
Is the Endpoint Well-Posed? Measurement Contracts for AI-Assisted Scientific Discovery

Odysseus Kaloidas
UC Berkeley

Abstract

1 Scientific-discovery benchmarks often score a proposed finding against the con-
2 clusion of a published paper. Interpreting this agreement requires knowing how
3 much of the conclusion is already present in the supplied text and whether other
4 conclusions could be scientifically valid. We study the first question by revealing
5 more source text and replacing it with text from another paper. We examine the
6 second by testing what retrieval of a historical conclusion can establish. On an
7 author-constructed CS/AI disclosure proxy ($N=86$), mean lexical recoverability
8 rises from 0.017 to 0.235 with disclosure. Replacing evidence with another paper’s
9 text of similar length reduces recoverability by 0.136; closer abstract matching
10 retains a positive effect (0.121, CI [0.096, 0.151]). Length-matched effects also
11 occur on AI Idea ($n=3,495$), FIRE ($n=35$), and a ResearchBench pilot ($n=12$).
12 Selecting sentences from fuller context without seeing the reference conclusion re-
13 covers 39% of the coverage gained by revealing methods-stage text rather than the
14 topic alone; selection guided by that conclusion recovers 51%. A two-reader pilot
15 supports coarse relative recoverability. Model scores remain positively associated
16 with recoverability under a different metric family, but aggregate ranking changes
17 are uncertain when models share the same items. FIRE’s historical conclusions
18 rank highly among other tasks’ conclusions, but this does not establish scientific
19 uniqueness. A mathematical example and a negated-sentence control show why
20 retrieval and overlap do not establish correctness. Our measurement protocol sepa-
21 rates these observable diagnostics from the independent criterion needed to judge
22 which scientific conclusions the evidence supports.

23 1 Introduction

24 Scientific-discovery benchmarks increasingly ask AI systems to propose findings and score their
25 agreement with a published paper’s conclusion, which we call the historical endpoint. Systems
26 receive literature context or progressively larger portions of the source paper (Lew et al., 2026; Liu et
27 al., 2026b; Guo et al., 2024; Qiu et al., 2025). If that text already summarizes the finding, a high score
28 may reward reconstruction rather than hypothesis formation. If several conclusions are defensible,
29 agreement with just one historical answer may miss valid alternatives. This matters when benchmarks
30 guide AI-system selection for scientific research. We use *measurement contract* for the relationship
31 between the supplied input, reference endpoint, scoring rule, and scientific interpretation assigned to
32 the score.

33 The central empirical question is how much historical-endpoint agreement can be explained by the
34 evidence supplied to the system. We measure *availability*: how much of the historical endpoint can
35 be recovered from the input using text similarity. We progressively reveal the topic, abstract and body
36 text, then test whether replacing that text with another paper’s evidence reduces recoverability. We
37 examine how this quantity relates to system scores. High availability identifies an opportunity for

reconstruction; it does not reveal which strategy a system used (Gururangan et al., 2018; Feng et al., 2019; Pacchiardi et al., 2024).

A second question requires different evidence: is the historical endpoint the only scientifically admissible answer? Retrieving it ahead of other papers’ conclusions does not answer that question. We report candidate discrimination as a conditional diagnostic, with explicit pool construction and stress tests. This distinction gives a practical order of operations: characterize the input–reference relationship before interpreting system scores, and establish a scientific criterion before treating retrieval rank as endpoint uniqueness.

Contributions. (1) An audit of historical-endpoint scoring through disclosure, paired evidence swaps, uncertainty estimates and blinded reader judgments; (2) controls using more similar replacement papers and sentence selectors with different access to the reference conclusion; (3) comparisons of model rankings on shared items and retrieval under different candidate pools. These tests bound what endpoint scores establish. The protocol in §6 links each diagnostic to its supported decision.

2 Related Work

Endpoint-scored discovery and ideation benchmarks include ResearchBench, ProjectionBench, AI Idea Bench, FIRE-Bench, and SDABench (Liu et al., 2026b; Lew et al., 2026; Qiu et al., 2025; Wang et al., 2026a; Shi et al., 2026). Related systems build or evaluate open hypotheses, while science-agent meta-evaluation probes what aggregate scores omit (Jansen et al., 2024; Majumder et al., 2024; Zhong et al., 2026; Nadgir et al., 2026); our object is instead the measurement contract connecting supplied evidence to one historical endpoint.

Our approach follows construct validation and analyses of how benchmark data shape score meaning (Messick, 1995; Bowman and Dahl, 2021; Bean et al., 2025; Kovatchev and Lease, 2024) and tests of shortcut-accessible or partial inputs (Gururangan et al., 2018; Feng et al., 2019; Pacchiardi et al., 2024). We combine disclosure dose, evidence interventions (Shao, 2026), extraction controls and cross-benchmark consistency (Liu et al., 2026a; Zhang et al., 2026), ranking consequences (Panthi and Abdelfattah, 2026), and the limits of candidate discrimination. Broad benchmark surveys and automated audits catalogue failure modes and mitigations (Li et al., 2026; Wang et al., 2026b; Tu et al., 2026; McGregor et al., 2025); our analysis estimates per-benchmark endpoint quantities with uncertainty. We use tests with specified expected responses (Ribeiro et al., 2020), controlled-truth diagnostics (Dekoninck et al., 2024), and comparisons that distinguish information access (Huang et al., 2024). The contribution is their application to scientific endpoints. Unlike a leaderboard sensitivity study (Alzahrani et al., 2024), our compact panel leaves aggregate ranking changes uncertain.

3 Measurement Framework

Availability. We first compare the supplied text with the historical conclusion, before evaluating any model. Let x be the input and y^* the reference conclusion (the gold endpoint). A model produces $\hat{y}$, scored by $s(\hat{y}, y^*)$. Our availability measure averages four text-similarity scores:

$$R_{\text{avail}}(x, y^*) = \frac{1}{|\mathcal{M}|} \sum_{m \in \mathcal{M}} m(x, y^*), \quad (1)$$

Here $\mathcal{M}$ contains lexical coverage (the fraction of distinct endpoint content tokens found in x), Jaccard token-set overlap, ROUGE-L sequence overlap, and bag-of-words (BoW) cosine between token-count vectors. Main-table R denotes this mean; it measures textual recoverability, not truth. The model-output score s_{lex} uses lexical coverage alone: the fraction of distinct endpoint content tokens found in $\hat{y}$. We also compare frozen-encoder representations $e(x)$ and $e(y^*)$ using $R_{\text{embed}} = \cos(e(x), e(y^*))$. Lexical R with lexical outcome score s_{lex} is *same-family*; R_{embed} with s_{lex} is *scorer-disjoint*. This reduces direct metric reuse but retains a shared gold endpoint and semantic difficulty; neither association identifies a causal mechanism, and the same-family value is not a mathematical upper bound. To test sensitivity to the supplied evidence, we replace each input with the same field from another paper (a donor) of similar length. Every donor is used once and no paper donates to itself; this permutation is called a derangement. Most donors are within 10% of source length. The paired loss is $\Delta_{\text{evid}} = R(x_{\text{orig}}, y^*) - R(x_{\text{intervened}}, y^*)$.

Candidate discrimination and scientific admissibility. We ask whether the input distinguishes its historical conclusion from other candidates in a declared pool $\{y_j\}_{j=1}^C$. We rank them by BoW cosine to x . The gold endpoint ranked first has reciprocal rank 1; one ranked second has $1/2$. Their mean is the mean reciprocal rank (R_{id} , MRR). We also report gold@1, the fraction ranked first, and the similarity margin over the best distractor. These quantities measure *retrieval within the specified pool*, not scientific identifiability. FIRE uses all 35 retained task conclusions as candidates for each query, including its own gold. Other tasks’ conclusions are distractors, not adjudicated alternatives to that task. ResearchBench uses each ranking item’s released hypothesis pool (Appendix C).

A simple endpoint example shows the missing assumption. If the evidence is $f(0)=0$, $f(1)=1$, both $f(t)=t$ and $g(t)=t + t(t-1)$ fit exactly, but predict 2 and 4 at $t=2$. An affine model contract admits only the first; a quadratic contract admits both. A pool omitting g can rank the historical f first under either interpretation. Thus high rank cannot certify uniqueness without a valid and sufficiently complete alternative set. This is a logical boundary, not an estimate of ambiguity in FIRE.

Extraction and information access. To test recovery by selection rather than generation, we split the input at sentence-ending punctuation, keep segments longer than 40 characters, and select at most eight. The gold-assisted diagnostic ranks sentences by BoW cosine to y^* ; an input-only baseline ranks them by cosine to the entire input; a lead baseline takes the first eight. Stable ties follow input order; if no segment qualifies, use the first 1,500 characters. The gold is used only for evaluation of the latter two selectors. An echo baseline returns the supplied text unchanged. Let S denote mean lexical endpoint coverage, not the four-measure mean R . To express extraction relative to the coverage gained from topic-only to methods-stage (S3) input, define $\text{Closure} = (S_{\text{extract}} - S_{\text{topic echo}})/(S_{\text{S3 echo}} - S_{\text{topic echo}})$. Closure is the fraction of that disclosure gap recovered by selected sentences, not the fraction of a model-performance gap explained. S3 and S4 extraction are reported separately; their output lengths are not matched to those of the model panel.

4 Experimental Protocol

Populations. Our ProjectionBench-style disclosure proxy is an author-constructed convenience pool of 2023–2026 arXiv CS/AI-for-science papers: 97 collected, 11 without full ar5iv HTML excluded, $N=86$ retained. It substitutes for unavailable official ProjectionBench materials. The disclosure stages are topic-only (S0), abstract (S1), early body (S2), middle-body text (S3, the methods-stage proxy), and fuller context (S4). Targets are the last conclusion-facing body segment. Automated extraction checks find 7/86 hard failures, retained as flags; excluding them preserves the S3 effect ($\Delta=0.142$, CI $[0.122, 0.165]$). Sampling, IDs, and extraction taxonomy are in Appendices A, D, and K.

We repeat the evidence-swap test on the full public AI Idea target set ($n=3,495$) (Qiu et al., 2025), the FIRE verified-rediscovery suite ($n=35$) (Wang et al., 2026a), and a small public ResearchBench split ($n=12$) (Liu et al., 2026b). FIRE supplies the principal candidate pool for joint availability–candidate-discrimination analysis; ResearchBench contributes a small cross-benchmark pilot.

Inference and model panel. We resample paired item results to obtain bootstrap confidence intervals (CIs), using 5,000 replicates; sign-flip tests use 5,000 draws with plus-one correction. The five confirmatory tests (disclosure dose, three identity interventions, and the hard-negative control) all survive Holm correction (Holm, 1979) at $\alpha=0.05$ (all $p < .001$). We also report item-cluster intervals, stage-stratified associations, leave-one-out stability, and changes in the highest-scoring model. The compact panel uses Gemma-3-4B, Llama-3.2-3B, Qwen-2.5-7B, and GPT-4o-mini at temperature 0.2 on 228 model–item–stage combinations (cells). Models cover different item subsets because of the available budget. Outcome scores use deterministic lexical coverage, not an LLM judge. Full seeds, assignments, controls, and replay paths appear in Appendix A. The closer-donor, candidate-pool, shared-item ranking and extraction analyses are retrospective checks selected after review. They use seed 20260928 and 5,000 item-bootstrap draws. Their intervals condition on the retained corpus, donor assignment and model outputs.

These populations and compact instruction models do not establish the behavior of research agents or official leaderboards.

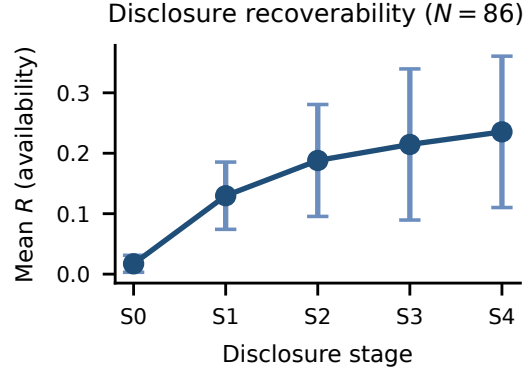


Figure 1: Historical-target recoverability as more source text is revealed ($N=86$). Points show mean lexical R ; bars show standard deviation across items. The mean rises from topic-only (S0) to fuller context (S4).

5 Results

5.1 Recoverability rises with progressive disclosure

We hold each historical target fixed and reveal progressively more source text. Figure 1 shows mean lexical R on the $N=86$ proxy rising $0.017 \rightarrow 0.130 \rightarrow 0.188 \rightarrow 0.215 \rightarrow 0.235$ from S0 to S4 ($\Delta=+0.218$, CI [0.195, 0.246]; 100% of items have higher S4 than S0 recoverability). The largest single step is topic $\rightarrow$ abstract ($0.017 \rightarrow 0.130$); later stages still add coverage beyond the abstract alone. As a negative control, random other-item targets at S4 score $R=0.081$ (gap $+0.154$), separating endpoint-specific recovery from generic long-text overlap. At S1, the true target also beats the most BoW-similar other target (the hard negative) for 66/86 items (77%). Worked high/low examples appear in Appendix N.

5.2 Evidence swaps reduce recoverability even with closer donors

Replacing S3 text with another paper’s same-role text of similar length reduces lexical R by 0.136 (paired-bootstrap 95% CI [0.118, 0.158]; 84/86 items decrease). It approximately matches length and preserves field role, but changes topic and content, so the effect alone cannot isolate evidence identity from relevance. A replay on the retained text gives 0.1356 (CI [0.1165, 0.1574]; 85/86 decrease), rather than the stored 0.1362. The original input version is unavailable; all comparisons below use the same retained text.

To test whether more similar replacement papers still reduce recoverability, we maximize abstract similarity over a one-to-one assignment with no self-pairs. Allowed donors have S3 length within 20% of the recipient; original length-nearest pairs remain allowed to ensure a complete assignment. Matching uses neither the target nor the outcome. Cached abstract embeddings raise mean donor cosine from 0.452 to 0.646, while the swap effect remains 0.121 (CI [0.096, 0.151]; 81/86 decrease). All assigned lengths satisfy the 20% constraint. Matching current abstracts by term frequency-inverse document frequency (TF-IDF) instead gives 0.127 (CI [0.102, 0.156]; 84/86 decrease). This weights uncommon terms more heavily. It also avoids relying solely on cached embeddings whose original input strings cannot be verified.

Twenty assignments with perturbed abstract-similarity scores give mean effects 0.129–0.139. The optimal semantic assignment attenuates the length-only effect by 0.015, but its paired CI [−0.008, 0.033] includes zero. The evidence supports a drop under these closer swaps, not an identity effect beyond exact entity and semantic-relevance matching. Restricting input to the topic lowers R by 0.197 versus S3; removing all gold content tokens yields zero overlap by construction. Excluding the seven extraction failures preserves the original effect (0.142, CI [0.122, 0.165]). Correlations among lexical measures are 0.67–0.97; omitting each lexical measure in turn preserves stage order (rank correlation $\rho \geq 0.99$).

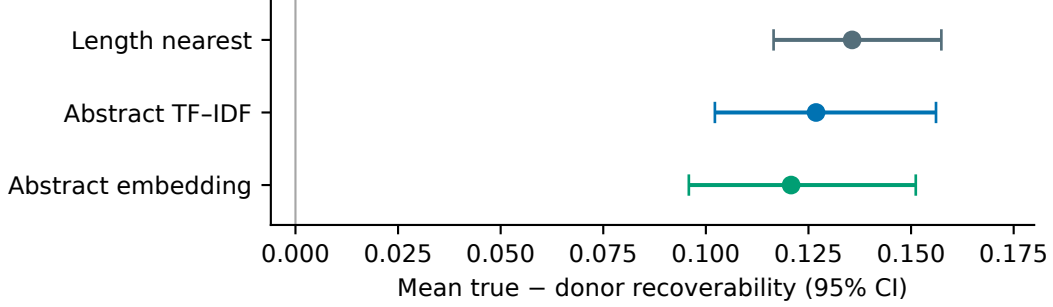


Figure 2: Paired S3 recoverability losses on the same retained 86 items. Bars are item-bootstrap 95% CIs, conditional on each donor assignment. Closer abstract matching retains a positive effect. The 20 perturbed semantic assignments measure assignment sensitivity, not 20 independent populations.

Table 1: Sentence extraction on the same 86 items. Closure is the fraction of the topic-only to S3 full-text coverage gap (0.266) recovered by extraction. CIs resample paired items. Gold access affects selection, not just scoring; output lengths are not matched to model outputs.

Input	Selection information	Mean coverage	Closure [95% CI]
S3	Input centrality	0.109	0.384 [0.356,0.410]
S4	First eight sentences	0.110	0.388 [0.342,0.435]
S4	Input centrality	0.112	0.394 [0.352,0.434]
S4	Gold-assisted	0.142	0.508 [0.471,0.544]

5.3 Embedding similarity and sentence extraction

Lexical R can miss semantic paraphrase of the endpoint. On the same $N=86$ items, mean R_{embed} jumps from topic-only (0.239) to abstract-level disclosure (0.682), then ranges from 0.745 to 0.753 across S2–S4; true targets beat random at S3 for 95% of items. Broad alignment thus emerges early. The later plateau concerns the frozen encoder and its 6,000-character window, not full recovery of scientific meaning; lexical coverage continues to rise (Fig. 1). Lexical and embedding R agree only moderately ($r=0.32$), so we retain them as separate measures. At S3, embedding recoverability and lexical outcome score have $r=0.65$; the pooled same-family association is 0.75. Their different populations and shared endpoint preclude interpreting the difference as a causal correction.

We next ask how much endpoint text can be recovered by simply selecting input sentences (Table 1). Gold-assisted S4 selection had historical closure 0.508; replay on retained text gives 0.508 (CI [0.471, 0.544]). Selecting S4 sentences most similar to the input as a whole (input centrality) closes 0.394 (CI [0.352, 0.434]), and taking the first eight sentences closes 0.388. Gold access increases mean lexical coverage by 0.030 (CI [0.022, 0.039]). Thus substantial overlap is accessible without access to the gold during selection; gold assistance gives a larger result. Centrality is not shown superior to the simpler lead baseline.

5.4 Scores track recoverability; aggregate rankings remain uncertain

Figure 3 plots system endpoint score s_{lex} against input R for 228 panel cells (four models $\times$ staged inputs). Higher recoverability goes with higher scores on average: pooled $r=0.749$ (CI [0.676, 0.830]), driven by mid/high disclosure (S1 $r=0.736$; S3 $r=0.600$); at S0 the association is negative, as expected when little endpoint text is available and scores are near floor. The stage split matters: pooled r mixes a hard floor with informative mid/high stages, so we report stratified correlations as well as the pool.

The score association persists after adjusting for stage, model, publication year and text lengths: $\beta_R=0.247$ (CI [0.146, 0.409]) per unit increase in R . Comparing conditions within each paper, with stage, model and input-length controls, gives 0.179 [0.100, 0.311]; both are associations. An input-only extractive baseline also shows a positive score– R association ($r=0.837$ across item–stage cells); the incremental LLM contribution remains unidentified. Without research context, empty-input

Table 2: Reader agreement with the higher- R ordering, by the two items’ recoverability thirds (disclosure-proxy S3; metric values hidden). Overall bootstrap CIs are [0.617, 0.850] (A) and [0.567, 0.800] (B); agreement is 55/60 (91.7%; $\kappa=0.83$). This pilot supports coarse relative ordering only.

Pair groups	n pairs	Reader A	Reader B
LOW-HIGH	30	83.3%	80.0%
LOW-MID	10	80.0%	60.0%
MID-HIGH	20	55.0%	55.0%
All	60	73.3%	68.3%

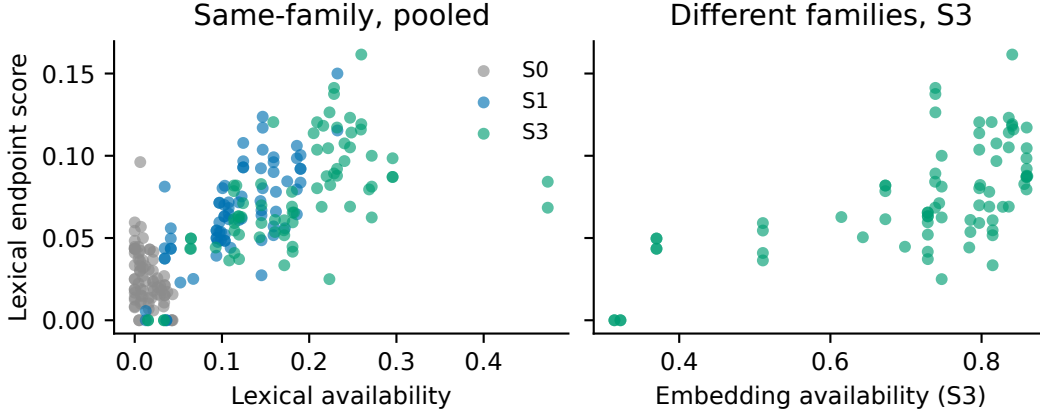


Figure 3: Endpoint score versus lexical availability (left, 228 retained cells) and embedding availability (right, S3 only). Each mark is a model–item–stage cell. The left panel spans disclosure stages; the right holds S3 fixed. Repeated items and stage differences preclude causal interpretation; clustered and adjusted estimates are reported in text.

199 $s_{\text{lex}}=0.031$ (near S0 0.026) and title-only 0.015 (vs panel S3 0.074; App. O). Grouping items into
200 thirds by S3 R , mean panel s_{lex} rises from low to high recoverability (App. E, Fig. 5).

201 We next ask whether disclosure changes which model scores highest. The historical 13/20 per-item
202 winner changes use an unequal model panel and resolve exact ties by row order. Counting a change
203 only when no highest-scoring model is shared between stages gives 12/20. Neither statistic is an
204 aggregate leaderboard reversal. On the 12 items shared by all four models at S0 and S3, the highest
205 mean changes from Gemma (0.036) to Qwen (0.080). Yet the top-minus-runner-up paired intervals
206 include zero: 0.006 [−0.003, 0.016] at S0 and 0.003 [−0.012, 0.019] at S3. Only two items have
207 unique winners separated by more than 0.01 at both stages; both flip. On the 20 common GPT/Qwen
208 items, Qwen has the higher mean at both stages. Thus the retained outputs warrant uncertainty about
209 model choice, not a stable reversal or a scientifically better selection. Repeated-generation variability
210 is unmeasured (Appendix C).

211 5.5 Readers agree with coarse recoverability differences

212 Using the S3 comparison interface, two readers each judged all 60 pairs while R and the analysis
213 key were hidden. Each pair asks which evidence side makes the historical endpoint *more recon-*
214 *structible*, a relative evidence judgment distinct from system-output grading. Reader A chose the
215 continuous higher- R side on 44/60 (73.3%, bootstrap CI [0.617, 0.850]); reader B on 41/60 (68.3%,
216 CI [0.567, 0.800]). They agreed on 55/60 pairs (91.7%; Cohen’s $\kappa=0.83$). Table 2 groups pairs by
217 their items’ low, middle or high S3 recoverability thirds. Agreement with the metric is 80–83% for
218 LOW-HIGH pairs but near chance (55%) for MID-HIGH. The pilot supports coarse relative ordering
219 on this proxy; fine discrimination and absolute correctness remain unvalidated.

Table 3: Original versus swapped evidence under the same four-measure R_{avail} (Eq. 1). Donors have the same field role and similar length; topic and relevance can also change. RB = ResearchBench; RQ = research question; refs = references; insp. = inspirations. Decrease = share with $\Delta > 0$. CIs are item-paired bootstrap 95% intervals. Permutation $p < 10^{-3}$ for disclosure, FIRE, and RB; the AI Idea CI excludes 0.

Setting	Original→swapped R	n	Δ	95% CI	Decrease
Disc. proxy	S3 0.215→0.078	86	0.136	[0.118, 0.158]	98%
AI Idea	topic+refs 0.233→0.168	3495	0.065	[0.063, 0.066]	94%
RB pilot	RQ+insp. 0.247→0.135	12	0.112	[0.095, 0.129]	100%
FIRE	instr. 0.252→0.065	35	0.187	[0.165, 0.209]	100%

Table 4: FIRE availability and retrieval among 35 historical conclusions (Wilson intervals for gold@1). R_{avail} measures input–gold recoverability; gold@1 and MRR summarize the gold endpoint’s rank among candidates. Disclosure/AI Idea have no candidate pool; the ResearchBench pilot ($n=12$) is appendix-only (App. E).

Input	R_{avail}	MRR	gold@1 count [95% CI]	n
Instruction	0.252	0.93	31/35 [0.74, 0.95]	35
Research question	0.522	0.98	29/30 [0.83, 0.99]	30

5.6 Evidence swaps reduce recoverability across benchmarks

Table 3 tests whether the decrease extends beyond the disclosure proxy. It reports the original estimates using the same four-measure R_{avail} but different evidence fields in each benchmark. Replacing evidence with similar-length text lowers recoverability in every setting; the loss is largest for FIRE instructions and the disclosure proxy, and smaller but precisely estimated for AI Idea ($n=3495$).

AI Idea and ResearchBench evidence fields. On AI Idea, length-nearest reference derangement drops topic+refs R by 0.065 (CI [0.063, 0.066]; 94.2% decrease): which references are shown matters under the implemented length/role control; topic and relevance also change. The topic fields already contain language shared with the gold. Swap losses Δ are positive in every keyword group (NLP/vision/ theory/science; full table App. I). On the ResearchBench pilot ($n=12$), replacing the supplied evidence also lowers R ($\Delta=0.112$ CI[0.095, 0.129]; decrease on 12/12), with inference scoped to this small split; dual-axis RB numbers remain in App. E.

FIRE retrieves historical conclusions; admissibility remains separate. The candidate pool contains 35 conclusions from different FIRE tasks. Instruction queries put the gold first on 31/35 tasks (MRR 0.928); 30 nonempty research questions yield 29/30 (MRR 0.983). These are discrimination results within a heterogeneous pool, not expert judgments that competing scientific answers are invalid. Keeping the query and scorer fixed, we retain the gold and sample four distractors. Across 200 such pools per task, instruction gold@1 is 97.0%. Choosing the four highest-scoring distractors instead gives 88.6%, the same as the complete pool. Pool construction changes apparent difficulty. Lexical coverage instead of BoW puts the gold first on 32/35 instructions and 30/30 questions. This scorer variation does not validate either metric.

There is also a direct correctness failure: our tokenizer removes both “does” and “not”, so “Treatment does improve survival” and “Treatment does not improve survival” have identical content tokens, BoW cosine 1, and lexical $R=1$. This constructed polarity contrast is a falsifier of using overlap as scientific truth, not a measured FIRE error rate. Candidate retrieval and availability therefore remain diagnostics to accompany a separate admissibility criterion. FIRE’s score association is strong within the same family (instruction $r=0.64$; question $r=0.76$), but scorer-disjoint instruction association is weak (0.17, not significant; question 0.54). This negative result limits transport of the disclosure-panel association.

6 Discussion

Implications for scientific-discovery evaluation. We quantify how historical-endpoint overlap responds to disclosure and evidence replacement, complementing reviews of benchmark practices (Bean et al., 2025). Extending evidence-dependence analyses (Liu et al., 2026a; Shao, 2026), we distinguish historical-answer retrieval from validation of scientifically admissible alternatives. These diagnostics identify reconstruction opportunities before scores are attributed to scientific synthesis; they imply neither autonomous discovery nor universal benchmark invalidity. Whether they improve scientific system selection requires an independent outcome criterion.

Measurement protocol. Declare evidence, historical target, scorer and intended interpretation. Pair availability measures with input-only baselines; report gold-assisted diagnostics separately. Swap evidence under explicit length, topic and role controls, reporting achieved matching, paired intervals and diagnostic failures. For candidate retrieval, publish pool provenance, size, hardness and ties. To claim unique scientific validity, obtain a separate admissibility criterion that can recognize valid alternatives. Finally, compare systems on common items and information budgets, separating per-item changes from aggregate uncertainty. High-availability cases can remain as labelled reconstruction controls; low retrieval rank should trigger inspection, not automatic rejection of the endpoint contract.

7 Conclusion

Across four settings, replacing supplied evidence lowers historical-endpoint recoverability; closer matching in the disclosure proxy retains the effect. Input-only sentence extraction recovers about 39% of the specified disclosure gap. These findings support testing availability before interpreting endpoint agreement as scientific synthesis. Gold-assisted extraction has additional information, model-ranking changes remain uncertain, and historical-answer retrieval does not certify scientific uniqueness. Scientific-discovery evaluation therefore needs both inspectable endpoint diagnostics and an independent criterion for which conclusions the evidence supports.

Limitations

The disclosure proxy is a convenience pool of 86 public CS/AI-for-science papers (97 collected; 11 unavailable), so its results are pool-specific. Holding observed pairs fixed, the mean swap loss remains positive even if every excluded item favors swapped evidence by the maximum possible amount. ResearchBench is descriptive ($n=12$), FIRE-Bench is the main candidate pool ($n=35$), and model-score associations cover the disclosure proxy and FIRE. The blinded 60-pair reader pilot supports only coarse relative S3 ordering: MID-HIGH remains near chance, and absolute, cross-population, and scientific validity remain open. No expert-adjudicated alternative endpoints or external protocol usability study are available.

The primary R family is lexical. Embedding recoverability and leave-one-measure-out stability reduce, but do not remove, semantic misses or scorer overlap; embedding inputs truncate at 6,000 characters. All score-versus- R results are associational: controls do not identify causal shortcut use. Candidate pools omit unknown valid alternatives; closer donors do not exactly match entities or relevance. Winner changes are conditional on one retained output per cell. Seven of 86 extraction hard fails are retained, and low input floors cannot exclude memorisation.

Shipped ledgers support offline inspection (App. A), but exact API-panel replay is unavailable because full responses and provider routing metadata were not retained. The disclosure ledger has a small exact-replay discrepancy; new checks retain separate input hashes and results (Appendix C).

Ethical Considerations

Public scholarly texts are used under their licenses; gated corpora are not redistributed. Two readers provided 120 blinded judgments of relative reconstructibility, not independent external validation or scientific correctness. Pairwise records omit names and demographics; reader-background and governance metadata were not retained. OpenAI Codex and Anthropic Claude assisted coding, editing and verification, not annotation; authors remain responsible.

References

- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *NAACL*.
- Shi Feng, Eric Wallace, and Jordan Boyd-Graber. 2019. Misleading failures of partial-input baselines. In *ACL*.
- Lorenzo Pacchiardi, Marko Tešić, Lucy G. Cheke, and José Hernández-Orallo. 2024. Leaving the barn door open for Clever Hans: Simple features predict LLM benchmark answers. arXiv:2410.11672.
- Yanhang Li, Zhichao Fan, and Zexin Zhuang. 2026. Auditing the audit: Five failure modes in benchmark-validity audits. arXiv:2607.02586.
- Xinming Tu, Tianze Wang, Yingzhou Lu, Kexin Huang, Yuanhao Qu, and Sara Mostafavi. 2026. BenchGuard: Who guards the benchmarks? Automated auditing of LLM agent benchmarks. arXiv:2604.24955.
- Sean McGregor, Vassil Tashev, Armstrong Foundjem, Aishwarya Ramasethu, Sadegh AlMahdi Kazemi Zarkouei, Chris Knotz, Kongtao Chen, Alicia Parrish, Anka Reuel-Lamparth, and Heather Frase. 2025. Risk management for mitigating benchmark failure modes: BenchRisk. In *NeurIPS Datasets and Benchmarks Track*.
- Yao Liu, Tien-Ping Tan, and Zhilan Liu. 2026a. Charge as a construct-validity factor in Chinese legal case retrieval: A cross-benchmark audit. arXiv:2606.12993.
- Kan Shao. 2026a. Metadata predictability is not evidence dependence: An intervention-based audit for weak-label benchmarks. arXiv:2605.23701.
- Sugam Panthi and Rabab Abdelfattah. 2026. Same ranking, different winner: How scoring targets shape LLM memory benchmarks. arXiv:2605.24060.
- Jiaqi Shao, Hanck Chen, Wei Zhang, Maxm Pan, and Bing Luo. 2026b. Do agent benchmarks measure capability? Protocol validity in the age of agentic AI. arXiv:2607.22368.
- Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2026b. ResearchBench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 13187–13207. doi:10.18653/v1/2026.findings-acl.644.
- Andrew J. Lew, Yuan Cao, and Markus J. Buehler. 2026. ProjectionBench: Evaluating scientific hypothesis generation in LLMs under progressive information disclosure. arXiv:2605.30284.
- Sikun Guo et al. 2024. IdeaBench: Benchmarking large language models for research idea generation. arXiv:2411.02429.
- Yansheng Qiu et al. 2025. AI Idea Bench 2025: AI research idea generation benchmark. arXiv:2504.14191.
- Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, Zhiting Hu, and Eric P. Xing. 2026a. FIRE-Bench: Evaluating AI agents on the rediscovery of scientific insights. arXiv:2602.02905.
- Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *ACL*.
- Oscar Sainz et al. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of EMNLP*.
- Samuel Messick. 1995. Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9).
- Samuel R. Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *NAACL*.

- 344 Andrew M. Bean, Ryan Othniel Kearns, et al. 2025. Measuring what matters: Construct validity in
345 large language model benchmarks. In *NeurIPS Datasets and Benchmarks Track*.
- 346 Yuzhe Zhang, Chihui Chen, and Lina Yao. 2026. Consistency evaluation of benchmarks used for
347 causal discovery. arXiv:2606.01789.
- 348 Nitya Nadgir et al. 2026. Life after benchmark saturation: A case study of CORE-Bench.
349 arXiv:2606.26158.
- 350 Peter Jansen et al. 2024. DiscoveryWorld: A virtual environment for developing and evaluating
351 automated scientific discovery agents. In *NeurIPS*.
- 352 Bodhisattwa Prasad Majumder et al. 2024. DiscoveryBench: Towards data-driven discovery with
353 large language models. arXiv:2407.01725.
- 354 Sture Holm. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of*
355 *Statistics*, 6(2).
- 356 Venelin Kovatchev and Matthew Lease. 2024. Benchmark transparency: Measuring the impact of
357 data on evaluation. In *NAACL*.
- 358 Tianyun Zhong, Wangyi Jiang, Wei Wang, Xuanang Chen, Yaojie Lu, Shiwei Ye, Yuzhen Shi, Boyu
359 Yang, Jinghang Wang, Han Li, Weiqi Zhai, Bing Zhao, Hu Wei, Haiyang Yu, Yongbin Li, Hongyu
360 Lin, Le Sun, and Xianpei Han. 2026. Before the action: Benchmarking LLMs on prospective
361 hypothesis discovery. arXiv:2607.15766.
- 362 Chuhan Shi, Xiaoquan Ren, Sicheng Song, Haobo Li, Rui Sheng, and Yushi Sun. 2026.
363 Are LLMs ready for scientific discovery? A capability-oriented benchmark for AI scientists.
364 arXiv:2607.11079.
- 365 Junlin Wang, Federico Bianchi, Shang Zhu, Fan Nie, Yongchan Kwon, Bhuwan Dhingra, and
366 James Zou. 2026b. Automated benchmark auditing for AI agents and large language models.
367 arXiv:2605.26079.
- 368 Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song,
369 and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *ICLR*.
- 370 Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy:
371 Behavioral testing of NLP models with CheckList. In *ACL*.
- 372 Jasper Dekoninck, Mark Niklas Müller, and Martin Vechev. 2024. ConStat: Performance-based
373 contamination detection in large language models. In *NeurIPS*.
- 374 Norah A. Alzahrani, Hisham Abdullah Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Z.
375 Alsubaie, Yousef Almushayqih, Faisal Abdulrahman Mirza, Nouf M. Alotaibi, Nora Al-Twairish,
376 Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing
377 the sensitivity of large language model leaderboards. In *ACL*.

378 A Reproducibility

379 Historical code and metrics: `src/recoverability/`, `scripts/`. Evidence ledgers:
380 `results/`; claim-central authorities include `canonical_disclosure_shuffle.json`,
381 `formal_significance.json`, `panel_ranking_stability.json`,
382 `dual_axis_rid_ravail.json`, and the oracle, `scorer-disjoint`, `falsifier`, `construction`, `ex-`
383 `traction`, `criterion-proxy`, and failure ledgers listed in the root SHA256SUMS manifest. Check-
384 sums: SHA256SUMS hashes the manuscript, figures, and all named evidence ledgers and
385 supplement-cited artifacts (verify with `shasum -a 256 -c SHA256SUMS`). Historical protocol
386 versions TRA-0.1–1.1; retrospective checks B254-v1. OpenRouter panel models (logged in
387 `results/openrouter_panel/` and `results/firebench_panel/`): `google/gemma-3-4b-it`,
388 `meta-llama/llama-3.2-3b-instruct`, `qwen/qwen-2.5-7b-instruct`,
389 `openai/gpt-4o-mini`; temperature 0.2 (prompts App. R). API caches retain requested
390 routes, prompts/settings, predictions, and vectors—not request IDs, returned model/routing metadata,

or full responses—so exact provider replay/version verification is unavailable. ResearchBench tiny is the public GitHub smoke split; full HF data remain gated. We do not redistribute gated corpora. Canonical disclosure identity-intervention bootstrap: seed 20260730, $n_{\text{boot}}=5000$, CI [0.118, 0.158] for mean $\Delta=0.136$ (matches abstract and §5); resampling is over *item-level* true—length-nearest-deranged differences. Re-running the documented pilot scripts on a clean copy does not exactly match the historical disclosure ledgers on retained inputs (App. C); new checks have their own input and script hashes. API panels remain provider-non-deterministic. The human instrument, anonymous returns, and scored ledger are included in the supplement for inspection.

Settings and seeds. Panel temperature 0.2; embedding encoder `text-embedding-3-small` with 6000-char truncation; disclosure identity intervention seed 20260730 ($n_{\text{boot}}=5000$); criterion instrument seed 20260731; extraction-validation seed 20260731. Prompts are fixed in App. R. API returns are non-deterministic at the provider, so panel point estimates should be read with the reported item-cluster CIs rather than as single-run truth.

Headline evidence. Primary identity $\Delta=0.136$ CI[0.118, 0.158] (`canonical_disclosure_shuffle.json`); $S0 \rightarrow S4$ +0.218 CI[0.195, 0.246]; pooled panel $r=0.749$ CI[0.676, 0.830] (`formal_significance.json`); historical gold-assisted $S4$ closure 0.508 CI[0.473, 0.544]; flips 13/20 CI[0.45, 0.85] (`panel_ranking_stability.json`); dual-axis MRR/gold@1 rows in `dual_axis_rid_ravail.json`; human A/B 73.3/68.3%, $\kappa=0.83$.

Null results. Several secondary associations are weak or compatible with noise and are reported as such: FIRE reverse-family correlations remain low (`firebench_panel/reverse_family.json`); $S0$ stage-stratified score $\times R$ is negative; system-side closure CI is wide and imprecise; embedding similarity changes little across $S2$ – $S4$ under the fixed input window.

B Formal Estimands and Intervention Contract

The main paper reports the operative lexical composite R_{avail} (Eq. 1). The complementary embedding measure is

$$R_{\text{embed}}(x, y^*) = \cos(e(x), e(y^*)), \quad (2)$$

where frozen encoder e is `text-embedding-3-small` (dimension 1536; inputs truncated at 6000 characters). Predictor and outcome are *scorer-disjoint* when they use different metric families: lexical R_{avail} with lexical s_{lex} is same-family, whereas R_{embed} with s_{lex} is scorer-disjoint. We therefore use the latter as a check against direct metric reuse. Both remain associations with a shared endpoint; neither identifies a mechanism or formally bounds the other.

For candidate pool $\{y_j\}_{j=1}^C$, candidate discrimination uses gold MRR, gold@1, and the contrastive margin

$$\text{margin}(x, y^*) = \text{sim}(x, y^*) - \max_{y \neq y^*} \text{sim}(x, y). \quad (3)$$

The evidence-identity estimand is the item-paired difference

$$\Delta_{\text{evid}} = R(x_{\text{orig}}, y^*) - R(x_{\text{intervened}}, y^*). \quad (4)$$

The confirmatory intervention is a deterministic one-to-one permutation of same-role donors with no self-pairs and a preserved donor multiset; donors are chosen by length proximity (most within 10% of source length). Random-target, hard-negative, topic-only, and full-mask controls test target specificity, partial-disclosure discrimination, topicality, and empty-input behavior.

C Retrospective Contract Checks and Exact Replay

These checks test donor similarity, candidate-pool difficulty, ranking stability and the information available to sentence selectors. They use the retained corpus and outputs with seed 20260928; all 5,000 bootstrap draws resample items, preserving their paired conditions. They are retrospective reanalyses selected after review, not independent confirmation. Scripts `run_b254_contract_checks.py` and `render_b254_figures.py` write complete item records, donor assignments, pool summaries, pairwise model contrasts and input hashes under `results/b254/`. No new API calls or human judgments were obtained.

Donor construction and uncertainty. We maximize the sum of recipient–donor abstract cosine under a one-to-one assignment, forbidding self-pairs and allowing S3 length ratios within 20% plus the original length-nearest edges. Every selected donor in the optimal semantic and TF–IDF assignments falls within 20%; the fallback edges do not weaken the achieved bound. The semantic assignment uses cached `text-embedding-3-small` vectors; the separate TF–IDF assignment uses current S1 tokens with raw term frequency and smoothed inverse document frequency $1 + \log((1 + N)/(1 + df))$, then cosine normalization. Mean abstract TF–IDF cosine rises from 0.044 to 0.140; mean S3 BoW cosine rises from 0.125 to 0.234. These measures establish closer text matching, not matched scientific relevance. Twenty additional assignments add independent Gumbel noise of scale 0.05 to semantic utilities before optimization. Their mean effects range 0.129–0.139; the optimized assignment is stronger matching and has the smaller effect 0.121. Intervals condition on each assignment; assignment variation is shown separately and is not another sample size.

Candidate pools and exact controls. FIRE uses 35 distinct retained conclusions with original stable-index tie handling; there are no gold-score ties under BoW in either query population. At pool sizes 5/10/20/35, the mean random-pool instruction gold@1 rates are 0.970/0.947/0.918/0.886 (200 draws per query and size). Hardest pools use query-to-candidate scores and are deliberately adversarial diagnostics, not an outcome-blind sampling scheme. BoW MRR is 0.930 with the hardest five and 0.928 with all 35. The ResearchBench ranking pilot instead has 16 candidates per item: its released gold plus fake and model-generated negatives. These are different candidate-generation policies, so absolute MRR values should not be read as a common scale of scientific uniqueness. For the main mathematical example, the family $f_c(t) = t + ct(t - 1)$ fits the two exact observations for every real c . An affine restriction forces $c = 0$; a degree-at-most-two restriction does not. The executable examples $c = -1, 0, 1$ yield predictions 0, 2, 4 at $t = 2$. No LLM judge or expert label supplies this elementary algebraic criterion. The polarity test separately demonstrates an exact tokenizer limitation; we preserve it as a failed validity test rather than modifying the metric after seeing the failure.

Common-item rankings. Table 5 reports common-item means. The original panel contains Gemma/Llama/Qwen/GPT outputs on 12/12/32/20 items respectively, each at S0, S1 and S3, giving 228 cells. The original 13/20 heterogeneous-panel statistic includes ties resolved by the first retained row; 12/20 have disjoint winner sets. Four-model S0 and S3 have respectively three and four tied items; only six items have unique winners at both stages, of which five change. Requiring margins greater than 0.005/0.01/0.02 at both stages leaves 3/2/0 eligible items and 3/2/0 changes. On the separate GPT/Qwen 20-item panel, 7/14 changes remain after excluding exact ties at either stage. These are descriptive sensitivity thresholds, not validated minimum scientific differences. All six four-model pairwise contrasts at each stage are retained, without selection for significance. Bootstrap intervals describe item variation, not provider or repeated-generation variability.

Table 5: Mean lexical endpoint coverage on common items. The four-model S3 winner differs from the unequal-panel mean ordering in the historical table.

Population / stage	Gemma	Llama	GPT-4o-mini	Qwen
12 common, S0	0.0363	0.0306	0.0215	0.0269
12 common, S3	0.0735	0.0769	0.0757	0.0802
20 GPT/Qwen, S0	—	—	0.0200	0.0268
20 GPT/Qwen, S3	—	—	0.0718	0.0732

Extraction access and replay differences. The historical 0.508373 closure comes from sorting S4 sentences by similarity to the gold, not input-only selection. With the retained corpus, the same procedure gives 0.508009. Input-only centrality gives 0.393561 and the lead baseline 0.387603. Mean output lengths are respectively 1870, 2293 and 1542 characters, so even these eight-sentence selectors do not have identical token budgets. A supplementary S3 comparison with retained model outputs uses the same input and lexical scorer, but different output lengths and an unverified old request-to-current-text binding; it cannot establish a matched system improvement.

Exact replay differs from the historical ledger. The retained corpus agrees across three release archives, yet current code gives disclosure swap loss 0.135637 rather than 0.136180, with 85 rather than 84 hurt items. The original input epoch cannot be reconstructed from these copies. Original

values remain labelled historical; all new contrasts use one hash-bound current corpus. Cached embeddings also lack per-text hashes, so exact binding to current strings cannot be certified. The reanalysis records input and script hashes and per-item outputs; its stopped-run log and amended protocol document the discrepancy. The small numeric change limits exact reproducibility without reversing the observed effect.

D Extraction-Fail and Low- R Taxonomy

Table 6: Extraction-fail and low- R taxonomy (disclosure proxy; fails retained in $N=86$; `failure_taxonomy.json`).

Fail / low- R mode	n	Definition
Prompt residue	3	instruction/prompt text, not endpoint prose
Bib./acks block	2	bibliography or acknowledgements block
Chrome-only	1	ar5iv footer/chrome, no usable body
Checklist form	1	checklist justification, not a conclusion
Low- R diffuse (S3 tertile)	23	long targets, low input overlap (not chrome)
Low- $R \cap$ hard-fail	5	of bottom S3 tertile ($n=28$)

E Availability and Candidate Discrimination

FIRE and ResearchBench differ in how their candidate endpoints are constructed. Table 7 adds the 12-item ResearchBench pilot to the 35 FIRE instruction queries and 30 research questions from Table 4; the pilot does not represent the full ResearchBench release. Figure 4 plots availability and gold retrieval for settings with candidate pools. The differing pools and metrics do not establish scientific uniqueness or validate opposite failure modes. Figure 5 shows disclosure-panel scores by S3 R_{avail} tertile.

Table 7: Dual-axis estimates including the RB tiny pilot (Wilson CIs in the main paper for FIRE). The RB row is descriptive and limited to $n=12$.

Setting	R_{avail}	R_{id}	gold@1	n
FIRE instr. (main)	0.252	MRR 0.93	31/35	35
FIRE RQ (main)	0.522	MRR 0.98	29/30	30
RB rank pilot (app. only)	0.279	MRR 0.45	4/12	12

F Responsible Research Summary

- **Risks:** Low; reproducible meta-evaluation of public benchmarks; no dual-use systems.
- **Data:** Public arXiv texts; AI Idea Bench, FIRE-Bench, ResearchBench tiny under their licenses; gated HF not redistributed.
- **Reader labels:** Two anonymous readers $\times 60$ blinded pairs on the disclosure proxy; relative-judgment scope only; labels and agreement in the supplement. Reader-background and governance metadata were not retained, so no independent/external-validation claim is made.
- **Compute:** API-only panels (disclosure 228 cells; FIRE 260 cells); model IDs and temperatures logged; no undisclosed training; no claim of full GPU-hour budget or provider-version replay.
- **Claims:** Ceilings stated; proxy/tiny splits labeled; reader-pilot evidence limited to relative S3 comparisons; model proxies remain distinguished; no universal invalidity claim.
- **Limitations:** Population, construct, scope, inference, criterion, contamination, and reproducibility threats are documented in the main paper.
- **Artifacts:** An anonymous software/data archive can accompany this workshop submission under the stated redistribution boundaries; a broader public release may follow after review.

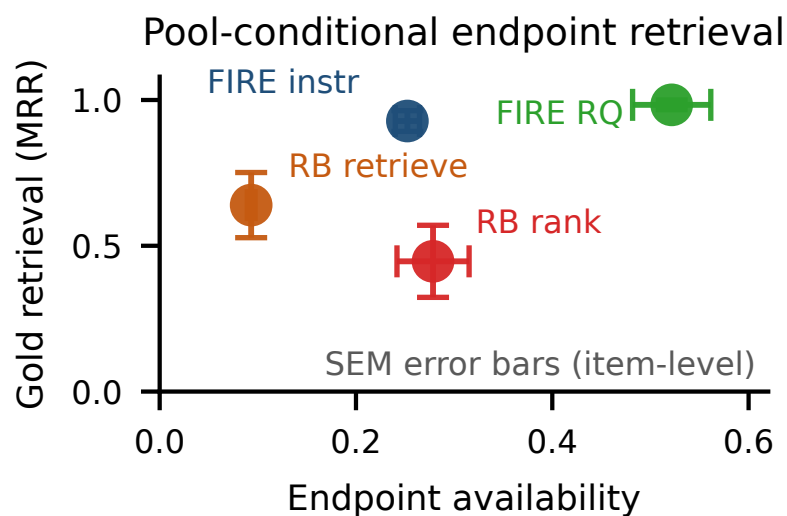


Figure 4: Input–endpoint availability (horizontal) and mean reciprocal rank within each candidate pool (vertical). Error bars show item-level standard errors. Different candidate pools preclude a common scientific-uniqueness scale.

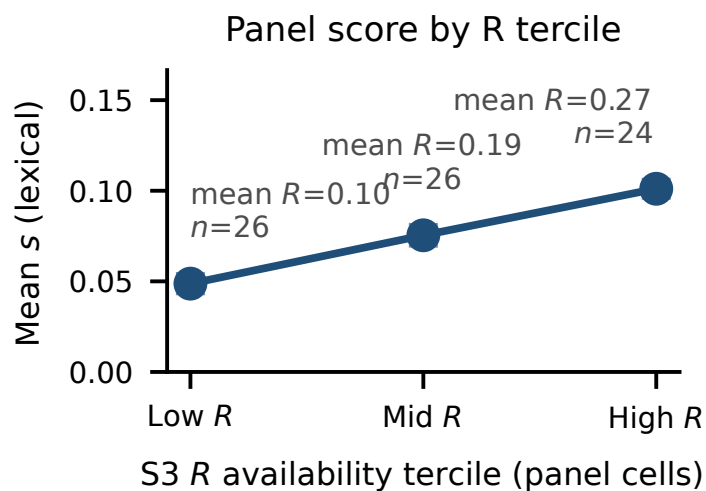


Figure 5: Mean panel s_{lex} by S3 R_{avail} tercile (SEM over panel cells).

512 G Additional Statistics

513 Ledgers (under results/): formal_significance.json, panel_ranking_stability.json,
514 cross_setting_transfer_scorecard.json, measure_agreement_S3.json.

Table 8: Cross-setting transfer estimates from frozen ledgers (F2=true>random, F3=true>hard; sc=no hard-neg pool; pass=control direction holds).

Setting	n	Δ_{shuf}	Hurt	F2 gap	F3 beat
Disc. proxy	86	0.136 [0.118, 0.158]	98%	0.154	77%
AI Idea	3495	0.065 [0.063, 0.066]	94%	0.172	98%
FIRE	35	0.187 [0.165, 0.209]	100%	pass	sc
RB tiny gen	12	0.112 [0.095, 0.129]	100%	0.225	92%

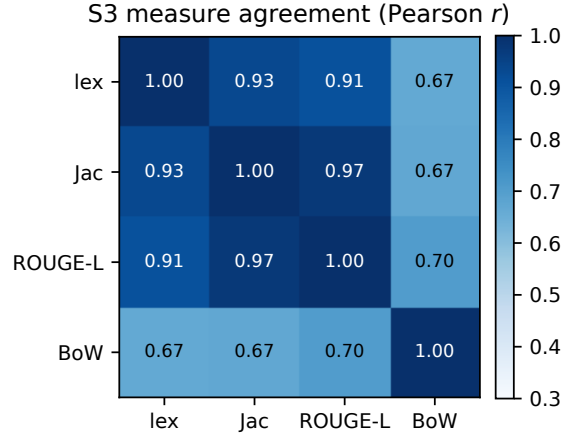


Figure 6: Convergent agreement among lexical R components at S3 ($N=86$; Pearson r ; ledger measure_agreement_S3.json). Discriminant contrast to R_{embed} is $r=0.32$ (main text).

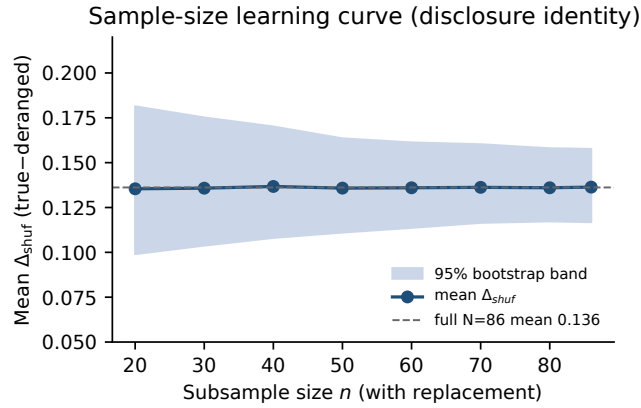


Figure 7: Sample-size learning curve for disclosure identity Δ_{shuf} (item bootstrap of n -subsamples; ledger sample_size_learning_curve.json). The band excludes zero for all $n \geq 20$.

515 RB conditional retrieval: gold@1 4/12 CI[0.14, 0.61], mean rank 6.75, margin -0.08 .

516 H Embedding Family Details

517 Encoder: OpenAI text-embedding-3-small via OpenRouter (dim. 1536). Texts truncated to 6000
518 characters before embedding; vectors cached under results/embedding_cache.jsonl. Summary

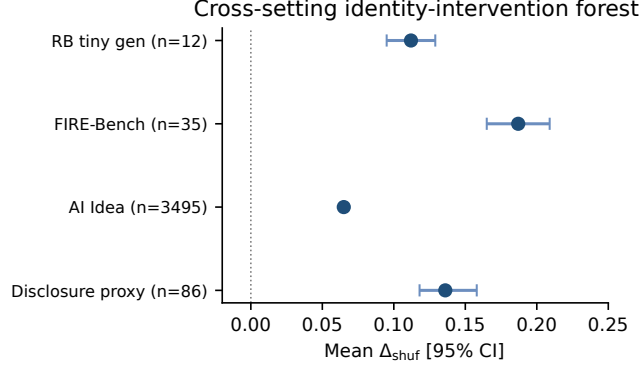


Figure 8: Forest of mean identity-intervention Δ_{shuf} across settings (CIs from frozen setting ledgers; same scorecard as Table 8).

Table 9: Contamination snapshot from closed-book outputs and publication-year counts. These diagnostics do not test training-set membership.

Contamination / leakage lens	Result	Note
Empty-input closed-book s_{lex}	0.031	vs panel S0 0.026
Title-only s_{lex}	0.015	vs panel S3 0.074
Proxy years 2023/24/25/26	2/33/45/6	$N=86$ convenience pool
Cutoff overlap $\Rightarrow$ leakage?	not entailed	training exposure unidentified

Table 10: Historical unequal-panel mean- s orderings (not common-item rankings; Ge/L/Q/GP = Gemma/Llama/Qwen/GPT-4o-mini; ledger panel_ranking_stability.json).

Contrast	Before order	After order	Item flips
S0→S3	Ge>L>Q>GP	L>Ge>Q>GP	13/20
S0→S1	Ge>L>Q>GP	L>Ge>GP>Q	top-1
FIRE cond.	cond.-dep. top-1	cond.-dep. top-1	19/30

Table 11: Gold-construction probe: share of gold tokens also in the named input field (topic or RQ). AI residual drops topic fields from the gold summary.

Setting	n	Constr. share	Topic/RQ→gold lex
AI Idea (full gold)	800	0.119	0.119
AI Idea (no topic fields)	800	—	0.083 residual
RB ranking	12	0.29	0.29
RB generation	12	0.32	0.32

Table 12: Disclosure panel correlations and flips (item-cluster CIs).

Quantity	Point	95% CI
Pooled $r(R, s)$	0.749	[0.676, 0.830]
S0 r	-0.405	[-0.603, -0.172]
S1 r	0.736	[0.557, 0.857]
S3 r	0.600	[0.372, 0.830]
Item-FE β_R	.179	[-.100, .311]
FE β_R (S1/S3)	.219/.174	[.131, .348]/[.086, .321]
Winner flips S0↔S3	0.65 (13/20)	[0.45, 0.85]
S0→S4 mean ΔR	0.218	[0.195, 0.246]

Table 13: Length-nearest evidence-identity derangement Δ by setting (all exclude 0 at $\alpha=0.05$).

Setting	mean Δ	95% CI / note
Disclosure (S3)	0.136	[0.118, 0.158]
FIRE instruction	0.187	[0.165, 0.209]
ResearchBench tiny	0.112	[0.095, 0.129]
AI Idea ($n=3495$)	0.065	[0.063, 0.066]

Table 14: Agreement among the lexical measures at S3: pairwise Pearson r . L/J/R/B denote lexical coverage, Jaccard, ROUGE-L and BoW cosine. Lexical versus embedding R has $r=0.32$ (§3); omitting one measure at a time preserves stage order ($\rho \geq 0.99$). These are consistency checks, not validation of scientific truth.

	J-R	L-J	L-R	B-R	B-J	B-L
r	0.97	0.93	0.91	0.71	0.67	0.67

Table 15: Panel coverage balance (mean S3 R_{avail} by models per item).

Assignment	Items	Mean S3 R
1-model	12	0.175
2-model	8	0.189
4-model	12	0.179

519 ledger: results/embedding_recoverability.json ($N=86$). Mean R_{embed} rises from about
520 0.68 at S1 to 0.75 at S2; S2–S4 differ by less than 0.008. This indicates early alignment under the
521 specified encoder and input window, not complete recovery of the conclusion’s meaning. Lexical
522 coverage continues to rise (Fig. 1).

523 I AI Idea Keyword Strata and FIRE-Bench

524 All seven keyword buckets sum to $n=3,495$ (results/parity_coverage_suite.json).

Table 16: AI Idea Bench keyword strata (full $n=3,495$; ledger parity_coverage_suite.json).

Stratum	n	Δ_{refs}	Δ_{shuf}	Hurt
NLP	1711	−0.013	0.057	92%
other	822	−0.002	0.074	95%
vision	742	−0.003	0.071	96%
RL	78	−0.002	0.067	99%
systems	72	−0.024	0.058	92%
theory	57	+0.007	0.093	100%
science	13	+0.031	0.113	100%

525 FIRE-Bench instruction→conclusion analysis: results/firebench_recoverability.json.
526 Panel residual accounting: results/parity_coverage_suite.json (panel_residual).

527 **ResearchBench tiny by discipline.** Shuffle Δ is positive in every discipline of the public tiny split
528 (Astronomy through Physics; $n=1$ each), ranging approximately 0.08–0.16 in mean $\Delta_{\text{derangement}}$
529 (full table in results/domain_rank_stability.json).

530 J Pathway Observability

531 ResearchBench is most process-decomposed yet still uses historical endpoints; high ob-
532 servability enables inspiration-identity tests but does not eliminate endpoint recoverability
533 (results/pathway_observability.json). The resulting observability levels are 1 for the dis-
534 closure proxy (no explicit inspiration or intermediate record), 2 for AI Idea Bench (inspiration and
535 decision fields), and 3 for FIRE-Bench and ResearchBench (inspiration, decision, and intermediate
536 records). This ordering is secondary context, not a numeric validity scale.

537 K Disclosure Proxy Item IDs

538 Convenience sample of public arXiv CS/AI-for-science HTML papers (2023–2026; $N=86$ after
539 excluding 11 without full ar5iv body text). IDs (ledger results/disclosure_item_ids.json):
540 2503.21248, 2411.02429, 2409.11363, 2502.05151, 2406.06769, 2407.01725, 2307.10635, 2412.17596, 2505.13259, 2503.13657, 2410.01792,
541 2501.05465, 2406.04598, 2311.07361, 2402.14034, 2501.09686, 2506.12345, 2410.19810, 2501.12948, 2409.05591, 2502.04364, 2412.05265,
542 2503.09516, 2504.01849, 2505.03422, 2408.12574, 2501.07542, 2411.14405, 2502.17416, 2503.16734, 2407.21783, 2504.08734, 2505.18098,
543 2410.05258, 2501.00663, 2502.06807, 2406.18532, 2504.16041, 2412.10360, 2503.04388, 2409.15277, 2502.09614, 2504.05186, 2411.05003,

544 2506.12993, 2605.23701, 2605.24060, 2607.22368, 2606.01789, 2606.26158, 2605.30284, 2503.13612, 2412.06438, 2501.05452, 2502.08791,
545 2411.14199, 2504.12371, 2505.00671, 2408.08309, 2501.17161, 2503.05555, 2410.24198, 2502.21231, 2504.21039, 2412.21154, 2501.19375,
546 2503.21778, 2407.07061, 2505.07271, 2410.07025, 2504.07640, 2411.17465, 2501.08528, 2409.18486, 2502.02561, 2504.03838, 2410.11939,
547 2501.13949, 2503.20009, 2408.04628, 2505.04358, 2411.09137, 2502.15707, 2504.12582, 2412.16178, 2501.04866.

548 Target rule: last conclusion-facing body segment after a deterministic section split (disclosure pilot
549 build script under `scripts/`).

550 L Cross-Setting Coverage and Score Sensitivity

551 Tables 17 and 18 consolidate coverage, falsification checks, score association, and rank sensitivity
across settings.

Table 17: Coverage and diagnostic controls. F1: added disclosure; F2: random targets; F3: hard negatives; F4: swapped evidence. Pan. counts model cells; sc denotes no hard-negative pool. These are diagnostic checks, not truth labels.

Setting	<i>n</i>	Pan.	F1↑	F2 rand	F3 hard	F4 shuf.	Extract / source
Disc.	86	228	pass	pass	77%	98%	7/86 ret.; arXiv HTML
AI Idea	3495	no	mixed	99.7%	98.3%	94%	gold; HF public
RB pilot	12	no	part.	100%	92%	100%	gold; GitHub tiny
FIRE	35	260	pass	100%	sc	100%	gold; public JSONL

552

Table 18: Score–recoverability associations and historical per-item winner changes, not aggregate rank reversals. Disjoint denotes $r(R_{\text{embed}}, s_{\text{lex}})$; reverse-family rows instead use embedding outcome scores.

Contrast	Same-fam. <i>r</i>	Disjoint <i>r</i>	Flips	Verdict
Disclosure	0.75	0.65 (S3)	13/20	tracks <i>R</i>
FIRE instr.	0.64	0.17 (n.s.)	—	same-family only
FIRE RQ	0.76	0.54	19/30	replicates
FIRE rev. fam.	—	0.17/0.15	—	weak <i>s_{embed}</i>

553 M Recommended Measurement-Validation Protocol

554 Minimum reporting steps for a discovery-endpoint benchmark $\mathcal{B}$:

- 555 1. Inventory input/target fields and pathway observability.
- 556 2. Define item-level *R* under an explicitly scorer-disjoint (or same-family labelled) family.
- 557 3. Run positive / random-target / hard-negative controls.
- 558 4. Intervene on evidence identity (derange, remove, or mask) with CIs.
- 559 5. Report input-only extraction and gold-assisted diagnostics separately, naming their inputs,
560 selection rule and gap denominator.
- 561 6. Relate official or panel scores to *R* (pooled and stratified).
- 562 7. Test common-item aggregate ranking stability, ties and margins separately from item-level
563 changes. Validate candidate admissibility independently of retrieval.

564 N Worked High/Low-*R* Examples

565 Clean-target subset examples at S3 (ledger `results/criterion_proxy_validation.json`; titles
566 from arXiv API). Snippets truncated; full texts in `data/disclosure_pilot/items.jsonl`. High-*R*
567 items show token/phrase reuse between disclosed input and gold endpoint; low-*R* items share
568 less endpoint text; the metric does not establish what additional reasoning would suffice. These
569 are automatic diagnostics rather than human labels; dual-model criterion proxy ledgers are under
570 `results/ (criterion_proxy_validation*.json)`.

Table 19: Worked high- and low- R disclosure-proxy examples at S3 (ledger criterion_proxy_validation.json).

Case	Item
Low $R=0.091$	arXiv:2502.02561 <i>Decision Theoretic Foundations for Conformal Prediction</i> . . . : methods-mid input shares little endpoint phrasing.
Low $R=0.092$	arXiv:2409.18486 <i>Evaluation of OpenAI o1</i> . . . : gold conclusion largely outside S3 window.
High $R=0.364$	arXiv:2505.13259 <i>From Automation to Autonomy</i> . . . : survey-style endpoint language reappears under methods-level disclosure.
High $R=0.391$	arXiv:2503.13612 <i>Searching for rotation in X-COP galaxy clusters</i> : technical claims already partially recoverable from S3 text.

O Contamination / Closed-Book Floor

Closed-book probes used eight R -stratified items and up to three models.

The ledger is qualitative_and_closedbook.json.

Table 20: Closed-book floors vs disclosure-panel stage means.

Input	mean s_{lex}	Note
Empty (closed-book)	0.031	vs panel S0 0.026
Title only	0.015	vs panel S3 0.074

The 86 proxy items span 2023 (2), 2024 (33), 2025 (45), and 2026 (6).

Public cutoff cards (approximate): Llama-3.2 Dec 2023; GPT-4o-mini family Oct 2023; Gemma-3/Qwen-2.5 2024 (vendor cards; not independently verified). Cutoff overlap with 2024+ papers does not prove leakage (Magar and Schwartz, 2022; Sainz et al., 2023); the low closed-book scores argue against memorisation as the sole S3 score driver.

P Target Extraction Quality Validation

The disclosure-proxy target rule was validated on 14 items per target-length tercile ($n=42$; seed 20260731; ledger target_extraction_audit.json).

In the stratified validation sample ($n=42$), 95% of targets are usable, 88% are conclusion-like, and 5% are hard failures (2/42). The full proxy retains all seven hard failures (7/86=8%).

Website footer text appears in 14% of the validation sample. Excluding hard fails, length-nearest intervention $\Delta=0.142$ (extraction_fail_sensitivity.json). These automated extraction checks assess the target text, not human validity of R .

Q Per-Item Detail Subsample (Stratified by S3 R)

Fifteen disclosure-proxy items stratified by S3 R_{avail} tercile (five per tercile; ledger results/per_item_detail_subsample.json).

Stratified per-item R detail (LOW/MID/HIGH S3 terciles).

	Tercile	item_id	R_{S0}	R_{S3}	R_{S4}
	LOW	arxiv:2402.14034	0.005	0.047	0.046
	LOW	arxiv:2407.01725	0.016	0.107	0.117
	LOW	arxiv:2501.19375	0.019	0.133	0.159
	LOW	arxiv:2412.17596	0.004	0.147	0.124
	LOW	arxiv:2410.07025	0.010	0.155	0.152
	MID	arxiv:2410.05258	0.018	0.158	0.154
591	MID	arxiv:2503.04388	0.016	0.170	0.196
	MID	arxiv:2504.05186	0.018	0.189	0.216
	MID	arxiv:2412.21154	0.023	0.210	0.244
	MID	arxiv:2502.15707	0.004	0.226	0.248
	HIGH	arxiv:2410.19810	0.006	0.231	0.268
	HIGH	arxiv:2409.15277	0.032	0.250	0.345
	HIGH	arxiv:2605.23701	0.014	0.268	0.315
	HIGH	arxiv:2406.06769	0.000	0.316	0.275
	HIGH	arxiv:2506.12345	0.043	0.762	0.783

592 Full $N=86$ rows remain in `results/pilot_per_item.json` (titles refreshed).

593 **R Panel Prompts and Extractive Rule**

594 **System prompt (verbatim):**

595 You are a scientific assistant. Given research context, project the paper's main findings/conclusions as a short paragraph of claims. Do not refuse.
596 Do not mention that you are an AI.

597 **User prompt template:**

598 Context:
599 `{stage_text truncated}`
600 Project the main scientific findings/conclusions of this research as a short paragraph of claims.

601 **Extraction access:** the historical panel uses input-only token-rich spans. The 0.508 S4 closure
602 instead used gold-guided sentence selection; the extraction comparison separates that diagnostic from
603 input-centrality and lead baselines.