CONSTRUCTING A PROVENANCE-ATTRIBUTION BENCHMARK OVER THE SCIENTIFIC LITERATURE: A SUPPLY CENSUS AND A SHORTCUT GATE

Anonymous authors
Paper under double-blind review

ABSTRACT

We set out to test whether a reported operating-regime boundary in provenance-constrained generation (attribution improves when a claim could plausibly come from several sources and collapses when it could not) transfers to the scientific literature. Building the benchmark produced a different result: across three successive constructions, a rule consuming no evidence could answer it. Gold was identifiable from question surface features in 335 of 335 items; the headline metric scored 0.873 in the treatment family against 0.250 in the control at zero capability, because its denominator was the manipulation; and publication year, reachable only through the provenance map under test, identified gold in 131 of 135 items, only in the arm the map is supplied to. A suite of 639 checks passed throughout. We contribute a shortcut gate that decides which such artifacts are fatal by conditioning each candidate rule on the information it consumes: question-only rules must not beat chance, evidence-reading rules may, with their maximum published as a ceiling, and provenance-reading rules must not beat the null of their own candidate set. Applied to our construction, the gate eliminated all 27 question-level rule–scope failures and left the evidence-level ceiling at 0.994 and 0.996, which we read as a diagnosis of this design rather than an impossibility. We also report a supply census and a decision pilot of 888 calls ($10.15) that failed its prespecified validity criteria; each triggered criterion was itself misspecified, so we did not proceed to the full experimental grid.

1 INTRODUCTION

A benchmark is an instrument, and an instrument can be miscalibrated in a way that survives every test of the code that produced it. This paper reports the construction of a provenance-attribution benchmark over the biomedical literature, and the instrument-level result we did not set out to find: across three successive versions of that construction, **a rule consuming no evidence at all could answer the benchmark**, and each artifact admits an explanation of the registered comparison other than capability.

The intended study was a registered replication of an industrial result: provenance-constrained assembly improves citation attribution when a claim could plausibly come from several sources, and the advantage collapses when it could not. Testing whether that boundary transfers to the scientific literature, where sources are published studies and claims are quantities, needs a benchmark with two families differing in one variable, and building it is where the paper's contribution turned out to lie.

What went wrong, three times. First, the gold passage was identifiable from the question's surface alone: "placeholder present, answer P_2; otherwise P_1" was correct on 335 of 335 items, a $9.0\times$ lift over chance, and equally so in both families. Second, the headline metric scored 0.873 in the treatment family and 0.250 in the control *at zero capability*, because its denominator was the manipulation, which compresses the treatment family's headroom and biases the registered contrast, map supplied against map withheld, negative. Third, publication year, rendered on the document card and reachable only through the provenance map the experiment was designed to test, identified gold in 131 of 135 items with zero errors, and only where the map was supplied, which is the one

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.

artifact running with the registered direction. None was caught by tests: a suite of 639 checks passed throughout, and one had codified a defective sentence boundary, actively protecting a defect.

Contribution. That construction can install artifacts is established, and two 2026 papers make the specific claim for retrieval and for hallucination benchmarks (Liu et al., 2026b; Hussain & Kantarcioglu, 2026). We contribute:

1. **A shortcut gate** (Section 4) that decides which shortcuts are fatal and which are legitimate by conditioning each candidate rule on the *information it consumes* rather than on where it looks. Rules reading only the question must not beat chance; rules reading the evidence may, and their maximum is published, because a ceiling of 1 means no improvement from the provenance map can be read as information the map supplied; rules reading provenance metadata must not beat the null of the set they are given, conditioned on what the evidence already licenses. The catalogue is a ratchet of 63 rules scored over 189 rule–scope pairs.
2. **A supply census** (Section 5) of what our rules retained from our corpus slice: of the quantity-bearing sentences extracted from systematic reviews, 82.2% carry no stored citation inside the sentence and only 6.3% carry one that resolves to a document we hold; the negative-class rules of the build we measured retain 36 of 1,390 documents carrying a recomputable numeric claim; and screening 378 full-text erratum pairs produced 27 candidates, whose audit identified 13 numeric disagreements, ten of them live supersessions; a hand-read sample of 41 excluded pairs found 2 genuine corrections the screen missed, so the 27 are not an enumeration.
3. **A negative result and its diagnosis.** Our benchmark's text-only ceiling is 0.994 in the treatment family and 0.996 in the control, measured with the rule given no access to the gold value: the gold passage is identifiable from the evidence text without the provenance map. We diagnose the design rather than declare the ceiling unavoidable. Our construction establishes origination by *numerical uniqueness* and builds its map from the same citation resolution that produces the labels, so the map is either redundant, the evidence text already settling origination, or not independently established from the labels; and the rule that attains the ceiling also reads the masked value off its flanking context in a rendering that includes the quoted passage. Removing any of these is a design change outside the construction measured here, and one we cannot price.

On the experiment. The decision pilot comprised 888 calls at a cost of $10.15 and failed the prespecified validity criteria (Section 6). Inspection showed that all three triggered criteria were themselves misspecified, two through defect classes already repaired in the construction: a registered band applied to the wrong family and metric, a scoring path rejecting 422 answers that contain the gold value verbatim, and a verifier criterion whose substantive triggers cannot fire on this item set. We therefore did not proceed to the full experimental grid. The instrument-level results above are properties of the construction, measured deterministically at no model cost, and do not depend on any of this.

2 Related Work

What is already established, and what we therefore do not claim. That benchmarks can measure something other than what they name is documented across 445 of them (Bean et al., 2025), with audit checklists and genuine nulls already in the genre: Zhu et al. (2025) report distortions "up to 100% in relative terms" and an empty-response agent succeeding 38% of the time. Shao (2026) draw the three-way separation we adopt, between dataset artifacts, model shortcuts, and protocols that reward metadata recovery rather than evidence use, and Luo et al. (2026) argue design validity "can be checked structurally before model inference."

Two 2026 papers reach the specific claim we would otherwise have made. Liu et al. (2026b) locate a legal-retrieval benchmark's measured effect in a stated construction rule and quantify it with a model-free oracle closing 99.2% of the BM25-to-best-trained gap; Hussain & Kantarcioglu (2026) find four of six hallucination corpora embed the ground-truth answer in the prompt and conclude that reported progress is "substantially explained by benchmark construction artifacts." **We therefore do not claim the observation**, and we do not claim its instruments. Dropping items a probe can

solve before release is established practice, from adversarial filtering (Le Bras et al., 2020) through the validated subsets of GPQA and Humanity's Last Exam (Rein et al., 2023; Phan et al., 2025) to the exclusion rules of LitQA2 and SeedRG (Skarlinski et al., 2024; Liu et al., 2026a). The closest procedure, Ovcharov (2026), combines a question-withheld probe, a binomial null over option permutations and an item-rejection gate; its central finding, that a bank filtered against one model stays open to a stronger one, argues for probing with enumerated, model-free rules rather than with a model, as Section 4 does. What we add is the combination Appendix F tabulates: blocking at build rather than item granularity, information tiers that separate a fatal shortcut from a legitimate one and publish the latter's ceiling, per-rule nulls, and one pass over several shortcut classes. The nearest formal treatment, Li et al. (2026), gives five defect classes "silently manufactured by implementation details" and then fences its own scope, "a validity audit is a second-order measurement," leaving build-time enforcement to others. That sentence is our demarcation.

Supply as a finding, and a gap in the attribution literature. Corpus-scale recomputation of published statistics is well developed: Nuijten et al. (2016) recompute 258,105 p-values over 30,717 articles, of which 16,695 (54.4%) yield at least one recomputable test; Brown & Heathers (2017) find only 71 of 260 sampled articles testable at all, the "supply is the finding" precedent we follow; Carlisle (2017) analyse 72,261 means across 5,087 trials. None concerns table-derived arithmetic claims that must survive a *same-table negative class*, the prevailing practice being to *insert* errors rather than find naturally occurring wrong-but-plausible values (Xi et al., 2025). One lexical observation: a recent attribution survey (Schreieder et al., 2026) contains none of the terms *shortcut*, *artifact*, *spurious*, *random baseline*, *chance level* or *leakage*, a fact about one survey's vocabulary that we do not read as evidence that attribution benchmarks go uncalibrated.

3 Benchmark Construction

3.1 Task and families

An item presents a masked quantitative claim and a closed set of nine passages, and asks the model to supply the value and cite the passage that is its original source. In the **treatment** family (`multi_source`) the question quotes a sentence from a systematic review that states a quantity and attributes it to one referenced study, eight of the nine passages come from studies the review cites, and the gold passage is the one from the study that produced the value. In the **control** (`single_source`) the question is a masked sentence from a single document and only that document's passage is admissible. The manipulation is the number of admissible candidate sources, eight against one; everything else, passage count, per-passage length bound and rendering, is held equal by construction, for reasons Section 4 makes concrete.

3.2 Evidence standard for the gold citation

A review sentence anchors an item only if all of the following hold. The standard is deliberately strict, and the supply it costs is reported in Section 5.

1. The sentence states a *distinctive* quantity, typed or carrying at least two printed decimals; bare integers collide constantly and matching on them is matching on noise.
2. It contains **exactly one attribution context**, either *structural*, a JATS `bibr` cross-reference whose `rid` resolves through the review's own reference list to exactly one corpus document, or *nominal*, a case-sensitive surname followed by a citation cue. Anything else is a discard, never an adjudication.
3. The named study states the same quantity, dimensionally normalised, in a sentence sharing at least two content words with the review's.
4. **No other resolved reference states that quantity**, since an item whose value appears in two referenced studies has no single correct answer.
5. **Origination**, an operational criterion: the gold sentence carries no citation of its own. A document attributing a value elsewhere is not its source; absence of a stored citation does not show it is, and 10 of the 200 shipped control golds are Introduction sentences.

defect	measurement that exposed it	disposition
Attribution leak	gold surname in 49/49 treatment questions, 0/200 control	blinding, plus a build assertion
Differential truncation	50.9% of treatment distractors dropped to fit the context budget against 20.8% of control	equalised passage count and length bound
Metric denominator is the treatment	random-citation null 0.873 treatment against 0.250 control	headline moved to the gold-exact metric
Gold the document may not have originated	40/200 control golds drawn from Introduction sections, as counted at the time; 10/200 shipped control golds are Introduction sentences, and all pass condition 5	condition 5
Corpus text corruption	1,978 citation markers fused onto adjacent digits, corrupting ≈214 numeric literals	offsets recorded at the removal point; abbreviation-aware segmentation

Table 1: Defects found and repaired before any model call. Each was found by asking what the result would look like if the mechanism under test were absent.

Conditions 2 and 5 use the same evidence in opposite directions: a citation marker in the *review's* sentence establishes attribution, one in the *primary study's* destroys origination. Resolving by `rid`, the publisher's machine-readable citation link, rather than by a printed surname is a strengthening, and it raised the treatment family from 49 to 148 items without widening any standard.

3.3 Blinding

Attribution evidence verifies the gold; it must never be usable to *find* it. Every resolved-reference surname is replaced by `[AUTHOR]` and every citation marker group by `[CITATION]`, in question and passage text, in every family, with a build-time assertion that no case document's label appears in any question. Document cards render opaque per-case aliases, and publication year is not rendered at all. The last two are not hygiene: before them, the *older* document's passage was gold in 131 of 135 items and never wrong where it fired, since a review cites older work, and that rule required the provenance map, which only one condition provides. The shortcut mimicked the hypothesis under test.

3.4 Repairs made before any experiment

The construction above is the third: two rounds of defects were found and fixed before any model call, by the gate of Section 4 and an independent code review, and the pattern, not the list, is the contribution.

Two properties matter more than the contents. **Every entry was found by a gate and not by inspection**, and none by the test suite of Section 1. And **the repairs introduced new defects of the same class**, the sixth and seventh being found in the construction built to repair the first five, which is why the contribution is a standing gate rather than a corrected benchmark.

4 A Shortcut Gate for Benchmark Construction

4.1 The rule

A benchmark measures what its construction permits, and construction rules can install the very contrast an experiment intends to measure. This is distinct from models exploiting annotation artifacts in fixed data: the question here is what a *rule* puts into the data before any model sees it. Two 2026 papers have already crossed that distinction (Section 2), so **we do not claim that auditing construction is new**, and none of the gate's ingredients is new either: restricted-input probes, null calibration and dropping probe-solvable items before release are each published (Appendix F). What we claim is their combination. Blocking is at the granularity of the build, not the item. Legitimate and fatal are separated by information tier, under which T1 is *permitted* to beat its null and publishes a

Tier	Information	Disposition
T0	question text only	must not exceed $1/k + 3\,\mathrm{SE}$; exceeding it is a **hard build failure**
T1	question + passage text	**permitted** to beat its null, this is legitimate evidence reading. Its maximum, t_1^{ceiling}, is published, and Section 4.2 states what it licenses
T2	+ provenance map and document metadata	must not exceed the null of the candidate set *its own inputs license*, which for a rule that filters on the evidence first is the null of the filtered set; **hard build failure**

Table 2: The three tiers. T1's permissiveness carries the substantive content: a benchmark solvable from its evidence text is working as intended, so the ceiling is published rather than the rule suppressed.

ceiling; Appendix E.1 measures what that distinction changes in what ships. The probe is a versioned, add-only catalogue of enumerated, model-free rules. Nulls are per rule, computed from each rule's own candidate set per item, with abstention earning the null. Several shortcut classes are audited in one pass.

The natural first formulation, "no rule may predict the answer above chance", is too narrow, since three of the four defects we report are invisible in question surface features, and unsatisfiable, since *some* rule must beat chance or the benchmark carries no signal: the construct is itself a rule. The operative form conditions on information rather than on surface:

> **No rule may exceed the null of the information it consumes.** Rules are enumerated in a frozen, versioned, add-only catalogue, partitioned by information tier, and each is scored against the null implied by the candidate set its own inputs license.

The catalogue is a **ratchet**: a rule discovered once is added permanently and never removed, so the gate strengthens monotonically and cannot be weakened by forgetting. It holds 63 rules, 18 at T0, 13 at T1 and 32 at T2, scored over 189 rule–scope pairs. Tier discipline is enforced by type rather than by convention, a T0 rule receiving a view object that cannot reach passage text at all. The same mechanism makes the tiers separable: a T2 rule reproducing a T1 rule earns exactly its own null and cannot fail. Appendix A gives the operational specification: the declared evidence-only reduction that makes that guarantee hold, the conditioning-validity contract every cell must satisfy, the three blocking dispositions, the pointwise failure threshold, and the pricing of abstention, of degenerate cells and of fitted rules. Ranking instead by lift over a common null would condemn the legitimate rule and rank the genuine build failure second, on our own cells (Appendix E).

4.2 What t_1^{ceiling} licenses, and what it does not

t_1^{ceiling} is the maximum accuracy *attained* by an enumerated rule reading question and passage text. Being attained, it is a lower bound on optimal text-only solvability and not an upper bound on anything a model does.

> A ceiling at 1 establishes that the evidence text is *sufficient* to identify the gold passage. It follows that no improvement a model shows when given the provenance map can be attributed to information the map supplies and the text lacks, because there is no such information. It does *not* follow that the improvement is small.

What the ceiling forbids is calling a with-map gain a measurement of what the map *knows*. Such a gain is an *elicitation* effect, weaker than the capability claim this benchmark was designed to test, and publishing the ceiling forces the choice between them into the open before the experiment runs (Appendix G.1).

	before repair	after repair
items scored (treatment + control)	$135 + 200$	$148 + 200$
failing rule–scope pairs	55	4
T0 (question text alone)	27	**0**
T2 (map and metadata)	28	4
distinct rules behind the failures	27	**2**
t_1^{ceiling} executable, treatment / control		**0.994 / 0.996**

Table 3: The gate applied to our own construction before and after the repairs of Section 3. Eliminating every T0 shortcut is a real repair; the near-complete ceiling is the finding.

4.3 What the gate found

Before repair, a rule consuming *only* the question text, if the question contains a citation placeholder answer P_2, otherwise answer P_1, identified the gold passage for 335 of 335 items, a $9.0\times$ lift over the item-level random-citation null of 0.111. It is the composition of three individually correct construction conditions, and it survives the gold-position permutation the construction applies (Appendix A.1).

After repair, T0 is empty. Four rule–scope pairs still fail, from **two** distinct rules. "The document contributing the fewest passages" fails on all three scopes, largest at $z = 19.3$ pooled, because the gold document contributes exactly one passage in $148/148$ treatment and $200/200$ control items while its neighbours average 1.11–1.21. "Retraction status" fails on the control scope alone at $z = 3.5$, *falsifying* the assumption under which retraction had stayed on the rendered document card.

4.4 The ceiling did not move, and we report why

Oracle and executable. The catalogue builds the value-stating set from the gold value, so a catalogued T1 accuracy is an *oracle* accuracy; re-deriving that set label-blind from question and passages and re-running every rule gives the *executable* ceiling, $\mathbf{0.994}$ in the treatment family against an oracle 1.000, and $\mathbf{0.996}$ in the control. We use the executable figures throughout, since quoting 1.000 would present an oracle number as a demonstrated text-only result. Appendix A gives the derivation, its agreement rates and the per-rule comparison.

What the ceiling diagnoses. Our construction requires that no other referenced study state the gold value, so that the gold citation is unambiguous. That requirement is what makes the gold *text-identifiable*, and text-identifiability is exactly the shortcut. Stated at the scope it actually has: for a construction that establishes origination by **numerical uniqueness**, and whose map is built from **the same citation resolution that produces the labels**, the two requirements are in direct tension. If origination is determinable from the text, the provenance map adds no information; if it is determinable only from the map, the labels and the map were not independently established. Both conditions hold of ours; neither is forced on a provenance benchmark.

This diagnoses our design; it does not bound every design meeting the two conditions. Uniqueness says that one referenced study states the quantity, not that an algorithm can find it, and ours finds it by recovering the masked value from the literal context flanking it in a rendering that supplies the quoted passage (Appendix A), both of them choices. Neither qualifier was in our first statement; Appendix G argues each and names the design the result does not reach. The usable zone lies between, origination inferable from text but requiring inference (Appendix G.1), and reaching it requires publishing the text-only ceiling as a headline row, a branch pre-registered before t_1^{ceiling} was measured.

5 A Supply Census of the Yield Under Our Rules

Benchmark papers report what they built and rarely the yield of their own admissibility rules, the count of items a construction retains from its corpus. We report it for three families over a corpus of 14,982 full-text open-access biomedical articles; the genre is established (Section 2), and to our knowledge no prior work reports the attribution supply or the erratum-instantiation rate below.

citation falling inside the quantity-bearing sentence	occurrences	share
present and resolved to a corpus document	9,190	6.3%
present but not resolvable to a corpus document	16,777	11.5%
none	**120,171**	**82.2%**
total	146,138	100%

Table 4: Citation attribution over the quantity-bearing sentence occurrences our extractor produces, measured from the structured citation record on the corpus build the benchmark uses. The dominant case is an extracted quantity sentence carrying no stored citation of its own.

5.1 Attribution supply: most extracted quantity sentences carry no citation

For a review sentence to anchor an attribution item, it must state a quantity and identify the referenced study it came from. We measure the citation cue from the corpus build's *structured citation record* rather than by pattern-matching rendered prose: each citation is stored with the offset at which its marker was removed, so a sentence is attributed exactly when a stored offset falls inside its own span (Table 4, over 146,138 quantity-bearing sentence occurrences in 1,686 systematic reviews).

Of the quantity-bearing sentence occurrences our extractor produces, 82.2% have no stored citation inside the sentence, and only 6.3% have one that resolves to a document we hold. The second figure is the operative one, because an item needs a citation that resolves before it can be checked at all. Both are measured over extracted occurrences and stored offsets, the population an item is drawn from, and both are claims about that population rather than about the articles as written.

Two scope qualifications stay attached: the count concerns the sentence carrying the quantity, since an item needs the cue *on the sentence* to be machine-checkable (Appendix B.3 lists the uncounted channels), and it counts citations, the nominal cue of an inline surname being a separate channel our earlier text-level pass put near 1%. Dropping the attribution requirement entirely would ceiling at 1,229 anchors; requiring it yields 148. The figure replaces an earlier 88.1% from a regular expression over rendered text on a superseded corpus build (Appendix B).

Three pipeline defects initially suppressed the yield further, and separating them from the corpus fact required measuring each; Appendix B records all three, among them a surname rule that excluded 937 of 940 unresolvable references, all two-character surnames and overwhelmingly Chinese, whose equity consequence is not incidental and whose naive repair is worse than the defect.

5.2 The other two families, and why we report any of this

The recomputable family loses 1,390 documents to 125 on the same-table negative class and then to **36** on our 50% composition cap, a figure produced by the admissibility rules of the build we measured rather than a limit the published record imposes. Screening the conflict family's 378 erratum pairs produced 27 candidates, whose audit found **13** numeric disagreements including ten live supersessions, and commentary pairs give 23 of 131, roughly 33 items against a target of 300. A hand-read sample of 41 excluded pairs held two corrections the screen missed, so the 27 are not an enumeration. In all three families our rules retained substantially fewer items than the intended scale, and reporting the per-rule yield makes a family this construction did not fill a finding rather than a silence. Appendix C gives both derivations, per derivation type and per notice class.

6 The Pilot, and What It Found

The pilot used 888 calls to `gpt-5.4-mini`, cost $10.15 against a prespecified $30 budget ceiling, and produced 24 evaluation cells at 148 items per family. It failed the prespecified validity criteria. The decision not to continue was warranted, but inspection showed that each triggered criterion was itself defective: two reproduced defect classes already identified in the benchmark construction, now in the scoring and validity checks, while the third could only have been satisfied by a model failure.

6.1 THE PASS-CONDITION EVALUATOR DOES NOT IMPLEMENT THE REGISTERED CONDITION

Condition (a) is registered as a non-degeneracy band on `attribution_correct`, applied to the treatment family, at $[0.15, 0.85]$. A pre-outcome amendment fixes its scope: the band "cannot hold in the CONTROL on any competent model, because the control informative null is 1.000," so applying it there fails the pilot for an unrelated reason.

The evaluator applies it to the control, and on `citation_correct`. Both substitutions push toward failure, since `attribution_correct`, the conjunction of a correct value and a correct citation, is the stronger of the two. All four offending cells are control cells, at 0.878 to 0.973, the amendment's predicted failure realised exactly.

Re-evaluated as registered, condition (a) **passes**. Treatment `attribution_correct` across all six cells is 0.527, 0.237, 0.493, 0.270, 0.554, 0.264, every value inside the band.

6.2 THE STRICT SCORING PATH REJECTS 422 ANSWERS THAT CARRY THE GOLD VALUE

Re-scoring the stored answers on both registered paths is deterministic and free. Of 888 scored answers, **422** (47.5%) are accepted by the companion path and rejected by the strict one, against 45 in the other direction, and in **422 of 422** (100.0%) the gold numeral is literally present in the answer. The rejections track surface rendering, a unit or separator differing between the gold value and the answer string. Numeral presence is not semantic correctness, so we audited all 422; Appendix D gives the worked examples, the negation-window and citation counts (Table 9), and the seeded hand-read of 30, which was consistent with rendering mismatch and exposed two mis-extracted gold values.

Strict is one of the two paths condition (a) is scored on, and it moves treatment `attribution_correct` in the `full` cell from 0.527 to 0.237. Our checks are pooled over both families, so they cannot apportion that 29-point movement to either arm, and the audit establishes numeral presence in the 422, not semantic equivalence item by item. What they support is enough: a scorer rejecting 47.5% of answers that carry the gold value is not a neutral instrument, so a contrast read off this column would be confounded with its normalisation and we report none.

6.3 A CONDITION ONLY A MODEL FAILURE COULD SATISFY

Condition (c) requires the citation verifier to flag at least one item; it flagged zero of 888. The verifier rejects an answer on four paths (Appendix I). The pilot has no attack items and every case a non-empty authentic set, so **three of the four cannot fire here**; the fourth, a citation to a passage id outside the case, can, and none of the 888 answers did. A clause carried entirely by a model failure mode that did not occur is not a health check on the verifier, and it was registered as gating.

The zero has a further consequence: the with-verifier arm, derived offline from the same answers at no extra call, is *vacuous*; with nothing to detect, 12 of the 24 cells duplicate their counterparts exactly.

6.4 IMPLICATIONS OF THE PILOT

Condition (b) registers an expected ordering, `full` $\geq$ `no_prov` $\geq$ `raw` on the treatment family, and records deviation without failing. It does not hold: `raw`, carrying no provenance scaffolding, is the best variant on every path, which the registered hypothesis did not predict. **We do not claim it**, because a result read off an instrument just shown to be broken is not evidence (Appendix H).

We therefore did not proceed to the full experimental grid. Condition (a) passes as registered and (e) passes with a $13.65 projection against a $30 ceiling; (b) and (c) are as above. Condition (d) is verifier–human raw agreement at a threshold of 0.85. The verifier predicted no flag on all 888 answers. Human verification labels were not collected, so the corresponding agreement statistic is unavailable rather than undefined. Its registration is the larger problem: against an all-negative detector raw agreement is one minus the human flag rate, so the threshold is met whenever fewer than 15% of answers deserve a flag, and a detector that never fires passes the condition meant to validate it, the null-model defect this paper reports catching in a benchmark, found in our own pre-registration. Running the full grid with this instrument would have produced a numerical result without supporting the intended measurement; the $10.15 pilot was sufficient to establish that limitation.

7 DISCUSSION

7.1 WHAT A NEAR-COMPLETE CEILING MEANS

Our benchmark's text-only ceiling is all but complete, a mean per-item score of 0.994 on treatment and 0.996 on control items for a rule with access to neither the provenance map nor the gold value. That invites engineering the ceiling away; Section 4 gives the design choices it would have to undo, and a capability reading of any effect measured here must survive a ceiling this high (Appendix G.1).

7.2 LIMITATIONS

The regime-transfer question is still open. The pilot established that the instrument, as gated, cannot answer it, not what the answer is (Section 6); the one row pointing against the registered hypothesis is not a finding, because an instrument that has failed its own gate cannot support one.

Two residual shortcuts, over four scopes. "Fewest passages" and "retraction status" fail the gate on the shipped build, over four rule–scope pairs (Section 4). Both repairs are filters over the shipped pool: one passage per document returns its rule to exactly the null on the 183 items already satisfying it, and seeded removal of gold-retracted items until gold and distractor retraction rates agree moves the control cell from $z = 3.479$ to 0.73 at a cost of 14 of 200 items. With both applied, all 18 T0 and all 32 T2 rules are re-scored and none crosses $z > 3$ in any scope: the repaired build passes its own gate at 179 of 348 items, 48.6% of supply removed (Appendix J). That is the shipped pool filtered under the repaired rules; a construction drawing under them from the start is a different build with its own supply and catalogue run, and the subset does not clear the shipped benchmark.

Single corpus, single domain, and one detector. All measurements are over open-access biomedical articles, and the 82.2% should not be assumed to transfer; it counts citations stored inside the sentence stating the quantity (Appendix B.3), not how reviews are written.

The catalogue is not complete and cannot be. The gate tests only rules someone enumerated, so a shortcut nobody imagined passes it. A green gate means only that no enumerated rule triggered a blocking disposition under the stated threshold policy, not that the tested associations are absent; several of ours sit above their null without crossing, so we publish the catalogue, not only the verdict.

Recommendation. Two practices, both cheap and deterministic. Report the null of every derived rate beside the rate, which for a benchmark means the accuracy of a rule consuming no evidence. And publish the text-only ceiling, executed without label access, because it fixes what a later effect may be called.

Reproducibility Statement

Code and instructions are provided in the supplementary material. The construction of Section 3, the shortcut gate of Section 4 and the supply census of Section 5 are deterministic and make no model call: given the corpus build, the released scripts deterministically regenerate every item, rule score, and census count, with associated build and seed metadata recorded in JSON reports. The gate is specified in Appendix A (catalogue version 1.1.0, Monte-Carlo seed 20,260,818, $B = 2{,}000$), the census and its pipeline defects in Appendices B and C, and the residual-shortcut repairs in Appendix J. The corpus is assembled from public bibliographic and open-access full-text sources by the released fetch scripts. The pilot of Section 6 is the only step that calls a model (888 calls to `gpt-5.4-mini`); re-scoring its stored answers on both registered paths is deterministic and free (Appendix D).

References

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, Maria Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Kaili-May Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrom, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Inioluwa Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*, 2025. doi: 10.48550/arXiv.2511.04703. URL https://arxiv.org/abs/2511.04703.

Nicholas J. L. Brown and James A. J. Heathers. The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4):363–369, 2017. doi: 10.1177/1948550616673876.

J. B. Carlisle. Data fabrication and other reasons for non-random sampling in 5087 randomised, controlled trials in anaesthetic and general medical journals. *Anaesthesia*, 72(8):944–952, 2017. doi: 10.1111/anae.13938.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.408.

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. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.398.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 107–112. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2017.

Khizar Hussain and Murat Kantarcioglu. PARALLAX: Separating genuine hallucination detection from benchmark construction artifacts. *arXiv preprint arXiv:2605.17028*, 2026. doi: 10.48550/arXiv.2605.17028. URL https://arxiv.org/abs/2605.17028.

Divyansh Kaushik and Zachary C. Lipton. How much reading does reading comprehension require? a critical investigation of popular benchmarks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 5010–5015. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1546.

Jakub L'ala, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. PaperQA: Retrieval-Augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023. doi: 10.48550/arXiv.2312.07559. URL https://arxiv.org/abs/2312.07559.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi. Adversarial filters of dataset biases. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1078–1088. PMLR, 2020. doi: 10.48550/arXiv.2002.04108. URL https://proceedings.mlr.press/v119/bras20a.html.

Yanhang Li, Zhichao Fan, and Zexin Zhuang. Auditing the audit: Five failure modes in benchmark-validity audits. *arXiv preprint arXiv:2607.02586*, 2026. doi: 10.48550/arXiv.2607.02586. URL https://arxiv.org/abs/2607.02586.

Jiayi Liu, Jiaxing Zhang, Bowen Jin, and Jennifer Neville. Generating leakage-free benchmarks for robust RAG evaluation. *arXiv preprint arXiv:2605.08838*, 2026a. doi: 10.48550/arXiv.2605.08838. URL https://arxiv.org/abs/2605.08838.

Yao Liu, Tien-Ping Tan, and Zhilan Liu. Charge as a construct-validity factor in Chinese legal case retrieval: A cross-benchmark audit. *arXiv preprint arXiv:2606.12993*, 2026b. doi: 10.48550/arXiv.2606.12993. URL https://arxiv.org/abs/2606.12993.

Junhao Luo, Ning Huang, Ziqi Sha, Wenxuan Tang, and Wei Deng. Beyond local accuracy: A protocol-level identifiability audit for controlled LLM reasoning evaluation. *arXiv preprint arXiv:2608.13326*, 2026. doi: 10.48550/arXiv.2608.13326. URL https://arxiv.org/abs/2608.13326.

Michèle B. Nuijten, Chris H. J. Hartgerink, Marcel A. L. M. van Assen, Sacha Epskamp, and Jelte M. Wicherts. The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48(4):1205–1226, 2016. doi: 10.3758/s13428-015-0664-2.

Volodymyr Ovcharov. Gated against one model, open to the next: Option-only solvability in legal multiple-choice benchmarks. *arXiv preprint arXiv:2608.15428*, 2026. doi: 10.48550/arXiv.2608.15428. URL https://arxiv.org/abs/2608.15428.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025. doi: 10.48550/arXiv.2501.14249. URL https://arxiv.org/abs/2501.14249. Published in revised form as "A benchmark of expert-level academic questions to assess AI capabilities", Nature 649:1139–1146 (2026), doi:10.1038/s41586-025-09962-4.

Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191. Association for Computational Linguistics, 2018. doi: 10.18653/v1/S18-2023.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. *arXiv preprint arXiv:2311.12022*, 2023. doi: 10.48550/arXiv.2311.12022. URL https://arxiv.org/abs/2311.12022.

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 (Volume 1: Long Papers)*, pp. 30956–31000. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1430.

Kan Shao. Metadata predictability is not evidence dependence: An intervention-based audit for weak-label benchmarks. *arXiv preprint arXiv:2605.23701*, 2026. doi: 10.48550/arXiv.2605.23701. URL https://arxiv.org/abs/2605.23701.

Zachary S. Siegel, Sayash Kapoor, Nitya Nadgir, Benedikt Stroebl, and Arvind Narayanan. CORE-Bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *arXiv preprint arXiv:2409.11363*, 2024. doi: 10.48550/arXiv.2409.11363. URL https://arxiv.org/abs/2409.11363.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024. doi: 10.48550/arXiv.2409.13740. URL https://arxiv.org/abs/2409.13740.

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

David Wadden, Kyle Lo, Lucy Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 61–76. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-naacl.6.

Sarina Xi, Vishisht Rao, Justin Payan, and Nihar B. Shah. FLAWS: A benchmark for error identification and localization in scientific papers. *arXiv preprint arXiv:2511.21843*, 2025. doi: 10.48550/arXiv.2511.21843. URL https://arxiv.org/abs/2511.21843.

Yuxuan Zhu, Tengjun Jin, Yada Pruksachatkun, Andy Zhang, Shu Liu, Sasha Cui, Sayash Kapoor, Shayne Longpre, Kevin Meng, Rebecca Weiss, Fazl Barez, Rahul Gupta, Jwala Dhamala, Jacob Merizian, Mario Giulianelli, Harry Coppock, Cozmin Ududec, Jasjeet Sekhon, Jacob Steinhardt, Antony Kellermann, Sarah Schwettmann, Matei Zaharia, Ion Stoica, Percy Liang, and Daniel Kang. Establishing best practices for building rigorous agentic benchmarks. *arXiv preprint arXiv:2507.02825*, 2025. doi: 10.48550/arXiv.2507.02825. URL https://arxiv.org/abs/2507.02825.

AI USE STATEMENT

Generative AI tools were used to assist with retrieving and discovering related literature, code implementation, writing and polishing. The research ideas originated with the authors. The authors reviewed and verified AI-assisted content and take full responsibility for the paper's accuracy, originality, and conclusions.

Provenance of the numbers below. Every figure in this appendix is emitted by a script released with the paper, which writes a JSON report naming the quantity it computes, its seed where one applies, and the build it was computed on.

A THE GATE'S OPERATIONAL SPECIFICATION

A.1 THE NULL FOLLOWS A RULE'S OWN INPUTS

A cumulative tier definition invites an obvious objection: if T2 sees everything T1 sees, a T2 rule can ignore the metadata, re-execute the winning T1 rule, and be condemned for evidence reading that T1 expressly permits. The null is therefore set by the candidate set a rule's own inputs license and not by its tier. A metadata rule ranging over all passages is scored against $1/k$; a conditioned one is scored against the null of **its own declared evidence-only reduction**. Each of the 16 metadata rules is catalogued twice, unconditioned and conditioned, and the two ask different questions: whether the metadata predicts gold at all, and whether it adds anything the evidence does not already supply.

Why the guarantee holds rather than happening to hold. An earlier version of this gate conditioned every rule on one fixed set, the value-stating set, and that is not sufficient. The winning treatment rule does more than filter to that set: it also excludes the passage carrying the quoted sentence, taking two candidates to one. A T2 rule replicating that full evidence-only selection would have been scored against the two-member null of 0.500, would have looked like a perfect metadata shortcut, and could have been failed for evidence reading that T1 permits. Since the catalogue is an add-only ratchet that will accept rules nobody has written yet, a defect that is latent today fires on the first rule that declares a deeper reduction.

Each rule now declares the evidence-only reduction its conditioning is scored against, and **the signature is the enforcement**. Two restrictions are needed and the gate applies both, because either alone leaves a hole.

The reduction takes the question view and the passage view and nothing else, so it cannot reach the provenance map or the document cards: a permissive rule cannot absorb metadata-assisted selection into its own null, because such a reduction would not have the inputs. That much stops metadata flowing *in*. It does not stop evidence-only selection leaking *out* of the declaration, since a rule could accept the declared pool and then filter it further on the evidence, being scored against a null that does not account for the filter. So the second restriction closes the other direction: a T2 rule's selection function receives the pool, the question view and the *metadata* view, and not the passage view. Its only channel to the evidence is the reduction it declared.

With both, a rule reproducing a T1 rule earns exactly its own null and cannot fail, by construction rather than by inspection of the current catalogue. This is the mechanism that already enforces tier discipline, applied twice to conditioning.

Neither restriction moved a number. Every catalogued rule uses the default reduction, the value-stating set, whose null is 0.500 in the treatment family where the set has size two in $148/148$ items and 1.000 in the control where it has size one in $200/200$; and no catalogued metadata rule read the passage view, so withdrawing it removed an unused argument. We re-ran the whole gate after each change and diffed rather than asserting this: across 63 rules and every scope, accuracy, null, standard error, z, item count, abstention count, mean answer-set size, mean conditioning-set size and the degeneracy flag are unchanged.

The catalogue scores the invariant, not the spelling. The pre-repair T0 rule of Section 4, "placeholder present, answer P_2; otherwise P_1", survives the gold-position permutation the construction applies. A permutation that also renumbered the passage identifiers would defeat this particular rule without removing the shortcut, which is why the catalogue scores the invariant the rule exploits and not the spelling it happens to use.

The failure threshold is pointwise. Any single rule–scope pair at T0 or T2 with $z > 3$ fails the build, with no multiplicity correction across the catalogue. This is deliberately conservative in the direction of failing: with 189 rule–scope pairs a catalogue-wide correction would be the more usual choice, and would also make it cheaper to add rules until the correction absorbed a real shortcut.

What the conditioned cells currently establish. In the post-repair build **9 of the 16** conditioned variants abstain on every item in every scope, and which 9 is not arbitrary: they are exactly the variants consuming publication year, section, or the year embedded in a document label, which the repair removed by not rendering the year and by aliasing the cards. Abstention is scored as returning the whole conditioning set, which earns exactly the null by construction, so those cells are uninformative here rather than evidence that metadata adds nothing. The remaining **7 still execute**, and they are the variants whose only inputs are the passage-to-document map, the document identifier, the card order or the retraction flag. What the measurement shows about those channels is that they remain *available*, not that each is necessary. The map and the document identifier cannot be removed without removing the task, since the task is to attribute a passage to a source. The card order and the retraction flag can be: neither is required by the task of Section 3, and randomising the order or withholding the flag is a repair outside the build measured here, and one this section does not price. None of the seven crosses the failure threshold. Four sit exactly at their null with undefined z, and three depart from it pooled: the mirror pair "not retracted" at 0.7989 against 0.7874, $z = 1.886$, and "retracted" at 0.7759, $z = -1.886$, and "the first document card" at 0.7716, $z = -1.808$. A previous version of this paragraph said all 16 abstained, which was wrong. The control's conditioned null of 1.000 is likewise degenerate, with zero variance and undefined z, and we report it as a property of the family rather than as a passing grade.

A.2 Scoring

A rule returns a set of passage ids S, possibly tied, or abstains. An item's score is $1/|S|$ when the gold is in S and 0 otherwise, the accuracy a uniform choice from the rule's own output would achieve in expectation; a scope's figure is the mean of these fractional scores, not a rate of solved items, and the convention makes a tie-producing rule and a single-answer rule comparable on one axis. Where the validity contract above holds, abstention returns the whole conditioning set and so earns exactly the null, and can neither inflate nor deflate a rule.

The null is stated per item rather than assumed constant, because $|S|$ and the conditioning set C both vary. Under the null the gold is uniform on C, so for one item

$$\mathbb{E}[\text{score}] = \frac{|S \cap C|}{|C|\,|S|}, \qquad \mathbb{E}[\text{score}^2] = \frac{|S \cap C|}{|C|\,|S|^2},$$

and the item's null variance is the difference of the second moment and the square of the first. Items are treated as independent, so the two accumulate across the scope and $\mathrm{SE} = \sqrt{\sum_i \mathrm{Var}_i}\,/\,n$, with z formed against that. That assumption is measured rather than asserted: the 348 items spread over 344 distinct anchor reviews, a cluster count and not the item agreement count of Appendix A.3, and 333 gold documents, at most two sharing an anchor and at most four a document, so the worst-case design effect is 1.045, the value at an intra-cluster correlation of 1 and therefore an upper bound. Positive dependence within a cluster inflates the variance and so inflates a z computed as if items were independent: at that bound the standard error rises by the factor $\sqrt{1.045}$ and every reported z should be read as smaller by that factor. **What the bound covers is exactly those two units.** It does not bound dependence induced by shared distractor documents, by topical overlap between anchors, or by items built under a common extraction rule, none of which is measured here; a rule whose errors track one of those would still have fewer effective degrees of freedom than n. The cluster identity is the audit-only real document id, because the rendered `doc_id` is an anonymised per-case slot label reused in every case and would have merged unrelated items. When every item has $|C| = |S|$ the variance is zero, z is undefined, and the cell is reported as a degenerate null rather than as a pass.

Fitted rules, those of the form "best of n candidates", need a null that accounts for the selection. Theirs is a *simulated* max-null, not an analytic one: the gold is redrawn uniformly over each item's passages, the whole family is refit on the redrawn labels, and the best accuracy attained is recorded, over $B = 2{,}000$ draws seeded at 20,260,818. Write $a_1 \dots a_B$ for those values. The null is their mean and the reference spread is their sample standard deviation s, $z = (a_{\mathrm{obs}} - \bar{a})/s$, and a conservative $p = (\#\{b : a_b \geq a_{\mathrm{obs}}\} + 1)/(B + 1)$ is reported alongside. Two quantities are easy to conflate and we do not: s is the spread of the null *distribution*, which is what z requires; the Monte-Carlo error in locating its mean is $s/\sqrt{B}$, roughly a forty-fifth of s at this B, and it is a property of our estimate rather than of the null. Simulation prices selection over the hypothesis family *as it is actually performed*, where a Bonferroni correction bounds it without reference to the family's correlation

structure. The decision is taken on z, and p is reported rather than gating, so that the threshold is one rule for every tier.

The conditioning-validity contract. Every guarantee in this section is conditional on it: the scoring rule, the abstention property, the degeneracy disposition. For each scored item, writing C for the conditioning set and S for the rule's answer:

$$C \neq \varnothing, \quad \text{gold} \in C, \quad S \neq \varnothing, \quad S \subseteq C.$$

The first two are what makes the null a null: a stipulated uniform draw over a set that cannot contain the answer describes no experiment. The second two are what makes the score defined, since $1/|S|$ and $|S \cap C|/(|C|\,|S|)$ divide by both sizes. **Any violation makes the cell *invalid*, which blocks the build**, and the remedy is to fix the reduction and re-evaluate rather than to interpret the number. Each cell publishes `n_invalid_items` and `n_gold_outside_conditioning_set`; both are **0** in all 189 rule–scope pairs, so every comparison we report satisfies the contract.

Three dispositions. A cell that satisfies the contract either exceeds its null, fails to produce a comparison, or is fine.

Exceeds is the ordinary failure: $z > 3$ at T0 or T2, pointwise.

Unresolved covers a comparison the gate could not make, and it *blocks*. A fitted rule whose *simulated* spread is zero lands here: $s = 0$ means all B redraws attained the same best accuracy, which is also what a small B or a coarse accuracy grid would produce, so it does not establish that the family had no selection to price. Recording a pass there would let an unavailable comparison read as a satisfied one. The disposition is to raise B, and until someone does, the build does not go through. No fitted cell in our report is unresolved.

Analytic degeneracy does not block, and *given validity* that is not an exception. The item variance vanishes exactly when $|S \cap C| = |C|$, and with $S \subseteq C$ guaranteed by the contract the condition is $S = C$: the rule returned the whole conditioning set, which is what abstention does. A rule that made no selection has nothing to exceed. The same conditioning is why abstention earns exactly its null and can neither inflate nor deflate a rule. Neither statement survives a contract violation, which is why the violation blocks.

what the cell looks like	disposition	build
$S = C \ni \text{gold}$, variance 0	analytic degeneracy	**proceeds**
fitted rule, simulated $s = 0$	unresolved	**blocked**; raise B
$C = \{P_2\}$, gold is P_1	invalid	**blocked**; fix the reduction
$z > 3$ at T0 or T2	exceeds	**blocked**; fix the construction

Table 5: The gate's decision table. The first row is the only one that lets a build through without a comparison, and it does so because the rule demonstrably made no selection.

All of this is computed at catalogue version 1.1.0, Monte-Carlo seed $20{,}260{,}818$ and $B = 2{,}000$.

A.3 The executable ceiling, per rule

Table 6 compares each T1 rule's oracle accuracy, where the value-stating set is built from the gold value, against its executable accuracy, where the set is re-derived from the literal context flanking the mask in the passage that prints the quoted sentence. The masked value is recovered from the flanking context on 347 of 348 items, one item failing; among those 347, the re-derived value-stating set equals the oracle set on 344, so 344 of 348 agree overall, and the three recovered-but-mismatched items are all in the control family. (The 344 here counts agreeing items; the 344 of Appendix A.2 counts distinct anchor reviews.)

	treatment		control	
rule	oracle	executable	oracle	executable
value-stating, not containing the quote	1.000	**0.994**	0.116	0.115
the passage containing the quoted sentence	0.000	0.000	0.996	**0.996**
the unique value-stating passage	0.111	0.111	1.000	0.987
shortest among value-stating	0.878	0.872	1.000	0.986
longest among value-stating	0.122	0.122	1.000	0.991
fewest citation placeholders among value-stating	0.970	0.964	1.000	0.991

Table 6: Oracle against executable accuracy for the T1 rules that attain or approach the ceiling. The executable column is the one the structural argument rests on. The second row is the cleanest case: that rule never consults the value-stating set, so the two columns coincide by construction.

B SUPPLY CENSUS: PIPELINE DEFECTS AND A SUPERSEDED MEASUREMENT

The derivation that takes each supply family from corpus to items is given once, in Appendix C. This section records what the census turned up about the pipeline on the way: three defects that suppressed the attribution yield, and an earlier attribution figure that the current one replaces.

B.1 THREE PIPELINE DEFECTS, AND WHY THE NAIVE REPAIR IS WORSE

Three defects initially suppressed the attribution yield further, and separating them from the corpus fact required measuring each. A sentence splitter breaking on `et al.` detached the attribution from the quantity; a surname rule rejecting stems under three characters excluded 937 of 940 unresolvable references, all two-character surnames and overwhelmingly Chinese; and passage flattening concatenated bracketless citation markers onto the preceding character, losing markers for 697 of 1,686 anchors.

We report these because the equity consequence of the second is not incidental, the rule systematically excluding a subpopulation of first authors, and because the naive repair is worse than the defect: admitting two-character stems matched 63 short tokens appearing 382,043 times as ordinary words, including the English preposition *in* 173,696 times.

B.2 WHY THE ATTRIBUTION FIGURE REPLACES AN EARLIER ONE

We previously reported 88.1% from a regular expression over rendered passage text on an earlier corpus build, and both halves of that were wrong in ways worth stating, because the failure is the paper's own subject.

On the earlier build the flattening defect above concatenated bracketless markers onto the preceding character for 697 of the 1,686 anchors, covering 80,414 body cross-references, so the detector under-counted attribution and the "no cue" share was too high. That defect is one of the five this paper reports repairing. The rebuild that repaired it moved citation structure out of the rendered text and into the structured record, at which point the same detector found markers in 0.06% of sentences and would have reported 98.5%.

build	method	cited	surname only	none
`full_20260817`	regex over rendered text	10.9%	1.0%	88.1%
`full_20260818_citfix`	the same regex	0.06%	1.5%	98.5%
`full_20260818_citfix`	**structured record**	**17.8%**	not read	**82.2%**

Table 7: The same quantity measured three ways. Neither of the first two measures attribution: one is computed under a defect, the other points a text detector at a representation that no longer carries what it looks for. The regex rows carry a third class, an inline surname with no marker, which the structured record does not read, so each row sums to 100% up to rounding; the first row's denominator is 147,227 occurrences on the earlier build, the other two rows' 146,138.

The offset test is not a repair of the regular expression but a replacement for it. The second and third rows are both computed on `full_20260818_citfix`, the build the benchmark is drawn from; only the third reads the structured record, and only it measures attribution. Of the sentences it finds cited, 9,190 resolve to a corpus document and 16,777 do not, so most in-sentence citations in these reviews point outside the open-access corpus.

B.3 What the 82.2% does not count

The figure's scope is narrower than a claim about attribution in the literature. It counts citations falling inside the sentence that states the quantity, read from the corpus build's structured citation record, and it says nothing about attribution carried by an adjacent sentence, a paragraph-level citation, a table, or a quantity the review computed itself. That scope is the right one for a supply census, since an item needs the cue on the sentence, and it is the wrong one for a statement about how reviews are written, which the paper therefore does not make.

C The recomputable and conflict families

C.1 Recomputable-claim supply: what the negative class cost this build

A recomputable claim is one whose value a named set of table cells reproduces under a declared derivation. Extraction over the same corpus yields 13,598 viable candidates across 1,390 documents. But an item also needs a *negative*, a value present in the table yet wrong for the stated relationship, and that requirement, not extraction, is where the supply went in this build.

Three losses compound, and Table 8 separates them per derivation. Extraction yields 1,390 documents. The same-table negative class takes that to 125: a single admissibility gate accounts for most of the loss, since the rendered claim must contain exactly one table-derived numeral, and row and column labels routinely carry a second, most often a group size printed in the column header, which for a proportion is the claim's own denominator and therefore permits refuting the claim from its own text. A 50% single-derivation composition cap, which is our design choice and not a property of the record, then takes the family to **36 documents**.

The 36 is the output of one named construction, the `full_20260817` numeric build under that admissibility rule and that cap, and not a supply limit imposed by how tables are published. The rule's dominant rejection is a sample-size numeral printed in a column header, 11,823 of its record-level rejections, and a header numeral is a property of the rendering convention we adopted rather than of the claim. The report carries the counterfactual, measured and deliberately not applied: ignoring a column-header numeral would return 1,124 proportion documents and 25 difference documents to the pool at that stage, which is a statement about one gate and not about the family, since the later gates and the composition cap are then applied to them. The census below is therefore reported against its build.

derivation	viable documents	with an admissible pair	retained
proportion	1,319	107	8.1%
difference	64	13	20.3%
mean	26	0	0%
ratio	6	2	
percent_change	4	3	
surviving the negative class, uncapped		125	
after the 50% composition cap		**36**	

Table 8: From corpus to items for the recomputable family, separating the three losses: extraction, the same-table negative class, and the composition cap.

C.2 Conflict supply: 13 checkable disagreements from 378 screened erratum pairs

Erratum and commentary pairs promise machine-checkable conflicts: a notice states a corrected value, the original states the superseded one. Screening 378 full-text erratum pairs produced 27 candidates; auditing those candidates identified **13** numeric disagreements, a screen precision of 0.48, including **ten** live supersessions, 0.37. The stricter endpoint is the one an item needs. Both are discovered yield, not total supply, and the excluded side of the screen was read as well: a seeded sample of 41 of the 351 excluded pairs, stratified over the fail reasons, was read by hand. Of the 41, 35 were correctly excluded, 2 were genuine numeric corrections the screen missed, 3 were not adjudicable, and 1 was a real correction carrying no number. Sampling among screen exclusions estimates the share of genuine corrections among excluded pairs, not the screen's false-negative rate, and the archived audit records each row's verdict but not its stratum, so no population estimate over the 351 is formed from these counts. What the sample establishes is qualitative: the excluded set is not empty of genuine corrections, so the 27 candidates are not an enumeration of the family's supply. Commentary pairs give 23 of 131. Against a family target of 300, the discovered supply is roughly 33 items.

The notice classes locate part of the attrition: of the 378, 231 are substantive, 83 are purely administrative (authorship, funding, metadata), 41 concern images alone, and 23 are unclassified. Administrative and image-only notices together are $124/378 = 32.8\%$, a large minority and not a majority, so the classes account for some of the shortfall without establishing what the correction record is mostly about.

C.3 Why we report this

Each census answers a question a benchmark builder must otherwise guess: how many items the stated rules retain from the corpus at hand. In all three families the yield was substantially below the design target, and in two of them the shortfall was severe enough to change what the paper can claim. Reporting the census makes those decisions auditable, and makes the negative results, a family this construction did not fill, a finding rather than a silence.

D The strict-path rejection audit

Both registered scoring paths are computed from the same stored answers, so re-scoring is deterministic and costs nothing. Doing so isolates a disagreement. Of 888 scored answers, **422** (47.5%) are accepted by the companion path and rejected by the strict one, against 45 in the other direction. In **422 of 422** (100.0%) the gold numeral is literally present in the answer.

The rejections track surface rendering. A gold value of 25.7 carrying unit `yr` against an answer reading "25.7 years"; $5{,}230$ against "5,230 patients"; 100 against "100 patients"; 5.0 carrying unit % against "5%".

Numeral presence is not by itself semantic correctness, so all 422 companion-accepted, strict-rejected answers were checked, not a sample (Table 9). In **34** the numeral sits inside a negated or contrastive window of ± 90 characters and cannot be assumed correct on this evidence; the check is positional and does not certify the other 388 as semantically correct, only as free of the pattern that would make them doubtful. **353 of the 422** also have the citation scored correct on the same record, so in five cases out of six the rejected answer names the right passage as well as the right number.

gold numeral literally present in the answer	422 (100.0%)
numeral inside a negated or contrastive window (± 90 characters)	34 (8.1%)
numeral with no such marker in that window	388 (91.9%)
citation also scored correct on the same record	353 (83.6%)

Table 9: Structured checks over all 422 disagreements.

A seeded sample of 30 was read by hand. Every one is consistent with rendering mismatch. Two exposed a different problem, where the gold value itself looks mis-extracted: 320 typed as `d` for days,

taken from the gene name *miR-320d*, and 2,000 typed as `yr` where the source sentence appears to mean a calendar year. Those are extraction defects rather than scoring ones, and we report them as observed rather than quantified.

The 422 spread evenly over both families and all three variants. This makes a family-specific explanation implausible, though even prevalence does not guarantee cancellation in a contrast, since the rate at which a rejection flips an item's score need not be equal across arms.

D.1 The consequence for condition (a)

The consequence is the point. Strict is one of the two paths condition (a) is scored on, and it moves treatment `attribution_correct` in the `full` cell from 0.527 to 0.237. The audit of the 422 establishes numeral presence in every one; semantic equivalence item by item is a stronger property it does not settle, and our checks are pooled over both families, so they cannot apportion that 29-point movement to either arm. What they do support is narrower and sufficient: a scorer that rejects 47.5% of answers carrying the gold value, 91.9% of them with no negation or contrastive marker within 90 characters of the numeral and 83.6% with a correct citation on the same record, is not a neutral instrument for measuring attribution. A contrast read off this column would therefore be confounded with the scorer's normalisation, by an amount these checks do not determine, and we report none.

E What the tiers decide that a flat audit does not

The tiering is not decoration, and our own build supplies the case. Two rules are scored against the same item-level null of 0.111. "The passage carrying the quoted sentence" is a T1 rule scoring $\mathbf{0.996}$ on the control family, a $9.0\times$ lift, and it consults neither the gold value nor the map, so the figure is executable as it stands. "The document contributing the fewest passages" is a T2 rule scoring $\mathbf{0.230}$ pooled, a $2.1\times$ lift.

A gate of the form "no rule may predict the answer above chance" ranks these by lift and condemns the first, which is four times the second. That verdict is wrong twice over: it would reject a benchmark for being solvable from its own evidence, which is what a benchmark is for, and it would rank the genuine build failure second. Our gate inverts the ranking. The T1 rule is *permitted* and published as the ceiling, which is what puts the capability reading of any later effect in question. The T2 rule is a **hard build failure** at $z = 19.3$, because the metadata it consumes is supplied by the experimental condition under test. Lift against chance does not separate these; information tier does.

E.1 Does the tier distinction change what ships?

Section 4 claims the combination rather than any ingredient, and Table 11 places this paper alone in the build-granularity, tiered cell. Whether that cell matters is checkable on our own pool by holding the catalogue fixed and varying only the decision policy. Three policies are compared over the 348 attribution items. An item-level policy drops an item when its probe names the gold passage uniquely; it deploys one probe per tier, the rule with the most unique hits among the catalogued rules of that tier, since the published item-level filters of Table 11 act on one probe's verdict per item. Because the probe is chosen on the observed labels, the null prices the selection and not the chosen rule, as Appendix A requires of the gate's own fitted rules: in each of 2,000 seeded draws gold is redrawn uniformly over the item's full candidate pool of k passages, the best probe is re-selected inside the draw, and its drops are counted. At T1 the catalogued probes are oracle rules, their value-stating set built from the gold value (Section 4), so no label-blind filter could deploy them. The executable candidate family is those same 13 rules re-executed with the value-stating set re-derived from the literal context flanking the mask, the substitution of Appendix A.3; selection runs over that family, separately in each scope, and the executable row is the row the comparison rests on. The oracle row is kept only to mark what a filter cannot deploy.

The item-level filter at T0 drops 51 items and ships 297. Choosing the best of 14 T0 probes on redrawn gold drops 48.7 on average, so the observed 51 sits at $z = 0.73$ against the selection: this policy's removals are statistically indistinguishable from picking the best of 14 probes by luck, and the 297 it ships still carry the T2 shortcut of Appendix J. The oracle T1 filter drops 343 and ships 5, but it reads the gold value to build its answer set and is reported

policy	probe (pooled selection)	drops	null drops	ships
item filter at T0	`t0_const_P7`, best of 14	51 (14.7%)	48.7 (14.0%), $z = 0.73$	297
item filter at T1, oracle	`t1_fewest_citation_ph_among_value_stating__unconditional`, best of 13	343 (98.6%)	41.7 (12.0%)	5; not deployable
item filter at T1, executable	the same, `__executable`, best of 13	340 (97.7%) pooled; by family 147 of 148 and 199 of 200	41.4 (11.9%), $z = 54.6$	8 pooled; 1 treatment and 1 control under their own probes
tiered build gate	whole catalogue, by tier	build-level	per rule	348 as shipped, fails; 179 after both repairs, passes

Table 10: Three decision policies over the same 348 attribution items and the same catalogue. An item is dropped when the designated probe names its gold uniquely. The null is the selection-aware drop count: over 2,000 seeded draws, gold is placed uniformly over the item's full candidate pool, the best probe is re-selected within the draw, and its drops are counted; z is against that distribution (T0 null sd 3.18; T1 executable null sd 5.47). Selection runs per scope, so the family columns of the executable row are two different probes and their ships do not sum to the pooled figure; both are named in the text.

only to mark that it is not deployable. The executable T1 filter is the comparator the argument rests on, and it removes both families rather than one. On the treatment family the selected probe is `t1_value_stating_not_containing_quoted__executable`, which drops 147 of 148 items (99.3%) against a selection null of 19.4, $z = 39.0$; on the control family it is `t1_contains_quoted_sentence__executable`, which drops 199 of 200 (99.5%) against 22.3, $z = 38.5$. One item of each family survives. Those two probes are exactly the two rules Table 6 reports as their families' executable ceilings, and their unique-hit counts reconcile with the published means: a mean fractional score of a requires a unique-hit fraction of at least $2a - 1$, since a correct answer set of size two or more contributes at most $1/2$, so 0.994 on 148 items requires at least 147 unique hits and 0.996 on 200 requires at least 199. The counts are 147 and 199.

An earlier version of this row read the comparison the other way, as dropping no treatment item and removing the control family alone, and concluded that a deployable T1 filter leaves the treatment family whole. That conclusion was ours and it is withdrawn. What the row executed was not the ceiling rule but the derived value-stating set without that rule's quote-exclusion filter, a weaker probe that names a passage uniquely only where one value-stating passage exists; the label promised one rule and the behaviour was another's, and the published mean of 0.994 is what makes the two incompatible. With every T1 rule re-executed on the derived set, the oracle and executable rows agree on what the policy does, so the near-total removal is not an artefact of the oracle's access to the gold value. A deployable item-level filter at T1 therefore does not shrink the benchmark, it consumes it: 8 items ship pooled, one per family under the family's own probe.

The tiered build-level gate reaches neither outcome. T0 passes after repair; T1 is permitted to beat its null and its executable ceiling is published as an estimand bound, 0.994 treatment and 0.996 control; T2 fails as shipped over four pairs and passes after both repairs at 48.6% of supply removed, in both families.

Why the item filter is given one probe and the gate the whole catalogue. The obvious objection to Table 10 is that it does not hold evaluated coverage fixed: the item policies consult one designated probe per tier while the gate evaluates every rule. That asymmetry is a property of the policies and not a defect of the comparison, and the first formulation of this comparison is what established it. That formulation asked whether *any* rule in a tier solves an item, which is the item-level policy that does use a catalogue, and it is dominated by catalogue size: with m rules each independently naming

one of nine passages at random, the chance that at least one names gold is $1 - (8/9)^m$, which is 0.81 at T0's $m = 14$ and 0.98 at T2's $m = 32$. Against signal-free rules such a policy drops almost everything, so it was withdrawn as a measurement. As a statement about policy design it is the point: an item-level filter cannot carry a catalogue, because dropping any item some rule names removes the pool before any rule's signal is read; it must designate one probe, and designating one probe is what makes it blind to whatever the rest of the catalogue would catch. Build-level blocking carries no such constraint, since a rule that fires at chance leaves the build's verdict untouched, which is why the catalogue can be add-only. The two figures are a toy-model null: the actual catalogue's rules are correlated and set-valued, so $1 - (8/9)^m$ bounds independent singleton predictions and not this catalogue. What Table 10 holds fixed is the rule catalogue, the items and the labels; what it varies is the granularity at which a verdict acts, and the coverage difference is a consequence of that granularity rather than a free parameter we could have set otherwise.

A second formulation priced a fixed probe where the selection was the policy, placed the null over the probe's answer set instead of the candidate pool, and deployed the oracle T1 probe. A third corrected those three but executed one weaker probe under the ceiling rule's name. Table 10 corrects all four defects, and names the executed rule for every figure it reports, which is what the fourth defect cost us.

F METHOD-LEVEL COMPARISON WITH PUBLISHED AUDITS AND FILTERS

Section 2 states that none of the gate's ingredients is new. Table 11 makes that statement checkable on the four axes the gate turns on, for the systems named in the venue context, for the item-level filters a literature sweep surfaced, and for the audit line of Section 2. The table and the per-cell evidence below need a page, which the main text does not have.

Column definitions. The columns are deliberately narrow. *Simult.*: two or more distinct shortcut classes are audited in one pass over one build, each with its own reported cell. A single aggregate predictability criterion removes many biases at once without separating them, which is filtering rather than auditing under this definition and is scored N for that reason, not as an absence of effect. *Access*: the probe is denied part of the item and scored under that restriction; a closed-book condition on the *system* under test counts as partial. *Null*: a zero-capability reference is reported beside the audited score. *Blocking*: an audit failure changes what ships, at item, build, claim or no granularity. Y, P and N mean yes, partial and no, each evidenced from the cited full text; a cell the full text does not settle is marked ? (unknown) and is left there rather than scored N.

Row evidence. Locators are the section, table or figure of the cited full text; quoted phrases are the sources' own words.

SciFact. Two partial-input probes, claim-only and abstract-only (Table 3), each trained with the other half withheld; Table 4 adds oracle-abstract and oracle-rationale settings. Label accuracies of 44.5 (claim-only) and 53.3 (abstract-only) are reported with no chance or majority line, and the authors read the claim-only figure as giving "some reassurance" that the negation procedure introduces no "obvious statistical artifacts" (§6.1). Nothing is scored against a threshold or stopped; the construction-side control is procedural (§3.2).

MultiVerS. A verification model, kept because the venue context names it. Its full text reports no partial-input, random or majority baseline, so the three audit cells are N on the paper's own evidence. Its data handling (§A.1) drops HealthVer evidence without a CORD-19 match and claims with support and refutation in one abstract, which is admissibility filtering rather than an audit-triggered stop. It is not a comparator on these axes; it is the kind of system a contaminated benchmark would inflate.

PaperQA / LitQA. One shortcut class, parametric recall, handled by a training-cutoff rule (§5.1). Closed-book GPT-4 and Claude-2 are conditions on the system, not a restricted probe. Table 2 carries a Random row (10.2 correct, 29.5 incorrect, 10.3 unsure of 50) and the text notes that the closed-book models "perform better than random" on post-cutoff questions; the items ship unchanged. This is the cleanest published case of a T0-style excess observed and not acted on.

PaperQA2 / LitQA2. Two exclusion channels, alternative-source and abstract answerability (§2, §8.2.1). The exclusion probe has more access than the task, not less: questions were removed where

Work	Object audited	Simult.	Access	Null	Blocking (granularity)
Systems named in the venue context					
SciFact (Wadden et al., 2020)	sci. claim benchmark	P	Y	P	N (none; post-hoc reassurance)
MultiVerS (Wadden et al., 2022)	verification model	N	N	N	N (none; §A.1 is admissibility)
PaperQA / LitQA (L'ala et al., 2023)	agent + 50 items	N	P	Y	N (none; excess over random shipped)
PaperQA2 / LitQA2 (Skarlinski et al., 2024)	agent + 248 items	P	P	P	Y (item; alt.-source probe)
CORE-Bench (Siegel et al., 2024)	reproducibility tasks	N	P	N	P (item; design constraints only)
ALCE (Gao et al., 2023)	citation benchmark	P	P	Y	N (none; §3.4, App. D)
Item-level gates surfaced by the sweep					
AFLite (Le Bras et al., 2020)	dataset filter	N	N	P	Y (item; threshold τ)
GPQA (Rein et al., 2023)	expert MCQ	N	N	Y	Y (item; human probe)
HLE (Phan et al., 2025)	expert closed-ended	N	N	P	Y (item; LLM probe)
SeedRG (Liu et al., 2026a)	RAG generator	P	Y	Y	Y (item; reject-regenerate)
Option-only gate (Ovcharov, 2026)	legal MCQ bank	P	Y	Y	Y (item; binomial null)
Audit line (Section 2)					
Charge validity (Liu et al., 2026b)	legal IR benchmarks	P	Y	Y	N (report; P for roll-up)
PARALLAX (Hussain & Kantarcioglu, 2026)	hallucination corpora	N	N	Y	N (none)
Auditing the audit (Li et al., 2026)	audit pipelines	Y	N	N	P (claim withheld)
ABC checklist (Zhu et al., 2025)	agentic benchmarks	Y	N	Y	P (checklist)
Partial-input baselines (Kaushik & Lipton, 2018; Gururangan et al., 2018; Poliak et al., 2018)	NLI / RC data	P	Y	P	N (none)
This paper (Section 4, App. A)	own provenance build	Y	Y	Y	**Y (build; $z > 3$ fails)**

Table 11: Method-level comparison on the four axes the gate turns on. Column definitions and the per-cell evidence locators are in the text of this appendix; Y/P/N are evidenced from the cited full texts. Item-level blocking, restricted-input probes and null calibration are each published; the combination at build granularity with a legitimate/fatal tier distinction is what remains, and Appendix E.1 measures what it changes on our own pool.

"an existing AI system or a human annotator could answer the question using an alternative source" (§2). Figure 2B reports five no-retrieval LLMs against a human line and no chance line. Exclusion is applied before release, with no scored threshold and no build-level disposition.

CORE-Bench. No shortcut class is probed. Easy, Medium and Hard grade what the agent receives (Table 3), a difficulty ladder rather than an audit restriction, and Table 4 reports no guess or random score. The ten admissibility criteria (Table 2; 5,090 to 90 capsules) and the rule that each task has "at least one question that cannot be solved by guessing" (§3) are design constraints enforced at build time, not probe-triggered failures. It answers a different question, whether agents can reproduce papers, and is not a comparator on these axes.

ALCE. Two "shortcut cases", the top-1 passage and its first two sentences returned as the answer (§3.4, Appendix D), both the same metric-gaming class. CLOSEDBOOK and oracle retrieval are conditions on the system; CLOSEDBOOK is reported beside every open-book row (Table 4), and Table 11 gives the copy-the-passage reference for citation quality. The paper's claim is that its metrics are "robust to shortcut cases"; nothing is stopped.

AFLite. One predictability criterion aggregates all biases, and the probes are "typically a linear classifier" over embeddings of the full input (Algorithm 1), weak by capacity rather than by information.

Predictability is compared against a stopping threshold τ and against human accuracy, not against chance. The algorithm removes the k most predictable instances and repeats (§3.2); the build never fails, it shrinks (SNLI 550k to 182k).

GPQA. Non-expert validators had "unrestricted access to the web". "25% accuracy is random chance" is stated beside the 34% non-expert and 39% GPT-4 figures (§1). Diamond keeps items where two of two experts agree and at most one of three non-experts is correct (Table 2); the same table notes that validator accuracies on filtered splits "are skewed by selection effects".

HLE. Frontier LLMs see the whole question. "Questions are rejected if LLMs can answer them correctly" (§3.1); for multiple choice the criterion is that models on average "do worse than random guessing" (§3.2). Residual solvable items are left in by choice, "instead of strongly adversarially filtering" (§4.1).

SeedRG. Two checks, knowledge leakage and reasoning-graph structure (Fig. 2). The leakage probe is "the LLM with q alone (no context)" (§4.1); Table 1 defines leakage error as no-context accuracy and answerability as gold-context minus no-context accuracy. "If any response contains the correct answer a, the sample is rejected" over three queries; a single hit rejects, so the decision itself carries no statistical null.

Option-only gate. The probe sees "the four options with the question withheld" (§5.1) and is scored against a binomial null over eight option permutations, "P(X $\geq$ 5) = 0.027"; 1,419 items were answered blind on all eight permutations against 0.2 expected (Fig. 1). "We reject an item at five or more hits" (§5.1), retaining 8,128 of 11,990. The answer-position habit is measured and subtracted per model (Table 4). The paper's central result is that "filtering a bank against one model leaves the strongest models in the set reading the options exactly as before" (§6): a model-based gate does not transfer, which is the argument for the enumerated, model-free rules of Section 4.

Audit line. Liu et al. (2026b) exclude out-of-specification cells from their primary roll-up (Table 2, rule 5), a gate on the report over third-party benchmarks with no build of their own. Li et al. (2026) frame their gate as "a withholding rule, not a scalar metric" and scope themselves away from construction, "a validity audit is a second-order measurement" (§1). Hussain & Kantarcioglu (2026) report a TF-IDF null; Zhu et al. (2025) audit several classes without gating a build; the 2018 partial-input baselines are the instrument the Access column descends from.

This paper. 63 rules over question surface (T0, 18), passage text (T1, 13) and provenance and metadata (T2, 32), 189 rule–scope pairs, one pass (Section 4). Tiers are typed views (Appendix A); nulls follow each rule's own candidate set per item; T0 and T2 excess at $z > 3$ pointwise is a build failure, and the shipped build fails the gate, which we report rather than shipping a smaller set.

Named systems that are not comparators, and one omission. MultiVerS and CORE-Bench appear because the venue context names them, and their rows are answered from their full texts; neither audits a shortcut. MultiVerS is a model, and the question it bears on is what a contaminated benchmark would inflate. CORE-Bench's tiers grade the agent and its filters are admissibility criteria; its ten-criterion funnel is a supply-census precedent (Section 5), not a gate. ERASER (DeYoung et al., 2020) was considered and left out: its sufficiency and comprehensiveness metrics restrict what a *model* sees in order to audit the model's rationales, not the benchmark. Dynabench, SWAG and HellaSwag, WinoGrande and SWE-bench Verified are named in the read texts as further adversarial-collection or verified-subset precedents and are not tabulated.

G THE SCOPE OF THE STRUCTURAL RESULT

Numerical uniqueness is sufficient for one identification strategy, not necessary for valid origination labels: two documents may state the same value while the original remains identifiable through attribution language or independently established publication history. We did not build such a construction, so we report the tension for the strategy we used and flag the alternative as untested rather than as ruled out.

The second limb needs its own assumption, because uniqueness and independence are different properties. It holds here because our map and our labels share a witness: the gold label is the corpus document that the review's citation resolves to, and the map is the passage-to-document assignment

over those same documents, so a map that alone determined origination would be determining it from the resolution that produced the label. A construction that sourced its map elsewhere, from separately collected publication records, and rendered excerpts omitting decisive material, could have a map that carries genuine information and is not circular. The second limb does not reach that design, and we do not claim it does.

G.1 What the ceiling diagnoses, and the usable zone

Within this identification strategy the ceiling and the ground truth are not independent: relaxing the uniqueness condition restores headroom and destroys the labels at the same time. The constraint stated in Section 4 is therefore not an artifact of this corpus, though it is specific to establishing origination by numerical uniqueness. It is also not an impossibility result. The executable rule that attains the ceiling uses two further properties of the shipped rendering, that the masked value can be read off the literal context flanking it and that the passage printing the quoted sentence is among the candidates (Appendix A); uniqueness alone would leave the quantity to be found. A construction that kept uniqueness and changed either rendering choice lies outside what was built here, and we cannot price it. A construction grounding origination in attribution language or in independently established publication history would not face it in this form; it too lies outside this build, so we name that route untested rather than closed.

The usable zone is between: origination inferable from text but requiring inference, with the map as a shortcut models are *observed* to take rather than one they are forced into. A benchmark in that zone measures *elicitation*, whether scaffolding causes a model to perform attribution it could have performed unaided, rather than capability. A model may well go from 0.4 without the map to 0.8 with it while a deterministic text-only rule scores 1.0; nothing in the ceiling forbids that, only the capability reading of it. Reaching the zone changes what the experiment can claim. We do not claim the elicitation reading for this construction's aggregate: at 0.994 and 0.996 the text suffices on all but a handful of items, and on that handful our evidence does not say whether the shortfall is an implementation limit or genuinely missing information.

H The ordering deviation

Condition (b) registers an expected ordering, `full` $\geq$ `no_prov` $\geq$ `raw` on the treatment family, and records deviation without failing on it. The ordering does not hold.

On the treatment family `raw` is the best variant on every path: 0.554 against `full`'s 0.541 and `no_prov`'s 0.493 on companion `citation_correct`, 0.622 against 0.574 and 0.554 on the strict path, and 0.554 against 0.527 and 0.493 on companion `attribution_correct`. Supplying no provenance scaffolding at all beats supplying it, throughout.

This is a direction the registered hypothesis did not predict, and it would be the most quotable number in the paper. **We do not claim this result**, because the instrument failed its own validity gate, the scoring defect above moves one of these columns by tens of points, and the separation between `raw` and `full` is a few items at $n = 148$. A result extracted from an instrument just shown to be broken is not evidence, and the discipline that produced the three findings above is the discipline that forbids reporting this one.

I The verifier's four rejection paths

The citation verifier rejects an answer on four paths: an empty authentic-passage set, a citation to an injected passage, affirmation of a payload-carried false claim, and a citation to a non-authentic passage id. The pilot carries no attack items and every case has a non-empty authentic set, so the first three cannot fire on the pilot item set; those surfaces belong to the injection experiment the pilot does not include. The fourth fires on a model that emits an identifier outside the case, and none of the 888 answers did. Condition (c), which requires at least one flag, was therefore carried entirely by a model failure mode that did not occur.

J THE TWO RESIDUAL SHORTCUTS: MECHANISM, NULL CALIBRATION, AND THE REPAIRED SUBSET

After repair the gate still fails four rule–scope pairs, from two T2 rules (Section 4). This section measures what can be measured about each on the shipped build with the gate's own scorer, so every cell is directly comparable with the published ones. The report is seeded at $20{,}260{,}914$ with $B = 2{,}000$ draws where a draw is needed. Scopes are named as in the gate report: `multi_source` is the treatment family ($n = 148$), `single_source` the control ($n = 200$), and the attribution pool is their union ($n = 348$). Both repairs are realised as filters over the shipped items, one by keeping the items that already satisfy the property and one by seeded removal; the section keeps the two apart, then applies both and re-runs the whole catalogue .

J.1 FEWEST PASSAGES IS A COUNTER

The rule returns the passages of whichever document contributes the fewest to the item, ties included, and it fires for an arithmetical reason. The gold document contributes exactly one passage in every item, $148/148$ treatment and $200/200$ control, so gold is always inside the argmin set. The rule does not identify gold. It exploits a property the construction guarantees of gold, the single contributed passage, to exclude every distractor document that contributed more, and each exclusion raises the expected score above chance. Where every document in an item contributes one passage, the argmin set is the whole nine-passage pool and the rule earns $1/9$; whenever some distractor document contributes two or more, those passages leave the set, the set shrinks, and $1/|S|$ rises. The shipped mean answer-set size is 6.74 against a pool of 9. A repair therefore has to hold the returned set at the pool, which is what one passage per document does.

J.2 THE REPAIR, EVALUATED ON THE SHIPPED ITEMS THAT ALREADY SATISFY IT

No rebuild is needed to evaluate it, because 183 of the 348 attribution items already have one passage per document. On that subset the rule returns all nine passages in every item and scores exactly 0.1111, its null, with zero variance; the gate marks the cell degenerate, which under the validity contract of Appendix A is the non-blocking disposition. The two families behave identically, 91 treatment items and 92 control items each at the null. The shipped cells are 0.1562 ($z = 7.746$) treatment, 0.2852 ($z = 17.700$) control and 0.2303 ($z = 19.321$) pooled, and the 165 violating items alone score 0.3625 pooled (Table 12). The whole of the shortcut lives in the violating subset.

	as shipped		one passage per document		violating		
family	accuracy	z	n	accuracy	n	accuracy	filtered share
treatment ($n = 148$)	0.1562	7.746	91	**0.1111**	57	0.2281	38.5%
control ($n = 200$)	0.2852	17.700	92	**0.1111**	108	0.4334	54.0%
pooled ($n = 348$)	0.2303	19.321	183	**0.1111**	165	0.3625	47.4%

Table 12: "The document contributing the fewest passages" on the shipped build, split by whether an item already satisfies one passage per document. The null is 0.1111 throughout. The one-passage-per-document cells have zero null variance and undefined z, the analytic degeneracy of Table 5. Filtered share is the fraction of the scope removed by keeping only the items that already satisfy the property.

J.3 WHAT FILTERING COSTS

Keeping only the items that satisfy one passage per document removes the 165 violating items: 47.4% of the shipped attribution pool, 54.0% of the control family and 38.5% of the treatment. That is the price of the filter on the shipped pool, and it is the quantity this section reports. A construction that draws distractor passages under the rule from the start is a different construction with its own supply, and its yield is a property of that build rather than of this one. We do not adopt the filter, and we make no claim about which of the two instruments, the 348-item pool with a disclosed T2 channel or the 183-item subset, is the better one; validity and sample size are not traded by assertion. The factual position is that the repaired subset is characterised in this section, the shipped benchmark does not satisfy the validity gate, and we do not proceed to the full experimental grid.

J.4 Retraction status was left uncontrolled

The retraction rule returns the passages of the item's retracted documents and abstains when there are none, which it does in 192 of 348 items. In the control family the gold document is retracted at 0.1650 against a distractor rate of 0.1069, and the rule scores 0.1551, $z = 3.479$. In the treatment family gold is retracted at 0.0608 against 0.0954, and the rule scores 0.0904, $z = -1.813$. Pooled, the rates are 0.1207 against 0.1020 and the rule scores 0.1275, $z = 1.881$. The construction left retraction status uncontrolled, and the shipped cells measure the imbalance it inherited. The two families' signs differ, and we draw nothing from that. The treatment cell does not establish a reliable negative association, and it does not offset the control failure. A family-dependent selection bias is compatible with both cells, and a bias of that kind is itself an interaction that can contaminate a treatment–control comparison, so the reversal adds to the case for treating the T2 column as uncontrolled.

J.5 A null calibration, which is not a repair

The repair the channel needs is a construction that preserves genuine origination labels and matches the gold document's retraction rate to the distractors'; the next subsection realises it as a filter on the shipped build. This subsection computes a different object, a random-label null calibration: gold is re-drawn uniformly over each item's own passages and the rule is scored on each of $B = 2{,}000$ seeded draws. The result is the null by identity. The rule's answer set S depends only on the retraction flags and never on gold, so for a fixed non-empty S over k passages

$$\mathbb{E}[\text{score}] = \Pr(g \in S)\,\frac{1}{|S|} = \frac{|S|}{k}\,\frac{1}{|S|} = \frac{1}{k},$$

which is $1/9$ here whatever the retracted set's size. The simulated means are 0.1113 pooled, 0.1113 control and 0.1111 treatment, and the intervals in Table 13 are 95% intervals across the simulated draws, not uncertainty about a repaired benchmark's performance. The calibration shows one thing: the shipped excess is carried by the gold–retraction association and not by the size of the retracted set. It does not show that a rate-matched construction with real labels would close the channel; that is the question the balanced filter of the next subsection answers on the shipped items. This is the paper's own null-model rule applied to the paper: a number that does not require the mechanism to be true is not evidence, and this calibration is such a number.

	retraction rate		as shipped		random-label null	
family	gold	distractors	accuracy	z	mean	95% interval
treatment ($n = 148$)	0.0608	0.0954	0.0904	-1.813	0.1111	$[0.0909, 0.1343]$
control ($n = 200$)	0.1650	0.1069	0.1551	3.479	0.1113	$[0.0872, 0.1371]$
pooled ($n = 348$)	0.1207	0.1020	0.1275	1.881	0.1113	$[0.0950, 0.1299]$

Table 13: "Retraction status" on the shipped build, with the random-label null calibration. The null is 0.1111 throughout; the calibration means equal it by the identity above, and the intervals are across the $B = 2{,}000$ simulated draws. The rule abstains in 95 treatment, 97 control and 192 pooled items, where no document is retracted.

J.6 The retraction repair, realised as a balanced filter

The repair balances retraction status between the gold document and the distractors. On the shipped build it is realised by seeded removal (seed $20{,}260{,}914$) of gold-retracted items until the gold rate no longer exceeds the distractor rate, with real labels throughout (Table 14). In the control family, the one that failed, 14 of 200 items (7.0%) are removed; the gold rate falls from 0.1650 to 0.1022 against a distractor rate that moves from 0.1069 to 0.1028, and "retraction status" goes from 0.1551, $z = 3.479$, a gate failure, to 0.1204, $z = 0.73$, which does not cross. Its mirror, "not retracted", moves from $z = -2.082$ to 0.372. Pooled, 9 of 348 items (2.6%) are removed and z falls from 1.881 to 0.565. In the treatment family nothing is removed, because the gold rate there is already below the distractor rate, and its cells are unchanged. This is a filter over items drawn under the unbalanced rule and then thinned; it is not a from-scratch balanced construction, and it says nothing about the yield such a construction would have.

	as shipped			balanced			removed	
family	n	gold / distr. rate	z	n	gold / distr. rate	z	items	share
treatment	148	0.0608 / 0.0954	-1.813	148	0.0608 / 0.0954	-1.813	0	0.0%
control	200	0.1650 / 0.1069	3.479	186	0.1022 / 0.1028	0.730	14	7.0%
pooled	348	0.1207 / 0.1020	1.881	339	0.0973 / 0.0985	0.565	9	2.6%

Table 14: "Retraction status" before and after balancing gold and distractor retraction rates by seeded removal of gold-retracted items. The null is 0.1111 throughout and z is the gate's own statistic. The pooled removal is computed on the pooled rates and is not the union of the per-family removals.

J.7 The T2 catalogue on the one-passage-per-document subset

Removing the fewest-passages channel does not by itself settle the T2 column, because every other metadata rule has to be scored on the same items and a rule that passed on the shipped build can fail on a selected subset. The report therefore runs the whole T2 catalogue, all 32 rules, on the 183-item subset with the shipped gold labels and no simulation. **No T2 rule crosses the gate's failure threshold of $z > 3$ in any scope.** Five cells, over four rules, sit above their null without crossing it: "not retracted" conditioned on what the evidence licenses, at 0.5330 against 0.5000, $z = 1.5$, on the treatment family and 0.7678 against 0.7514, $z = 1.5$, pooled; "retraction status" at 0.1293 against 0.1111, $z = 0.816$, on the control family; "the first document card" at 0.1196, $z = 0.258$; and unconditioned "not retracted" at 0.1124, $z = 0.234$. The largest of them is a z of 1.5.

The two residual rules of the shipped build are scored in Table 15. Fewest passages sits at 0.1111, degenerate, in all three scopes. Retraction status scores 0.1095 pooled against a null of 0.1111, $z = -0.111$; on the 92 control items 0.1293, $z = 0.816$; on the 91 treatment items 0.0896, $z = -1.263$. The control cell's observed excess over the null falls from $0.1551 - 0.1111 = 0.0440$ on the full shipped family to $0.1293 - 0.1111 = 0.0182$ here, which is a statement about the two point estimates; the accompanying fall in z also reflects the smaller n and is not itself a measurement of a weaker association.

The limits are stated here so that the result is not read past them. The measurement is observational: the subset is the set of shipped items that happen to satisfy the passage-count property, which the filter was not designed to produce. The per-family samples fall to 91 and 92, so the retraction test there has low power, and a control z of 0.816 is as consistent with a smaller residual association as with none. The result also says nothing about whether a one-passage-per-document construction would eliminate the retraction channel in general. What it establishes is a threshold outcome on these 183 items: over all 32 T2 rules, none crosses the gate's failure threshold. Non-crossing at this n is not a demonstrated absence of a residual metadata association, and the shipped benchmark the subset was drawn from still fails.

		fewest passages		retraction status		
family	n	accuracy	z	accuracy	z	abstains
treatment	91	0.1111	degenerate	0.0896	-1.263	48
control	92	0.1111	degenerate	0.1293	0.816	28
pooled	183	0.1111	degenerate	0.1095	-0.111	76

Table 15: Both residual T2 rules scored on the one-passage-per-document subset of the shipped build, with the shipped gold labels. The null is 0.1111 throughout. Fewest passages has zero null variance there and its z is undefined. The corresponding full-set values for retraction status are $z = -1.813$ treatment, 3.479 control and 1.881 pooled (Table 13).

J.8 Both repairs, and the whole catalogue

The two filters are then applied in sequence within each family, one passage per document first and retraction balancing on what remains, giving one membership manifest per family; the pooled scope is the union of the two manifests, so its count is the sum of theirs by construction. On that manifest

every catalogued rule is re-scored with the shipped gold labels: 18 of 18 T0 rules and 32 of 32 T2 rules, the 59 enumerated rules plus the 4 fitted rules, whose null is the simulated maximum over re-selection in each of 2,000 draws (Table 16). **No T0 or T2 rule crosses the failure threshold of $z > 3$ in any scope**; the repaired build passes its own gate. The dispositions that make a pass auditable are recorded in the same report, per scope: no cell is invalid, so the conditioning-validity contract of Appendix A.2 holds for every item of every scored cell, and no fitted cell is unresolved, so every simulated max-null had non-zero spread. Of the 46 enumerated T0 and T2 cells, 26 on the treatment family, 29 on the control and 26 pooled are degenerate, the rule returning its whole conditioning set on every item and therefore earning exactly its null with zero variance; all of them are T2 rules, and a degenerate cell is reported as degenerate rather than as a pass, so the pass above is carried by the cells that could have crossed. The price is 91 of 148 treatment items (38.5% removed) and 88 of 200 control items (56.0% removed), 179 of 348 pooled, a 48.6% reduction. Almost all of the cost is the passage-count filter, since retraction balancing removes four control items and no treatment item once the first filter has run. The cells that sit above their null without crossing are the constant position P_7 at $z = 1.929$ pooled and 1.631 on the treatment family, "not retracted" conditioned on what the evidence licenses at $z = 1.5$ on the treatment family and pooled, and on the control family the constant position P_6 and the question-length quantile rule, both at $z = 1.771$. What the 179 items are is the shipped pool after two filters. A construction that draws under both rules from the start would have its own supply, which could exceed 179, and could face shortcuts this catalogue has not enumerated, so a pass here is a pass of this catalogue on this subset and not a clearance of that build.

scope	shipped	after passage filter	after both	removed	T0 rules $z > 3$	T2 rules $z > 3$
treatment	148	91	91	38.5%	0 of 18	0 of 32
control	200	92	88	56.0%	0 of 18	0 of 32
pooled (union)	348	183	179	48.6%	0 of 18	0 of 32

Table 16: The gate re-run on the shipped build under both repairs, applied as filters within each family; the pooled row is the union of the two family manifests. All 63 catalogued rules are scored in every scope, the 4 fitted rules against their simulated maximum null. The largest remaining statistics pooled are $z = 1.929$ (T0, constant position P_7) and $z = 1.5$ (T2, "not retracted" conditioned).

J.9 What this section establishes, and its scope

On the shipped build both rules remain gate failures, over the same four rule–scope pairs, and the shipped build fails the gate. Under both repairs, applied as filters, the repaired build passes its own gate at 179 of 348 items, 91 treatment and 88 control. This section localises one channel, balances the other, prices both filters, and scores all 18 T0 and all 32 T2 rules on what they leave. Its subject is the shipped pool. A construction drawing under the repaired rules from the start could both recover supply and meet shortcuts this catalogue has not enumerated, and it would need its own catalogue run; a passing subset does not make the shipped benchmark satisfy the gate.