale, and wrong-version pairs in Appendices G and H
Positive and negative evidence differ	Witness–coverage proof in Appendix D	Citation, status, abstention, and coverage analyses in Appendices G and H
Collisions limit any verifier	Collision bounds in Appendix E	Answer-only ceilings and fixed verifier rules in Appendix G
Residual collisions design H	Set-cover result in Appendix F	Held-out field selection and interface ablations in Appendices G and H

Reproducibility map. Appendix G documents the primary paper artifact. Appendix H documents the independent private-world stress test. Exact model revisions, pairing keys, record schemas, field derivations, and attack assumptions are stated where they are first used.

A Extended related work

This appendix expands the compact positioning in the main text. The closest work improves answer quality, verifies displayed support, or authenticates provenance objects. Our question is complementary: which verifier-facing interface preserves the execution relation required by a declared evidential judgment?

A.1 Retrieval-augmented generation and local evidence verification

Retrieval-augmented generation brings external knowledge into generation (Lewis et al., 2020; Karpukhin et al., 2020). Later systems decide when to retrieve, critique their own responses, and reduce sensitivity to irrelevant context (Asai et al., 2024; Yoran et al., 2024). Evaluation frameworks diagnose retrieval quality, faithfulness, and generation quality (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024). Benchmarks also expose failures at the local-evidence boundary: RAGTruth annotates unsupported or contradictory spans in RAG outputs, while WHERE includes topically relevant questions whose supplied text contains no answer evidence (Niu et al., 2024; Henning et al., 2023). Recent published work probes this tuple more finely. Sufficient Context separates failures caused by insufficient retrieval from failures to use available context, while RGV validates retrieved subgraphs against external knowledge (Joren et al., 2025; Sun et al., 2025). Real-world claim-verification pipelines also make retrieval time and incomplete evidence-set construction explicit rather than assuming human-curated evidence (Chen et al., 2024). These systems improve or diagnose variants of the local tuple (q, E, a). We study a prior interface question: whether that exposed tuple distinguishes executions with different query binding, search coverage, or completion status.

A.2 Citations, factuality, and attribution

Attributed generation evaluates links between claims and displayed sources (Rashkin et al., 2023; Gao et al., 2023b; Liu et al., 2023). Cross-lingual question answering further shows that even exact-reference answers can lack support from retrieved passages (Muller et al., 2023). Automatic evaluators test whether cited references fully support generated statements, while long-document assistants study joint evidence extraction and abstention (Yue et al., 2023; Buchmann et al., 2024). Cross-dataset benchmarking further shows that strong automatic attribution evaluators remain imperfect (Li et al., 2024). Post-hoc systems retrieve support and revise unsupported text (Gao et al., 2023a). Atomic factuality and model-internal attribution evaluate claim support or context influence (Min et al., 2023; Qi et al., 2024; Phukan et al., 2024). RefLens provides an end-to-end post-hoc citation audit for scholarly documents (Lee et al., 2026). Recent survey work organizes this literature around attribution, citation, and quotation (Schreieder et al., 2026). These methods establish displayed support or construct a new support relation after generation. Generation-time entitlement asks whether the original retrieval–generation handoff carried the relevant relation. The two claims can differ even when the displayed source and final answer are unchanged.

A.3 Knowledge conflict and source trust

TCR makes the parametric–retrieved conflict decision observable and controllable (Ye et al., 2026). Trust-Aware KG-RAG assigns reliability weights and resolves cross-source conflicts during graph construction and retrieval (Shi et al., 2026). These methods choose, transform, or prioritize evidence. Such a policy can change the support matrix S or the target predicate T. Our results ask whether the interface preserves the resulting judgment after that policy and predicate have been declared.

A.4 Provenance and authenticated negative evidence

Database provenance characterizes how inputs contribute to outputs (Buneman et al., 2001; Green et al., 2007). Why-not provenance instead explains which operations excluded an expected result (Chapman & Jagadish, 2009). Authenticated-query systems distinguish result correctness from completeness (Xie et al., 2007), while authenticated dictionaries support compact membership and nonmembership verification (Anagnostopoulos et al., 2001). Secure provenance chains make history tampering detectable (Hasan et al., 2009); accountability logs and supply-chain frameworks bind actions or artifacts to recorded principals and process steps (Haeberlen et al., 2007; Torres-Arias et al.,

2019). Sigstore combines identity-bound signing with transparency-log evidence (Newman et al., 2022). Our prototype uses these mechanisms as carriers for hash-bound references to H. They do not by themselves define the query–evidence–claim relation. We bring provenance and authenticated nonmembership to the retrieval–generation handoff. The witness–coverage result explains why an absence claim requires a stronger interface than a witnessed membership claim.

A.5 Interfaces, information, and verification

Software interfaces hide implementation while preserving downstream obligations (Parnas, 1972). Statistical decision theory asks which observations support which decisions (Wald, 1950; Blackwell, 1953). Information theory quantifies representation loss (Shannon, 1948). Interactive proofs show that a verifier can learn a claim without receiving the full hidden object (Goldwasser et al., 1989). Selective classification separates prediction from abstention (Chow, 1970). These lines motivate controlled distinguishability. A useful interface preserves the distinctions needed for a declared decision while suppressing unrelated implementation details.

B Formal setup, notation, and factorization

This appendix collects the objects used across the theoretical and empirical sections. The main text introduces each object at first use. Here we state the deterministic setup in one place and make the factorization criterion explicit.

B.1 Core objects

Symbol	Meaning
$\mathcal{W}, w$	Space of complete retrieval–generation worlds and one execution world.
$X(w)$	Verifier-facing interface exposed by world w.
$T(w)$	Declared trust judgment, usually evidential entitlement.
$L = (X_R, X_G)$	Local retriever and generator views.
H	Interaction record that preserves query binding, support, coverage, status, or integrity.
$\mathcal{E}, \mathcal{C}$	Evidence items and generated claims for one query.
Γ	Execution-level evidence–claim coupling.
S	Claim-specific support matrix.
$\mathrm{TrC}_\gamma(X)$	Trust-relevant collision rate under paired-world distribution γ.
R_{hide}	World pairs whose differences should remain hidden.

Table B.1: Notation used in the main text and appendices.

Definition 7 (Interface adequacy)**.** Let $X : \mathcal{W} \to \mathcal{X}$ be an interface and let $T : \mathcal{W} \to \mathcal{T}$ be a judgment. We call X adequate for T when

$$X(w) = X(w') \Longrightarrow T(w) = T(w') \tag{B.1}$$

for all $w, w' \in \mathcal{W}$.

Proposition 8 (Factorization)**.** *The interface X is adequate for T if and only if a function $\tau : X(\mathcal{W}) \to T(\mathcal{W})$ exists with $T = \tau \circ X$.*

Proof. If X is adequate, define $\tau(x)$ as $T(w)$ for any world with $X(w) = x$. Adequacy makes this definition independent of the chosen world. Hence $T = \tau \circ X$. Conversely, if $T = \tau \circ X$, equal interface values produce equal judgments. □

B.2 Claim ladder

Definition 9 (Evidence availability, entitlement, and causal use)**.** Fix a query q and generated claim c. Evidence availability holds when some item in the target corpus supports c. Evidential entitlement holds when a supporting item was delivered and bound to (q, c) in the execution. Causal use holds when an admissible intervention on that item changes the distribution of the generated claim.

The three definitions induce different partitions of $\mathcal{W}$. An interface can be adequate for one claim and inadequate for another. Unless stated otherwise, the paper studies generation-time entitlement. The four-arm intervention in Section G.8 separately measures black-box response sensitivity to delivered evidence.

C EVIDENCE-COUPLING PROOFS

We prove the characterization used in the main text. The argument follows the cycle space of the transportation polytope: local marginals determine entitlement exactly when every coupling-preserving cycle has zero support gain.

C.1 PROOF OF THE COUPLING THEOREM

Proof. We prove the cycle condition first. Let $\mathbf{\Pi}(\mu, \nu)$ be the transportation polytope with strictly positive marginals. Its feasible perturbations include every sufficiently small 2×2 cycle

$$\Delta^{ii'jj'} = \epsilon \begin{pmatrix} +1 & -1 \\ -1 & +1 \end{pmatrix} \tag{C.1}$$

placed in rows i, i' and columns j, j'. The perturbation preserves every row sum and every column sum.

Assume that J_S is constant on the polytope. For an interior coupling Γ, both $\Gamma + \Delta^{ii'jj'}$ and $\Gamma - \Delta^{ii'jj'}$ are feasible for sufficiently small ϵ. Constancy gives

$$0 = J_S(\Gamma + \Delta^{ii'jj'}) - J_S(\Gamma) \tag{C.2}$$

$$= \epsilon \left(S_{ij} + S_{i'j'} - S_{ij'} - S_{i'j} \right). \tag{C.3}$$

Every cross-difference therefore vanishes.

Assume next that every cross-difference vanishes. Fix reference indices i_0 and j_0. Define

$$u_i = S_{ij_0}, \qquad v_j = S_{i_0 j} - S_{i_0 j_0}. \tag{C.4}$$

The vanishing cross-difference for (i, i_0, j, j_0) gives

$$S_{ij} = S_{ij_0} + S_{i_0 j} - S_{i_0 j_0} = u_i + v_j. \tag{C.5}$$

This proves the additive form.

Finally, assume $S_{ij} = u_i + v_j$. For every coupling Γ with marginals μ and ν,

$$J_S(\Gamma) = \sum_{i,j} (u_i + v_j) \Gamma_{ij} \tag{C.6}$$

$$= \sum_i u_i \mu_i + \sum_j v_j \nu_j. \tag{C.7}$$

The value depends only on the marginals. It is therefore constant over $\mathbf{\Pi}(\mu, \nu)$. □

C.2 PROOF OF THE LOCAL-ONLY COLLISION COROLLARY

Proof. A nonzero cross-difference violates condition 2 of Theorem 1. The theorem implies that J_S is not constant over $\mathbf{\Pi}(\mu, \nu)$. Hence two couplings Γ and Γ' share the same marginals and satisfy $J_S(\Gamma) \neq J_S(\Gamma')$. Construct two worlds that share the local evidence and claim views and differ only in the interaction assignment. The local interface is equal in the two worlds, while the entitlement score differs. □

C.3 A BINARY SUPPORT COROLLARY

Corollary 10 (Relational support creates a collision)**.** *Suppose there are evidence items $e_i, e_{i'}$ and claims $c_j, c_{j'}$ such that e_i supports c_j but not $c_{j'}$, while $e_{i'}$ supports $c_{j'}$ but not c_j. Then the local evidence and claim views do not determine the total support of an assignment.*

Proof. The corresponding 2×2 support submatrix has a nonzero cross-difference. Apply Theorem 1. □

D WITNESS–COVERAGE PROOFS

The proofs use a single indistinguishability construction. Any uninspected corpus location can carry the unique exact match without changing the exposed trace. The same construction yields the inspection lower bound and the nearest-candidate corollary.

D.1 PROOF OF THE WITNESS–COVERAGE THEOREM

Proof. Let the observed trace inspect $S \subsetneq D$ and report $m_q(d) = 0$ for every $d \in S$. Choose one uninspected position $d^\star \in D \setminus S$. Construct world w_0 by assigning $m_q(d) = 0$ to every item in D. Construct world w_1 by keeping every inspected value unchanged and assigning $m_q(d^\star) = 1$. The trace is identical in the two worlds because the execution never inspects $d^\star$ and exposes no authenticated constraint on it. The global predicate differs: $T_\varnothing(w_0) = 1$ and $T_\varnothing(w_1) = 0$. □

Corollary 11 (Worst-case inspection cost)**.** *In an unstructured corpus of N independent items, a zero-error algorithm that has no trusted index must inspect all N items in the worst case to certify $T_\varnothing = 1$.*

Proof. Suppose the algorithm terminates after inspecting fewer than N items. At least one item remains uninspected. The construction in Theorem 3 produces two worlds with the same transcript and opposite no-match predicates. The algorithm must return the same certificate in both worlds and is therefore wrong in one of them. □

Corollary 12 (Nearest candidate is not negative evidence)**.** *Let $d^\star$ minimize a similarity distance to the query. The identity of $d^\star$ alone does not determine $T_\varnothing(D, q)$.*

Proof. Fix the same nearest candidate in two corpora. Add an exact match at an unreported location in one corpus and omit it in the other. The nearest-candidate message can remain unchanged while the no-match predicate changes. □

D.2 COVERAGE-EQUIVALENT STRUCTURES

The theorem does not require a literal scan of every record. A trusted index can summarize the corpus. An authenticated dictionary can provide a compact nonmembership proof. A transparency log can bind the proof to a corpus version. These mechanisms move the coverage cost into preprocessing and maintenance. They do not remove the coverage obligation.

E COLLISION BOUNDS AND STOCHASTIC INTERFACES

This appendix converts interface collisions into verifier limits. We first prove the deterministic paired-world result, then state the distributional version used for stochastic or adaptive interfaces.

E.1 PROOF OF THE COLLISION BOUND

Proof. Condition on a paired world (W^+, W^-). On the collision event, both worlds produce the same interface value x. Let the randomized verifier output label 1 with probability $p(x)$. Its error probability is $1 - p(x)$ in W^+ and $p(x)$ in W^-. The average error across the equally likely pair is therefore $1/2$. Outside the collision event, the error is nonnegative. Averaging over pairs gives

$$\mathrm{Err}_\gamma(V \circ X) \geq \frac{1}{2} \Pr[X(W^+) = X(W^-)], \tag{E.1}$$

which is Eq. (10). □

E.2 PROOF OF THE POST-PROCESSING COROLLARY

Proof. Whenever $X(W^+) = X(W^-)$, applying the same function g yields $g(X(W^+)) = g(X(W^-))$. Every collision of X is therefore also a collision of $Z = g(X)$. Taking probabilities gives the result. □

E.3 Stochastic interface version

Theorem 13 (Distributional collision bound)**.** *Let P_+ and P_- be the interface distributions in two equally likely worlds with opposite labels. The minimum error of any verifier is*

$$\mathrm{Err}^* = \frac{1 - \mathrm{TV}(P_+, P_-)}{2}. \tag{E.2}$$

Proof. This is the standard Bayes risk for testing two simple hypotheses with equal priors. The optimal acceptance region is a likelihood-ratio region. Its advantage over random guessing equals the total variation distance. □

The deterministic collision theorem is recovered when each world induces a point mass. The distributional result also extends to adaptive protocols by treating the complete query–response transcript as the interface signal.

F Minimum interfaces, integrity, and active verification

The main text uses residual collisions twice: first as a lower bound on answer-only verification, and then as the design object for a refined interface. This appendix proves the finite synthesis result, adds hiding constraints, and states the integrity boundary that turns an informative record into a certificate.

F.1 Proof of the residual collision-cover theorem

Proof. Assume first that every residual collision edge is cut by at least one selected field. Consider worlds $w, w' \in \mathcal{W}_0$ with equal refined interfaces (L, H_J). Equality of L means that either $T(w) = T(w')$ or $\{w, w'\}$ is a residual collision edge. The second case is impossible because some selected field cuts every residual edge, while equality of H_J says that no selected field differs. Hence $T(w) = T(w')$, and the refined interface is adequate.

Assume next that some residual edge $\{w, w'\}$ is not cut by any selected field. The edge definition gives $L(w) = L(w')$ and $T(w) \neq T(w')$. The absence of a cut gives $H_J(w) = H_J(w')$. The refined interface is equal in two worlds with different judgments. It is not adequate. □

Proposition 14 (Computational form)**.** *The minimum-cost interaction-interface problem in Eq. (12) is a weighted set-cover problem. It is NP-hard in general.*

Proof. Each candidate field corresponds to the subset of residual edges that it cuts. Feasibility requires the selected subsets to cover the universe $\mathcal{E}_T(L)$. The objective sums field costs. This is weighted set cover. NP-hardness follows by direct identification with an arbitrary weighted set-cover instance. □

Proposition 15 (Threat-model monotonicity)**.** *Let $\mathcal{W}_0 \subseteq \mathcal{W}_1$ and use the same predicate, local interface, candidate fields, and costs. The minimum interface cost on $\mathcal{W}_1$ is at least the minimum interface cost on $\mathcal{W}_0$.*

Proof. Every residual edge induced by $\mathcal{W}_0$ also appears in the restriction of the design on $\mathcal{W}_1$. A field set feasible for $\mathcal{W}_1$ must cover all edges from $\mathcal{W}_0$. The feasible family can only shrink when worlds are added. □

F.2 Hiding constraints

Let $R_{\mathrm{hide}} \subseteq \mathcal{W}_0 \times \mathcal{W}_0$ contain world pairs whose differences should not be exposed. A selected field set respects hiding when

$$(w, w') \in R_{\mathrm{hide}} \Longrightarrow L(w) = L(w') \text{ and } H_J(w) = H_J(w'). \tag{F.1}$$

The finite synthesis problem becomes a constrained set cover. Trust edges must be cut. Hiding pairs must remain uncut. Exact feasibility fails when one pair is simultaneously a trust edge and a required hiding pair. This makes the disclosure boundary explicit rather than treating maximal logging as the default.

F.3 No self-certification without integrity

Proposition 16 (No self-certification without integrity)**.** *Suppose two worlds w^+ and w^- satisfy $T(w^+) = 1$, $T(w^-) = 0$, and $X(w^+) = X(w^-)$. Suppose an untrusted service in w^- can emit every interaction message that it can emit in w^+. No verifier that observes only (X, H) can accept w^+ with probability one and reject w^- with probability one.*

Proof. Let h be the message emitted in w^+. By assumption, the service can emit the same h in w^-. The verifier then receives the same pair (X, h) in both worlds and must use the same decision distribution. Perfect acceptance in one world and perfect rejection in the other are impossible. □

The proposition separates semantic sufficiency from integrity. A record can preserve every relation needed by T and still be forgeable. Practical systems can bind the record to the execution with a signature from an independent gateway, an authenticated data structure, a trusted execution measurement, or an interactive challenge. The mechanism can vary. The record must be harder to copy than the answer.

F.4 Claim-specific interaction records

Positive evidence. A compact positive record can expose the current query and atomic claim, the identity and version of delivered evidence, the claim–evidence support judgment, the payload that reached the generator, and an integrity binding among these objects. The record need not expose the full corpus, prompt, or model state.

Negative evidence. A compact negative record can expose the normalized query, target corpus and version, requested scope, searched scope, retrieval completion or error state, match count or match type, and an integrity binding. A single Boolean can be sufficient on a narrow world graph. Threat-model monotonicity explains why broader operating regions may require more structure.

F.5 Adaptive probing

An active verifier can issue follow-up queries and choose each query from the previous transcript. The complete transcript becomes the interface X_π. All adequacy and collision statements apply to this transcript. Active probing can refine the interface partition. It cannot separate two worlds that induce the same transcript distribution under every admissible policy.

G Real-corpus protocol and supplementary analyses

This appendix provides the complete protocol behind Section 5. It follows the evidence chain used by the verifier: sources and frozen traces, public record semantics, residual-collision structure, output-equivalence checks, interface selection, gateway citations, evidence assignment, and integrity. The released artifact contains corpus snapshots, queries, frozen traces, prompts, responses, receipts, summaries, logs, and SHA-256 manifests.

G.1 Corpora, models, and frozen execution

The Python inputs are the official 3.10 and 3.11 Sphinx inventories. The Kubernetes inputs are the OpenAPI v2 specifications at the official v1.29.0 and v1.30.0 release tags. The Federal Register inputs are complete annual API snapshots for 2022 and 2023; unlike the two software domains, their documents are human-authored regulatory notices, proposed rules, rules, and presidential documents. Retrieval traces are generated and frozen before model inference.

The four frozen model families are Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3, Granite-3.3-8B-Instruct, and OLMo-2-1124-7B-Instruct. Greedy decoding is used in the primary study. The short-answer protocol permits at most 64 new tokens, and the long-form extension permits at most 192. Prompts expose retrieved passages or operational status text, but never expose the private entitlement predicate, intervention name, or verifier receipt. The auxiliary robustness cells separately use Qwen2.5-32B-Instruct in native BF16 and five Qwen2.5-7B nucleus runs with temperature 0.7,

Retriever	Answer collision	Min. fields	Zero-residual subsets	Full residual
BM25	56.56%	3 (S_1–S_3)	64	0/960
E5 dense	68.75%	3 (S_1–S_3)	64	0/960
BM25–E5 RRF	59.06%	3 (S_1–S_3)	64	0/960
ColBERTv2	50.63%	3 (S_1–S_3)	64	0/960

Table G.1: Retrieval-stack sensitivity of the complete 2^8 interaction-field enumeration. A residual edge requires equality of both normalized answer and the selected public fields.

top-p 0.9, and seeds 17/29/43/59/71. They reuse the frozen Federal Register worlds and are excluded from the primary generation count.

G.2 Primary retriever and matched retrieval stacks

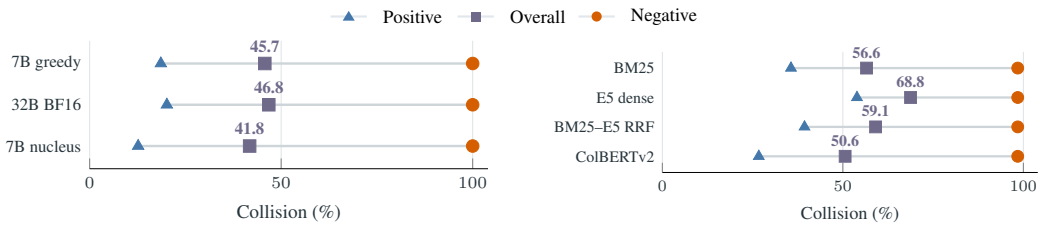
All three domains use the released `VersionedBM25Retriever`, a deterministic lexical BM25 implementation with $k_1 = 1.5$ and $b = 0.75$. The tokenizer applies the regular expression `[A-Za-z0-9_.-]+` and Unicode case folding; duplicate query terms are removed in first-occurrence order. BM25 scores the concatenated document title and text. Before scoring, the retriever retains documents whose version string exactly equals the routed version and whose scope set intersects the routed scopes. It returns the top two positive-score documents, sorting by decreasing score and then ascending `doc_id`; the nearest-candidate intervention instead returns the top one after excluding the expected document IDs. Complete negative claims bypass top-k ranking and scan every version- and scope-eligible document for a case-folded exact entity key.

The implementation is in-memory and unsharded: scopes are logical filters, not physical shards, and no dense encoder, approximate-nearest-neighbor index, or query rewriting is used. Each corpus version is an immutable snapshot loaded before trace generation. Thus freshness means exact binding to the named snapshot; the experiment makes no live-update, replication-lag, or read-consistency claim. Stale-cache and wrong-version routing are controlled interventions rather than measurements of an online cache.

For a matched retriever check on Kubernetes, we hold the 2,123-document corpus, 192 queries, top-k output, world interventions, exact-entity boost, Qwen2.5-7B generator, and greedy decoding fixed, and vary retrieval scoring and ranking. The dense cell uses `intfloat/e5-base-v2` with query/passage prefixes; the hybrid cell fuses BM25 and E5 by reciprocal-rank fusion (RRF constant 60, candidate depth 50); and the late-interaction cell uses ColBERTv2. For ColBERT, all 2,123 candidates are scored before the version/scope predicate, avoiding a top-256 truncation confound in partial-scope worlds.

Changing the retriever changes the empirical collision graph but does not eliminate it. Relative to BM25, exhaustive ColBERT changes overall collision by -5.94 percentage points (paired query-cluster 95% CI -11.04 to -1.04); the negative-pair difference is zero. The ordered BM25/ColBERT payload identity is 34.38%, mean top-k Jaccard is 41.93%, and nearest-candidate identity is 2.08%, so persistence is not explained by identical retrieved packets. Better recall may reduce how often individual failure worlds occur, but completion, snapshot, scope, status, and request-binding facts remain hidden unless the retrieval service exports them. These results are a matched robustness check, not a universal ranking of retrieval systems.

To test whether retrieval infrastructure changes the record itself, we rebuild H from each stack's frozen traces and enumerate the empty projection plus all 255 nonempty subsets. All four stacks select the same three tied minimum-cardinality covers, S_1–S_3 in Section G.6; every minimum and full H removes all 960 observed residual edges. The result is exact for this matched Kubernetes intervention graph, not a universal minimality claim for arbitrary retrievers.



(a) Model scale and decoding on fixed Federal Register pairs.

(b) Retrieval infrastructure on fixed Kubernetes pairs.

Figure G.1: Targeted robustness cells, excluded from the primary pooled count. Blue triangles, slate squares, and vermilion circles show positive, overall, and negative normalized-answer collision, respectively. Negative-pair collision is 100.00% in panel (a) and 98.44% in panel (b) for every cell: scale, sampling, and retrieval infrastructure change the magnitude without eliminating the interface failure.

G.3 Public interaction-record schema and canonicalization

The primary experiment separates generator-visible payloads from verifier-visible records. We therefore state the public record before reporting ablations over its fields. The interface object in receipt schema version `1.0` contains exactly the following eight keys. Each value is a JSON string; uppercase ASCII enum tokens are serialized verbatim.

The fields form four semantic groups. `query_binding` and `version_binding` connect the frozen search to the request that the verifier is judging. `scope_coverage`, `retrieval_status`, and `coverage` describe what the retrieval service actually attempted and whether that attempt completed. `claim_support` and `exact_match` state the claim-specific result of the delivered payload or exhaustive lookup. `integrity` records whether the trace carrying these relations was attested.

This grouping matters because no single object identifier carries all four obligations. A source ID can identify a document without showing that the document belongs to the current request. An empty source list can occur after a complete no-match, an error, or no search at all. The public record keeps these meanings separate while withholding raw prompts, private predicates, and source text.

Input identifiers, questions, versions, and scope strings are Unicode strings trimmed at their outer whitespace; empty strings and duplicate scopes are rejected. Version and scope comparisons are case-sensitive, while exact entity-key lookup is case-folded. The query commitment is the first 24 hexadecimal characters of SHA-256 over the UTF-8 compact-JSON array `["field-query", query_id, question, required_version, requested_scopes]`; no question case folding or scope reordering is applied. Compact JSON uses sorted object keys, no optional whitespace (`,` and `:` separators), and unescaped Unicode before UTF-8 encoding. The public interface is serialized under the same rule. Receipt projection follows the fixed field order shown in Table G.2.

Field	Allowed values	Derivation
`integrity`	`VERIFIED`, `UNVERIFIED`	`VERIFIED` iff the frozen trace's Boolean `trace_attested` is true.
`query_binding`	`MATCH`, `MISMATCH`, `UNAVAILABLE`	`UNAVAILABLE` for a missing or blank search commitment; otherwise byte-exact equality of search and request commitments.
`version_binding`	`MATCH`, `MISMATCH`, `UNAVAILABLE`	`UNAVAILABLE` for a missing or blank searched version; otherwise exact equality of searched and required version strings.
`scope_coverage`	`COMPLETE`, `PARTIAL`, `NONE`	`NONE` if the searched-scope set is empty; `COMPLETE` if every requested scope is in that set; `PARTIAL` otherwise.
`retrieval_status`	`SUCCESS`, `NO_MATCH`, `MATCH`, `EMPTY`, `TIMEOUT`, `ERROR`, `NO_SIGNAL`, `NEAREST_ONLY`	The gateway's terminal retrieval state. `MATCH`/`NO_MATCH` are outcomes of the exhaustive exact-key path; `NEAREST_ONLY` marks the exclusion-based nearest-candidate path.
`coverage`	`COMPLETE`, `INCOMPLETE`	`COMPLETE` iff the Boolean `coverage_complete` is true; in this study only the exhaustive negative lookup sets it.
`claim_support`	`PRESENT`, `ABSENT`, `NOT_CHECKED`, `NOT_APPLICABLE`	For a positive claim, `PRESENT` iff an expected document ID occurs among the delivered hits; timeout, error, or no-signal gives `NOT_CHECKED`. Negative claims give `NOT_APPLICABLE`.
`exact_match`	`PRESENT`, `ABSENT`, `NOT_CHECKED`, `NOT_APPLICABLE`	For a negative claim with complete coverage, `PRESENT` iff the exact-key result list is nonempty, else `ABSENT`; incomplete coverage gives `NOT_CHECKED`. Positive claims give `NOT_APPLICABLE`.

Table G.2: Exact schema of the eight-field public interface H. Container metadata such as `schema_version`, `record_id`, `query_id`, and `claim_mode` is outside the eight-field projection. Raw commitments, intervention labels, expected answers, and the private predicates T_+ and T_- are not exposed.

	Qwen	Mistral	Granite	OLMo	All
Stale cache	100	100	100	100	100
Wrong version	4.7	0	10.9	10.2	6.4
Partial scope	4.7	0	10.9	12.5	7
Timeout	7.8	0	10.9	10.2	7.2
Nearest only	4.7	0	10.9	18	8.4

(a) Python.

	Qwen	Mistral	Granite	OLMo	All
Stale cache	92.2	100	95.3	90.6	94.5
Wrong version	33.6	14.1	82	21.9	37.9
Partial scope	33.6	12.5	82	12.5	35.2
Timeout	33.6	14.1	82	25	38.7
Nearest only	32.8	13.3	78.1	39.8	41

(b) Kubernetes.

	Qwen	Mistral	Granite	OLMo	All
Stale cache	93	100	99.2	88.3	95.1
Wrong version	0	5.5	8.6	3.9	4.5
Partial scope	0	3.1	8.6	3.1	3.7
Timeout	0	3.1	8.6	3.9	3.9
Nearest only	0	21.1	18.8	5.5	11.3

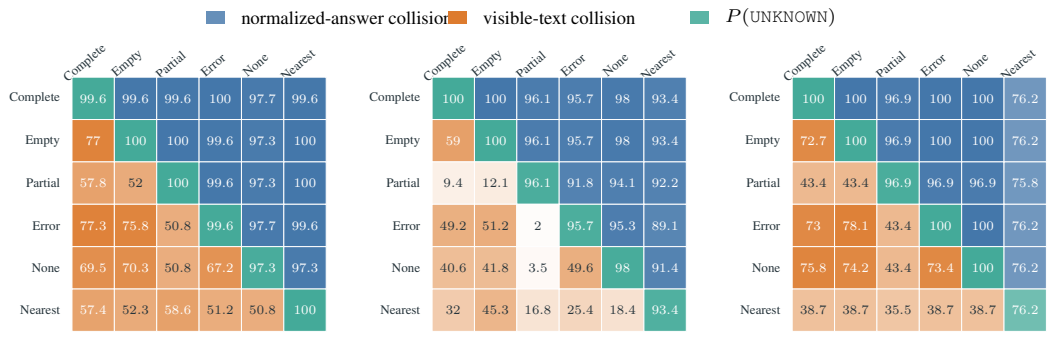
(c) Federal Register.

Figure G.2: Positive normalized-answer collision by alternative retrieval condition and model family. The ALL column pools the four models within each domain.

G.4 CONDITION- AND DOMAIN-LEVEL RESIDUAL STRUCTURE

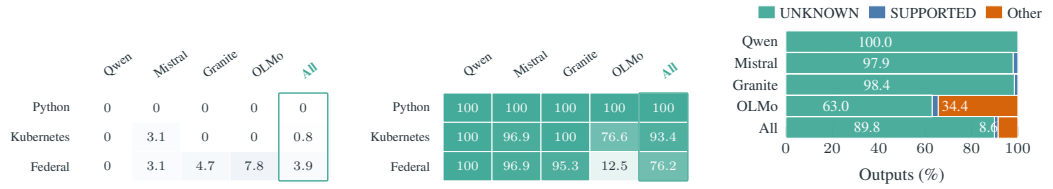
With the execution and public record fixed, we expose the condition-level collision geometry pooled in the main text. The main text reports the primary graph by model family; the following panels show the complementary domain and condition views using the same visual grammar as the private-world study. In positive pairs, stale-cache routing is almost completely invisible, while wrong-version, partial-scope, timeout, and nearest-only collisions vary sharply by corpus and model (Fig. G.2). In negative pairs, normalized endpoints merge most states even when the visible wording changes (Fig. G.3).

Nearest-only behavior differs materially by corpus: Python produces no `SUPPORTED` responses, while Federal Register reaches 3.91% pooled and also exposes OLMo parsing failures. Figure G.4 reports both the small false-support component and the much larger abstention component rather than transferring the private-world adoption rate to public traffic.



(a) Python. **(b)** Kubernetes. **(c)** Federal Register.

Figure G.3: Negative-evidence observation classes pooled over the four model families within each real corpus. The encoding matches Fig. H.2.



(a) Rate of SUPPORTED. **(b)** Rate of UNKNOWN. **(c)** Complete output composition by model family.

Figure G.4: Nearest-candidate-only negative queries. Domain and model stratification separates the rare false-support response from abstention and other or unparsed output.

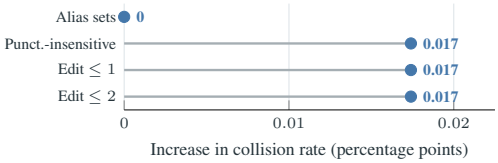
Domain	Exact answer	Punct.-insens.	Edit ≤ 2	Exact text	Text edit ≤ 2
Python	50.31	50.31	50.31	37.29	37.34
Kubernetes	65.18	65.18	65.18	24.14	24.27
Federal Register	47.34	47.40	47.40	33.59	33.96
Pooled	54.28	54.30	54.30	31.68	31.86

Table G.3: Collision sensitivity across four model families (percent). The deterministic relaxations change pooled normalized-answer collision by at most 0.02 percentage points.

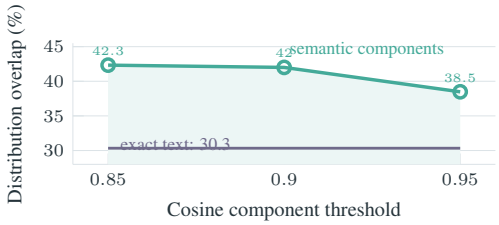
G.5 Output-equivalence sensitivity

The main result uses exact normalized-answer equality. We test whether conservative lexical or semantic equivalences materially change the measured overlap. The primary normalized interface uses case-folded, whitespace-normalized equality. We recompute collision using a conservative alias map, Unicode punctuation-insensitive equality, and Levenshtein thresholds one and two. The visible-text analysis uses exact equality and the same small-edit tolerances.

To allow free-form lexical variation, we also compare the five nucleus runs using `all-mpnet-base-v2` cosine components and report $1-$empirical TV between the paired output distributions. At threshold 0.90, exact visible-text overlap is 30.33%, while semantic overlap is 42.00% (query-cluster 95% CI 37.29–46.38); positive and negative semantic overlap are 20.38% and 85.25%. Overall semantic overlap is 42.33%, 42.00%, and 38.48% at thresholds 0.85, 0.90, and 0.95, respectively. Embedding components are an operational equivalence rule, not ground-truth semantic identity.



(a) Change from exact normalized-answer equality.



(b) Free-form distribution overlap under semantic components.

Figure G.5: Output-equivalence sensitivity. Conservative deterministic equivalences change the headline collision rate by at most 0.02 percentage points, while semantic components increase measured overlap relative to exact visible-text identity.

G.6 Held-out interface selection and fixed verification rules

The residual graph now becomes a design object. We enumerate all public-field projections on two domains and test the selected interfaces on the third. For each held-out domain, we enumerate the empty projection and all nonempty subsets of the eight public fields. On the two training domains we retain every minimum-cardinality set with zero residual collision. These tied sets are frozen before held-out evaluation. All three folds select the same three tied minima: $S_1 = \{\texttt{query_binding}, \texttt{retrieval_status}, \texttt{claim_support}\}$, $S_2 = \{\texttt{query_binding}, \texttt{coverage}, \texttt{claim_support}\}$, and $S_3 = \{\texttt{query_binding}, \texttt{claim_support}, \texttt{exact_match}\}$.

Held-out domain	Answer-only	Min. fields	Selected sets	Held-out residual	Full residual
Federal Register	1,818/3,840	3	S_1–S_3	0–0	0
Kubernetes	2,503/3,840	3	S_1–S_3	0–0	0
Python	1,932/3,840	3	S_1–S_3	0–0	0

Table G.4: Leave-one-domain-out evaluation over all $2^8 = 256$ public-field subsets. The held-out residual range covers all tied minimum-cardinality sets selected on the two training domains.

Dropping query binding from the full record restores 512, 484, and 487 collisions out of 3,840 when Python, Kubernetes, and Federal Register are held out. Dropping any other single field restores no collision in these finite folds because retained fields are redundant on the observed graph. Field responsibilities should therefore be read from the declared trust predicate, not from one finite drop-one result.

G.7 Gateway citations and their residual boundary

The long-form experiment asks how much object-level provenance can recover before execution relations are exposed. Each response begins with a parseable `SUPPORTED` or `UNKNOWN` endpoint and continues with two to four explanatory sentences. The gateway deterministically appends the sorted, deduplicated set of all delivered source identifiers and versions. An independent audit reconstructs these citations with 100% precision and recall and finds zero no-source violations.

Parsing the first rather than the last endpoint changes one answer pair and no answer-plus-citation pair: answer collision is 57.36% rather than 57.43%, while the citation result remains 28.26%. A separate pilot in which models copied their own citations failed its predeclared integrity gate; strict delivered-citation precision was 19.02%. This failure motivated gateway rendering and is not used in the main effect estimate.

Same-protocol executable verifier audit. For each of the 1,728 long-form outputs, the entitlement target is true only for a bound positive world or a complete-no-match negative world. The answer-only rule accepts `SUPPORTED` for a positive claim and `UNKNOWN` for a negative claim. The validated-citation rule additionally requires a positive answer to carry at least one syntactically valid gateway

Model	Pairs	Answer collision	Answer + citation
Qwen2.5-7B	360	46.94	28.33
Mistral-7B-v0.3	360	47.50	28.06
Granite-3.3-8B	360	54.44	28.33
OLMo-2-7B	360	80.83	28.33
Pooled	1,440	57.43	28.26

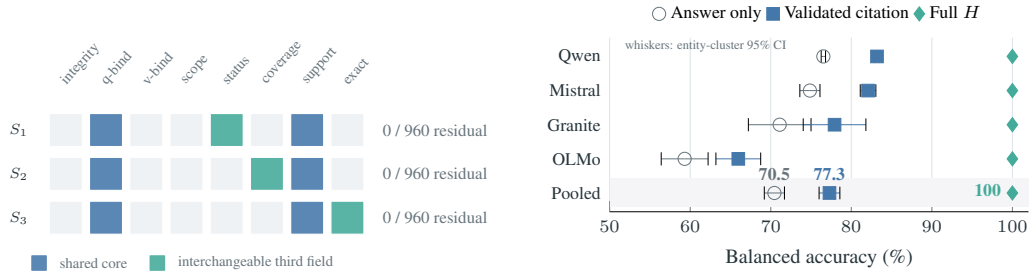
Table G.5: Long-form endpoint and gateway-citation collision by model family (percent).

Paired condition	Answer collision	Answer + citation	Reduction
Empty top-k	98.96	98.96	0.00
No signal	100.00	100.00	0.00
Retrieval error	100.00	100.00	0.00
Stale cache	97.40	62.50	34.90
Nearest candidate	47.57	0.00	47.57
Partial scope	46.18	0.00	46.18
Timeout	23.44	0.00	23.44
Wrong version	19.79	0.00	19.79

Table G.6: Where gateway-rendered source identity and version remove or retain long-form collisions (percent).

Model	Answer only	Validated citation	Full H
Qwen2.5-7B	76.53 [76.25, 76.67]	83.19 [82.92, 83.33]	100.00
Mistral-7B-v0.3	74.86 [73.61, 75.97]	82.08 [81.11, 82.92]	100.00
Granite-3.3-8B	71.11 [67.22, 74.58]	77.92 [74.03, 81.39]	100.00
OLMo-2-7B	59.31 [56.67, 62.22]	65.97 [63.33, 68.89]	100.00
Pooled	70.45 [69.20, 71.67]	77.29 [76.01, 78.58]	100.00

Table G.7: Balanced accuracy and entity-cluster bootstrap 95% intervals for deterministic entitlement-verifier rules on the same long-form world outputs. Each model contributes 432 worlds (72 entitled and 360 non-entitled); pooled resampling keeps each entity's four model outputs together. Full-H intervals are [100, 100].



(a) Tied minimum field sets, unchanged across retrievers.

(b) Observed balanced accuracy on the same world graph.

Figure G.6: Interaction-record robustness and executable verifier performance. Panel (a) reports a finite zero-residual cover, while panel (b) evaluates fixed decision rules; neither is an out-of-distribution deployment guarantee.

citation whose version equals the version named in the request; a negative answer must carry no citation. It observes no trace status, coverage, query commitment, or support relation. The full-H rule is the executable policy in Section G.3. All rules are fixed before scoring; no classifier is trained.

Pooled answer-only sensitivity/specificity are 85.42%/55.49%; validated citations retain 85.42% sensitivity and raise specificity to 69.17%, but leave 444 false positives. Full H has 100%/100%. Binary Brier scores are 0.395, 0.281, and 0, respectively. Thus citation validation removes source/version errors without certifying the execution relations that remain outside the citation interface.

G.8 EVIDENCE-ASSIGNMENT INTERVENTION

Entitlement is an interface claim rather than a causal-use claim. We therefore separately test whether payload assignment has a black-box effect on the generated endpoint. We freeze 192 positive entities across the three domains and generate four prompts per entity while holding the question and decoding configuration fixed: exact evidence, identifier-masked evidence, a length-matched entity-neutral placebo, and valid evidence swapped from another entity. Across four model families, the 3,072 outputs have pooled support rates of 68.62%, 10.68%, 6.77%, and 8.59%, respectively. Domain/entity-clustered exact-minus-control effects are +57.94 percentage points (95% CI 53.39–62.37), +61.85 (56.90–66.67), and +60.03 (55.60–64.32). The estimand is output sensitivity to payload assignment under the frozen prompt.

G.9 INTEGRITY, AUTHENTICATED ABSENCE, AND PORTABLE ARTIFACTS

The final analyses ask whether an informative record can be bound to a snapshot and transported as a verifiable artifact.

Authenticated negative coverage under churn. An Ed25519-signed snapshot manifest commits to the snapshot ID, corpus version, epoch, and exhaustive shard list for each searched scope. Every enumerated shard supplies an ordered-Merkle predecessor/successor adjacency proof for exact-key nonmembership, and verification binds the request, scope, version, snapshot, and completion state to verifier-selected freshness expectations. Across 64 negative queries and eight shards per scope, valid and unrelated-churn witnesses are accepted in 100% of trials, while omitted shards, proof tampering, stale replay, and insertion of the target key are detected in 100%. Median witness size is 31.2 KiB; median build and verification time are 1.68 ms and 0.28 ms. The certificate proves exact-key absence only in the enumerated scopes of one authenticated snapshot, not future absence, semantic non-relevance, or honest ingestion.

Manifest and transparency-log references. The interaction record can carry a hash-bound reference to a provenance manifest, together with a transparency-log entry digest, Merkle inclusion proof, log identity, and signed checkpoint as in the Rekor ecosystem (Newman et al., 2022). Across 256 records, valid acceptance and detection of payload, manifest, H-core, inclusion-proof, and checkpoint tampering are all 100%. The median referenced record is 2,452 bytes and median local verification is 81.1 microseconds. This tests binding and transport compatibility; it is not a public Rekor deployment.

Portable world graph. The release package contains 192 queries, 1,152 condition worlds, 4,800 nodes, 5,568 edges, and 960 paired entitlement-flip edges, together with frozen traces, responses, SHA-256 checksums, and an independent verifier. The verifier accepts the exported package and rejects checksum drift, count drift, dangling edges, or pairs that fail to flip the declared entitlement predicate.

G.9.1 SIGNED-RECEIPT THREAT MODEL

The application or generator may alter, omit, or replay a receipt but cannot access the independent gateway's Ed25519 private key. The verifier holds an authentic public key and independently knows the expected request identifier, query identifier, nonce, request commitment, and evidence digest. The 47,808 attacks comprise per-field modifications, unsupported-to-entitled rewrites, evidence-digest substitution, signature corruption or removal, cross-request replay, and expiry across the three domains. Authenticity and request-context verification precede the semantic entitlement policy.

The signed construction addresses application-level forgery under key separation. Operational deployments additionally require nonce freshness, rotation and revocation, gateway auditing, and access control over receipt fields. A dishonest gateway, active-key compromise, or collusion inside the trusted boundary lies outside this experiment.

Table G.8: Executable interaction-record checks pooled over three domains.

Check	Result
Rule-verifier agreement	13,824 / 13,824
Full-H residual collisions	0 / 11,520
Genuine signed envelopes accepted	3,456 / 3,456
Application-level attacks rejected	47,808 / 47,808
Unsigned entitlement forgeries accepted	2,880 / 2,880
Signed entitlement forgeries rejected	2,880 / 2,880
Median envelope / p95 verification	≈1.1 KiB / 0.23–0.36 ms

H PRIVATE-WORLD MECHANISM-ISOLATION STUDY

The real-corpus study in the main text is the primary empirical evaluation. This appendix asks a narrower question under stronger control: can the answer remain unchanged when we alter only the evidence relation? We replace public facts with two generated knowledge worlds, so the task answer and every trust label are known by construction. The study isolates positive query binding, negative search coverage, and nearest-candidate ambiguity without relying on parametric knowledge of public entities.

H.1 STUDY DESIGN

Private knowledge worlds. Worlds 303 and 404 are generated independently. Each contains 80 fictional agents, 40 receptors, 20 pathways, 20 conditions, 20 adverse events, and 16 contraindication targets. Every entity has two aliases. Five relations connect the entities: BINDS RECEPTOR, ACTIVATES PATHWAY, TREATS CONDITION, CAUSES ADVERSE EVENT, and CONTRAINDICATED WITH. Each query requests one canonical identifier.

Models and execution. We evaluate Qwen3-8B, Mistral-7B-Instruct-v0.3, Granite-3.3-8B-Instruct, and OLMo-2-1124-7B-Instruct on both worlds. The eight model–world units run independently with greedy decoding and at most 24 new tokens. Exact revisions and record counts appear in Table H.7.

Interfaces and predicates. The normalized interface X_{norm} contains the parsed canonical answer. The text interface X_{text} contains the full visible model output. Positive entitlement $T_+ = 1$ requires supporting evidence bound to the current query and claim. Negative entitlement $T_- = 1$ requires a complete, query-bound no-match search over the declared scope and corpus version. These predicates judge the evidence relation, not accidental answer correctness.

Table H.1: Controlled contrasts and stress conditions. Each contrast holds the task answer fixed while changing the evidence relation that the answer interface may hide.

Contrast	Held fixed	Changed	Target	N	Source
Ordinary positive binding	Query, target fact, model, and world	Query–claim binding of the same fact packet	T_+	5,120	`anchors_extended`
Exact-match positive control	Query, entity, fact, model, and world	Binding attestation for the same exact-match fact	T_+	768	`p8_core`
Complete vs. unsupported no-match	Unknown query, ground-truth answer, model, and world	Search completeness or retrieval status	T_-	7,680	`p8_core`
Nearest-candidate stress	Unknown query, absent entity, model, and world	A nearest known candidate is placed in the evidence position	diagnostic	1,536	`p8_core`

Note. T_+ asks whether the current answer has query-bound positive evidence. T_- asks whether an `UNKNOWN` answer has complete, query-bound negative evidence.

The ordinary positive contrast keeps the query, model, target fact, evidence identifier, source version, and answer fixed. It changes only query–claim binding. The exact-match positive control keeps the same exact retrieval result and changes its attestation. The negative contrast pairs a complete no-match record with unexplained empty output, partial search, a plain-text no-match message, retrieval error, or no retrieval signal. A separate nearest-candidate stress condition places one similar known entity in the evidence position.

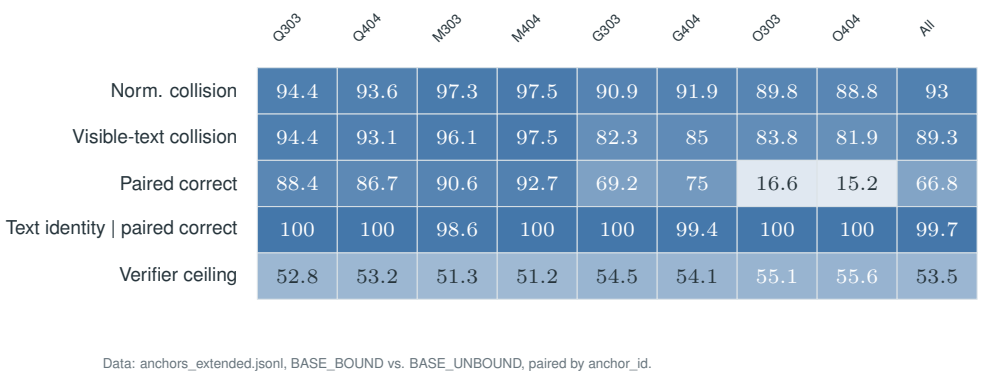
We compute trust-relevant collision for X_{norm} and X_{text}, paired correctness, collision conditional on valid output syntax, and the answer-only verifier ceiling from Theorem 4. For negative states, we additionally evaluate `UNKNOWN` as a certificate for $T_- = 1$ and measure adoption of the nearest candidate.

H.2 Positive binding remains invisible at the endpoint

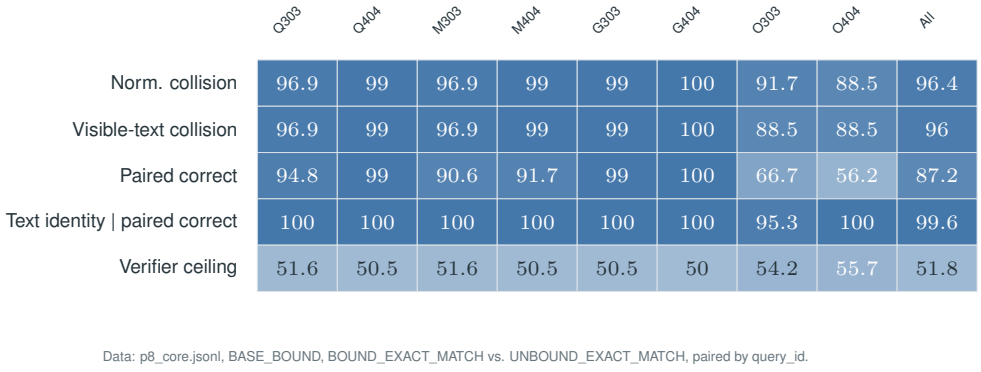
The main text uses public-corpus examples. This study supplies a stronger mechanism-isolation counterpart in which the paired fact and every trust label are known by construction. On 5,120 ordinary binding pairs, normalized answers collide in 93.03% and complete visible text collides in 89.26%. Both worlds are correct in 66.80% of pairs; within those paired-correct cases, the visible text is identical in 99.68%. The answer-only verifier ceiling is 53.49%.

The 768 exact-match controls strengthen the isolation. Their normalized collision is 96.35%, their visible-text collision is 95.96%, and 87.24% of pairs are correct on both sides. The resulting verifier ceiling is 51.82%, close to random guessing on the balanced paired task.

Per-run and relation-level estimates are reported with the complete stratifications in Tables H.5 and H.6. They show that model quality changes paired correctness, but no model family or relation type makes the endpoint informative about binding.



(a) Ordinary positive-binding anchors.



(b) Known-entity exact-match controls.

Figure H.1: Positive binding is largely invisible at both answer interfaces. Each column is one model–world unit; ALL pools the eight units.

H.3 NEGATIVE STATES COLLAPSE INTO THE SAME ABSTENTION

A complete no-match record and an unsupported alternative produce the same normalized answer in 82.94% of 7,680 pairs. The visible text collides in 49.61%. Both worlds are correct in 71.94% of pairs because every queried entity is absent. Conditional on both outputs satisfying the answer schema, normalized collision reaches 100%. Correctness survives while the basis for correctness disappears.

The five unsupported states are not equally visible. Complete no-match collides with unexplained empty output in 91.4% of pairs, with partial no-match in 94.0%, with retrieval error in 91.4%, with a plain-text no-match message in 74.5%, and with no signal in 63.3%. Wording changes model behavior, but it does not align the answer partition with T_{-}.

Abstention is not a certificate. Across the seven negative states, the prior probability of $T_{-} = 1$ is 14.3%. After observing `UNKNOWN`, it rises only to 16.7%. The rule `UNKNOWN`$\Rightarrow T_{-} = 1$ reaches 56.6% balanced accuracy. The model can abstain correctly without establishing why abstention is warranted.

A nearest candidate is not negative evidence. When the packet contains one similar known entity, the models adopt that entity's answer in 45.25% of cases and return `UNKNOWN` in 22.20%. The packet turns similarity into apparent evidence because the interface does not identify the item as a non-match.

Table H.2: Trust-relevant collisions at the answer interface. High collision is worse for interface adequacy. Boldface marks the pooled headline findings rather than a performance winner.

Controlled contrast	N	Norm. collision	Text collision	Valid-only norm. coll.	Paired correct	Text id. given correct	Verifier ceiling
Positive binding: ordinary anchors	5,120	**93.03**	89.26	96.03	66.80	**99.68**	**53.49**
Positive binding: exact-match controls	768	**96.35**	95.96	98.46	87.24	**99.55**	**51.82**
Negative coverage: pooled unsupported states	7,680	**82.94**	49.61	100.00	71.94	**63.04**	**58.53**

All entries are percentages except N. The verifier ceiling is $100(1 - C_T(X_{\text{norm}})/2)$. "Both valid" conditions the normalized collision on both outputs satisfying the answer schema.



(a) Qwen3-8B.

(b) Mistral-7B-Instruct-v0.3.



(c) Granite-3.3-8B-Instruct.

(d) OLMo-2-1124-7B-Instruct.

Figure H.2: Interface-induced observation classes over seven negative-evidence states. Upper triangles report normalized-answer collision, lower triangles visible-text collision, and diagonals $P(\texttt{UNKNOWN})$. Only CERT satisfies $T_{-} = 1$.

H.4 INTERFACE AUGMENTATION SEPARATES THE CONTROLLED WORLDS

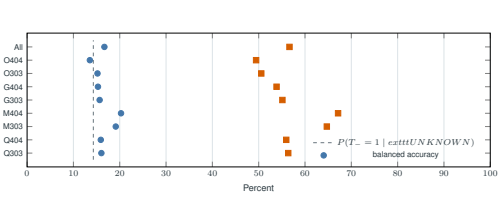
The controlled graph allows a direct interface ablation. Evidence identity and source version leave positive collisions intact because both are shared by the paired worlds. An unbound payload hash also leaves them intact. A digest that binds the current query, claim, and delivered payload reduces positive residual collision from 93.03% to zero. For negative coverage, corpus version gives partial separation, retrieval status adds more, and the complete scope–version–completion–status record reduces the residual from 82.94% to zero.

Model-visible citation strings do not substitute for this verifier record. Asking the generator to emit `evidence_id@source_version` produces more than 99% exact citations but leaves 93.09–93.24% answer-plus-valid-citation collision. Conversely, removing binding fields from the generator prompt while retaining them only in H makes paired prompts byte-identical and yields 100% answer and visible-text collision. These results isolate the role of the verifier-facing interaction relation.

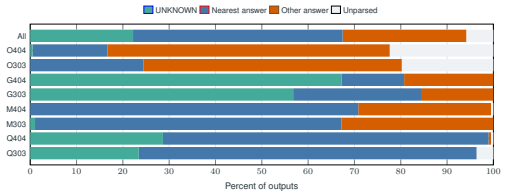
	Q303	Q404	M303	M404	G303	G404	O303	O404	All
Cert	100	100	99.5	99.5	100	100	16.1	17.7	79.1
Empty	100	100	100	100	100	100	28.1	33.9	82.7
Partial	100	100	100	99.5	100	100	25	28.6	81.6
Text	100	100	76.6	60.4	99	99	9.9	10.4	69.4
Error	100	100	100	99.5	100	100	24	34.4	82.2
None	100	100	42.7	31.8	82.8	87.5	3.1	5.2	56.6
Nearest	23.4	28.6	1	0	56.8	67.2	0	0.5	22.2

Rows are evidence states. Only Cert corresponds to $T_{-} = 1$.

(a) Probability of UNKNOWN under each evidence state.



(b) Diagnostic value of UNKNOWN for $T_{-} = 1$.



(c) Outputs when the packet contains only a nearest candidate.

Figure H.3: Abstention and nearest-candidate behavior. The endpoint reacts to evidence presentation but does not preserve the coverage relation required by the trust predicate.

Table H.3: Diagnostic value of the rule UNKNOWN $\Rightarrow T_{-} = 1$. Precision measures how often an observed UNKNOWN is backed by complete negative evidence.

Run	N	Prior $P(T_{-} = 1)$	UNKNOWN rate	TPR	FPR	Specificity	Precision	Balanced accuracy
Q303	1,344	14.3	89.1	100.0	87.2	12.8	16.0	56.4
Q404	1,344	14.3	89.8	100.0	88.1	11.9	15.9	55.9
M303	1,344	14.3	74.3	99.5	70.1	29.9	19.1	64.7
M404	1,344	14.3	70.1	99.5	65.2	34.8	20.3	67.1
G303	1,344	14.3	91.2	100.0	89.8	10.2	15.7	55.1
G404	1,344	14.3	93.4	100.0	92.3	7.7	15.3	53.9
O303	1,344	14.3	15.2	16.1	15.0	85.0	15.2	50.6
O404	1,344	14.3	18.7	17.7	18.8	81.2	13.5	49.4
All	**10,752**	**14.3**	**67.7**	**79.1**	**65.8**	**34.2**	**16.7**	**56.6**

Table H.4: Minimal interface-augmentation ablations. Residual collision requires equality of both the answer and the listed record. Citation values report answer-plus-valid-citation collisions from two generator-side runs; all other rows project fixed outputs onto verifier-side records.

Contrast	Added record	Pairs	Baseline collision	Residual collision
Positive binding	Evidence ID + source version (generated)	5,120	93.03%	93.09–93.24%
Positive binding	Delivered-payload hash	5,120	93.03%	93.03%
Positive binding	Query/claim-bound payload digest	5,120	93.03%	0.00%
Negative coverage	Corpus version	7,680	82.94%	37.08%
Negative coverage	Retrieval status + corpus version	7,680	82.94%	18.80%
Negative coverage	Scope + version + completion/status	7,680	82.94%	0.00%

	Q303	Q404	M303	M404	G303	G404	O303	O404
ED1	71.9	78.1	65.6	65.6	34.4	18.8	28.1	9.4
ED2	68.8	59.4	68.8	62.5	25	12.5	25	21.9
ED3	65.6	53.1	59.4	71.9	31.2	12.5	18.8	12.5
Phon	62.5	68.8	65.6	78.1	31.2	6.2	28.1	18.8
Morph	68.8	71.9	65.6	78.1	9.4	9.4	25	15.6
Far	100	90.6	71.9	68.8	34.4	21.9	21.9	18.8

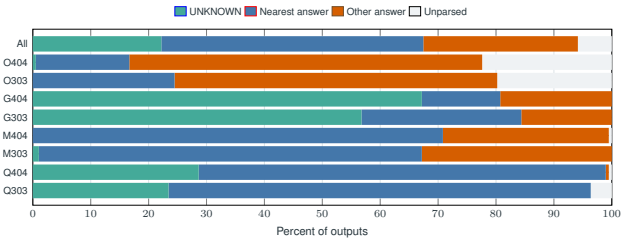
Data: p8_core.jsonl, BASE_BOUND, state = NEAREST_CANDIDATE_ONLY.

(a) Probability of returning the nearest known answer.

	Q303	Q404	M303	M404	G303	G404	O303	O404
ED1	21.9	21.9	3.1	0	46.9	59.4	0	0
ED2	31.2	34.4	0	0	65.6	71.9	0	3.1
ED3	34.4	46.9	3.1	0	59.4	71.9	0	0
Phon	34.4	31.2	0	0	46.9	75	0	0
Morph	18.8	28.1	0	0	75	75	0	0
Far	0	9.4	0	0	46.9	50	0	0

Data: p8_core.jsonl, BASE_BOUND, state = NEAREST_CANDIDATE_ONLY.

(b) Probability of returning UNKNOWN.



(c) Complete output composition by model–world unit.

Figure H.4: Nearest-candidate behavior across entity-boundary types. The model-family effect is larger than the distance-bin effect, which points to the evidence position rather than string distance alone.

H.5 Complete stratified results

This subsection records the remaining stratifications used to audit the pooled claims. The nearest-candidate figure separates model-family effects from lexical and phonetic distance. The run manifest then fixes model revisions and record counts, followed by the complete negative-state and relation-level tables.

Table H.5: Positive-binding collisions by model–world run. The two blocks use the ordinary-anchor contrast and the exact-match positive control, respectively.

Run	N	Bound acc.	Unbound acc.	Both valid	Norm. collision	Text collision	Paired correct	Text id. given correct	Ceiling
Ordinary positive binding									
Q303	640	93.0	89.1	97.5	94.4	94.4	88.4	100.0	52.8
Q404	640	91.4	86.7	97.5	93.6	93.1	86.7	100.0	53.2
M303	640	93.1	90.6	99.5	97.3	96.1	90.6	98.6	51.3
M404	640	93.4	94.4	100.0	97.5	97.5	92.7	100.0	51.2
G303	640	75.8	71.2	71.4	90.9	82.3	69.2	100.0	54.5
G404	640	81.9	75.9	75.5	91.9	85.0	75.0	99.4	54.1
O303	640	21.4	17.2	33.9	89.8	83.8	16.6	100.0	55.1
O404	640	22.3	15.6	26.9	88.8	81.9	15.2	100.0	55.6
All	**5,120**	**71.5**	**67.6**	**75.3**	**93.0**	**89.3**	**66.8**	**99.7**	**53.5**
Exact-match positive control									
Q303	96	97.9	94.8	96.9	96.9	96.9	94.8	100.0	51.6
Q404	96	100.0	99.0	99.0	99.0	99.0	99.0	100.0	50.5
M303	96	92.7	90.6	99.0	96.9	96.9	90.6	100.0	51.6
M404	96	92.7	91.7	100.0	99.0	99.0	91.7	100.0	50.5
G303	96	99.0	100.0	100.0	99.0	99.0	99.0	100.0	50.5
G404	96	100.0	100.0	100.0	100.0	100.0	100.0	100.0	50.0
O303	96	70.8	68.8	79.2	91.7	88.5	66.7	95.3	54.2
O404	96	60.4	60.4	71.9	88.5	88.5	56.2	100.0	55.7
All	**768**	**89.2**	**88.2**	**93.2**	**96.4**	**96.0**	**87.2**	**99.6**	**51.8**

Table H.6: Positive-binding robustness across relation types. Results use the ordinary-anchor contrast and pool all eight model–world runs.

Relation	N	Norm. collision	Text collision	Paired correct	Text id. given correct	Ceiling
Binds receptor	1,024	86.0	84.9	69.7	100.0	57.0
Activates pathway	1,024	94.1	93.1	77.4	99.6	52.9
Treats condition	1,024	96.4	91.4	68.8	100.0	51.8
Adverse event	1,024	96.6	91.5	62.0	100.0	51.7
Contraindication	1,024	92.0	85.4	56.0	98.6	54.0
All relations	**5,120**	**93.0**	**89.3**	**66.8**	**99.7**	**53.5**

Table H.7: Run manifest for the final 8×B200 experiment. Every worker completed its assigned model–world cell.

Run	GPU	Model	World	Revision	Anchors initial	Anchors extended	P8 core	Status
Q303	0	Qwen/Qwen3-8B	303	`b968826d`	5,120	5,120	4,944	**COMPLETE**
Q404	1	Qwen/Qwen3-8B	404	`b968826d`	5,120	5,120	4,944	**COMPLETE**
M303	2	mistralai/Mistral-7B-Instruct-v0.3	303	`c170c708`	5,120	5,120	4,944	**COMPLETE**
M404	3	mistralai/Mistral-7B-Instruct-v0.3	404	`c170c708`	5,120	5,120	4,944	**COMPLETE**
G303	4	ibm-granite/granite-3.3-8b-instruct	303	`51dd4bc2`	5,120	5,120	4,944	**COMPLETE**
G404	5	ibm-granite/granite-3.3-8b-instruct	404	`51dd4bc2`	5,120	5,120	4,944	**COMPLETE**
O303	6	allenai/OLMo-2-1124-7B-Instruct	303	`470b1fba`	5,120	5,120	4,944	**COMPLETE**
O404	7	allenai/OLMo-2-1124-7B-Instruct	404	`470b1fba`	5,120	5,120	4,944	**COMPLETE**

Table H.8: Negative-evidence collisions by run and unsupported state. Every row compares the stated alternative with BOUND COMPLETE NO-MATCH.

Run	Unsupported state	N	Norm. collision	Text collision	Paired correct	Cert. accuracy	Alt. accuracy	Ceiling
Q303	Empty / unexplained	192	100.0	100.0	100.0	100.0	100.0	50.0
	Partial no-match	192	100.0	100.0	100.0	100.0	100.0	50.0
	Retrieval error	192	100.0	100.0	100.0	100.0	100.0	50.0
	Plain-text no-match	192	100.0	100.0	100.0	100.0	100.0	50.0
	No signal	192	100.0	100.0	100.0	100.0	100.0	50.0
Q404	Empty / unexplained	192	100.0	100.0	100.0	100.0	100.0	50.0
	Partial no-match	192	100.0	100.0	100.0	100.0	100.0	50.0
	Retrieval error	192	100.0	100.0	100.0	100.0	100.0	50.0
	Plain-text no-match	192	100.0	100.0	100.0	100.0	100.0	50.0
	No signal	192	100.0	100.0	100.0	100.0	100.0	50.0
M303	Empty / unexplained	192	99.5	92.7	99.5	99.5	100.0	50.3
	Partial no-match	192	99.5	96.9	99.5	99.5	100.0	50.3
	Retrieval error	192	99.5	89.1	99.5	99.5	100.0	50.3
	Plain-text no-match	192	77.1	22.4	76.6	99.5	76.6	61.5
	No signal	192	42.2	1.0	42.2	99.5	42.7	78.9
M404	Empty / unexplained	192	99.5	94.8	99.5	99.5	100.0	50.3
	Partial no-match	192	99.5	97.4	99.5	99.5	99.5	50.3
	Retrieval error	192	100.0	87.0	99.5	99.5	99.5	50.0
	Plain-text no-match	192	60.9	18.2	60.4	99.5	60.4	69.5
	No signal	192	32.3	0.0	31.8	99.5	31.8	83.9
G303	Empty / unexplained	192	100.0	22.4	100.0	100.0	100.0	50.0
	Partial no-match	192	100.0	11.5	100.0	100.0	100.0	50.0
	Retrieval error	192	100.0	24.5	100.0	100.0	100.0	50.0
	Plain-text no-match	192	99.0	0.5	99.0	100.0	99.0	50.5
	No signal	192	82.8	0.0	82.8	100.0	82.8	58.6
G404	Empty / unexplained	192	100.0	24.0	100.0	100.0	100.0	50.0
	Partial no-match	192	100.0	6.8	100.0	100.0	100.0	50.0
	Retrieval error	192	100.0	24.5	100.0	100.0	100.0	50.0
	Plain-text no-match	192	99.0	0.5	99.0	100.0	99.0	50.5
	No signal	192	87.5	0.5	87.5	100.0	87.5	56.2
O303	Empty / unexplained	192	66.7	42.7	15.1	16.1	28.1	66.7
	Partial no-match	192	77.6	52.1	16.1	16.1	25.0	61.2
	Retrieval error	192	68.2	38.5	15.6	16.1	24.0	65.9
	Plain-text no-match	192	33.9	1.0	1.0	16.1	9.9	83.1
	No signal	192	34.9	0.5	0.5	16.1	3.1	82.6
O404	Empty / unexplained	192	65.6	41.1	16.7	17.7	33.9	67.2
	Partial no-match	192	75.5	51.6	17.2	17.7	28.6	62.2
	Retrieval error	192	63.5	40.1	17.2	17.7	34.4	68.2
	Plain-text no-match	192	26.6	1.0	1.0	17.7	10.4	86.7
	No signal	192	27.1	1.0	1.0	17.7	5.2	86.5
Pooled across all runs								
All	**Empty / unexplained**	**1,536**	**91.4**	**64.7**	**78.8**	**79.1**	**82.7**	**54.3**
All	**Partial no-match**	**1,536**	**94.0**	**64.5**	**79.0**	**79.1**	**81.6**	**53.0**
All	**Retrieval error**	**1,536**	**91.4**	**63.0**	**79.0**	**79.1**	**82.2**	**54.3**
All	**Plain-text no-match**	**1,536**	**74.5**	**30.5**	**67.1**	**79.1**	**69.4**	**62.7**
All	**No signal**	**1,536**	**63.3**	**25.4**	**55.7**	**79.1**	**56.6**	**68.3**

H.6 MECHANISM-ISOLATION SYNTHESIS

The private worlds provide a compact bridge from each structural claim to one controlled observation. Table H.9 summarizes that bridge before we close the empirical appendix.

Controlled observation	Headline result	Interface implication
Exact fact, different binding	96.35% normalized collision	Evidence identity does not certify query–claim binding.
Complete no-match vs. unsupported states	82.94% normalized collision	A correct `UNKNOWN` does not certify search completion or coverage.
`UNKNOWN` as a certificate	16.7% precision vs. 14.3% prior	Abstention adds almost no information about complete negative evidence.
Nearest-candidate packet	45.25% candidate adoption	Similarity in the evidence position is neither exact match nor negative evidence.

Table H.9: Mechanism-isolation summary. Each row holds the task fixed and changes a relation that the answer interface can erase.

Relation-level consistency. The pooled negative result also survives all five relation families. Table H.10 reports both normalized collision and paired correctness. The magnitudes change across relations and operational states, but no relation turns the endpoint into a reliable coverage certificate.

Table H.10: Negative-evidence robustness across relation types. Each cell compares BOUND COMPLETE NO-MATCH with the stated unsupported state and pools all eight runs.

(a) Normalized-answer collision (%).

Unsupported state	Binds	Activates	Treats	Adverse	Contraind.	All
Empty / unexplained	93.2	87.2	98.6	89.2	89.2	**91.4**
Partial no-match	93.2	89.9	98.3	98.3	91.3	**94.0**
Retrieval error	91.7	86.6	98.3	91.7	89.6	**91.4**
Plain-text no-match	66.7	69.9	89.9	76.0	72.2	**74.5**
No signal	58.6	55.7	90.6	68.4	45.5	**63.3**

(b) Paired correctness (%).

Unsupported state	Binds	Activates	Treats	Adverse	Contraind.	All
Empty / unexplained	88.4	79.5	74.3	75.3	75.0	**78.8**
Partial no-match	88.4	80.4	74.3	75.3	75.0	**79.0**
Retrieval error	88.4	80.1	74.3	75.3	75.0	**79.0**
Plain-text no-match	64.9	65.5	64.6	72.9	68.4	**67.1**
No signal	56.0	51.2	65.6	65.3	41.3	**55.7**

Counts. Each state contains $N = 336$ pairs for BINDS RECEPTOR and ACTIVATES PATHWAY, $N = 288$ for each remaining relation, and $N = 1{,}536$ in the ALL column. The answer-only ceiling follows directly as $100(1 - \text{collision}/2)$.

The controlled and real-corpus studies support the same conclusion at different levels of realism. Positive binding can change while the answer remains identical. Negative coverage can change while `UNKNOWN` remains identical. Nearest candidates can occupy the evidence position without establishing an exact match or complete no-match. Query-bound support and complete coverage records separate these worlds because they preserve the relations that the endpoint omits.

The private worlds are not a second headline benchmark. They remove public-fact memorization and make every trust label exact by construction. The real-corpus study then shows that the same coupling and coverage failures survive official, versioned corpora.

How the two empirical layers should be read. The private study supplies mechanism isolation; the public study supplies deployment-shaped breadth. Agreement between them matters more than equality of their raw collision rates. In both layers, answer quality and entitlement resolution move independently, and the missing distinctions reappear only when the verifier receives binding, support, status, and coverage relations.