



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



Odysseus Kaloidas¹

¹ National Technical University of Athens (ntua.gr)

29 Aug 2026 (modified: 14 Sept 2026)
 NeurIPS 2026 Workshop AI4MetaScience Submission
 AI4MetaScience, Reviewers, Authors Revisions CC BY 4.0

Edit ▾

Keywords: AI for science; measurement contracts; construct validity; scientific discovery; evaluation; recoverability

TL;DR: We introduce a dual-axis measurement-contract audit for scientific-discovery benchmarks, revealing when inputs over-determine historical endpoints or under-determine them—and when these failures distort scores and model rankings.

Abstract:

Claims about AI-assisted scientific discovery increasingly rest on benchmarks that reward agreement with a single historical paper endpoint. Before such scores are interpreted as evidence of hypothesis formation or scientific competence, the benchmark's measurement contract must itself be validated. We show that this contract can fail in two directions at once: inputs may over-determine the endpoint (availability R_{avail}) or under-determine it (identifiability R_{id}). We introduce a dual-axis measurement-contract framework using scorer-disjoint measures, falsifiers, model-free baselines, and length-nearest evidence-identity interventions. Empirically, on a CS/AI-for-science disclosure proxy ($N=86$), mean lexical R rises $0.017 \rightarrow 0.235$ under disclosure, and a length-nearest derangement reduces $S3$ recoverability by 0.136 (CI $[0.118, 0.158]$). The same identity effect transfers to AI Idea ($n=3,495$) and FIRE ($n=35$). In a blinded two-reader pilot over 60 frozen pairwise comparisons, readers agree with the continuous higher- R side on 73.3% and 68.3% of pairs (full-sample Cohen's $\kappa=0.83$). In a compact panel, endpoint scores are associated with R (same-family $r=0.75$; scorer-disjoint $r=0.65$), and the winning model changes between $S0$ and $S3$ for 65% of multi-model items—associational, not causal. FIRE combines high identifiability (gold@1 31/35) with moderate availability, indicating that the two axes capture distinct properties. We recommend validating both axes with evidence-identity interventions to delimit when endpoint scores can support claims about scientific discovery.

Track: Technical Track
Submission Type: Novel work (double-blind)
Ai Usage Disclosure: AI was extensively used for coding and improving the initial written draft.
Review Study Opt In: Yes, I opt in to the study
Email Sharing: We authorize the sharing of all author emails with Program Chairs.
Data Release: We authorize the release of our submission and author names to the public in the event of acceptance.
Submission Number: 49

Filter by reply type...

▼

Filter by author...

▼

Search keywords...

Sort: Newest First

▼

≡

≡

≡

-

=

≡

🔗

👁

Everyone

Program Chairs

Submission49 Reviewers

6 / 6 replies shown

Submission49 Authors

Submission49...

Submission49...

✕

Add:

Withdrawal



Paper Decision

🔗

Decision

by Program Chairs

📅 29 Sept 2026, 21:42 (modified: 30 Sept 2026, 12:06)

👁 Program Chairs, Reviewers, Authors

📄 Revisions

Decision: Accept (Poster)



Timely and important topic with a useful protocol, but the paper is hard to read and does not properly address possible data contamination.

Official Review

by Reviewer JdfX

📅 28 Sept 2026, 09:49 (modified: 30 Sept 2026, 12:06)

👁 Program Chairs, Reviewers, Authors

📄 Revisions

Review:

SUMMARY The authors argue that benchmarks scoring AI systems against a single historical paper endpoint can fail in two ways: the input may already give away the endpoint (availability), or several endpoints may fit the input equally well (identifiability). They propose a framework based on text-similarity measures, evidence-swapping interventions and simple baselines, and show across four benchmarks that recoverability rises with more disclosed input, drops when evidence is swapped, and is associated with model scores. They conclude with a protocol for validating such benchmarks before reading scores as evidence of scientific discovery.

STRENGTHS

- The question of what benchmark scores for AI-assisted discovery actually measure is timely and fits the workshop well.
- The statistical analysis is careful, with many control conditions and robustness checks.
- The seven-step protocol is practical and useful for benchmark developers.

MAJOR COMMENTS

- The paper is very hard to read and follow. It is extremely dense with information, and a clear storyline is partly missing, especially in the introduction. Many abbreviations are used without being introduced, both common technical ones (e.g., BoW, MRR) and ones specific to this paper (e.g., S_o, S_w and S_b in the closure formula).
- The authors built their own dataset with papers from 2023 to 2026, while the models used have knowledge cutoffs between late 2023 and 2024. A substantial part of the dataset (2 papers from 2023 and 33 from 2024) may therefore fall within the models' training data. Given the topic of the paper, it would have been more convincing to either use only papers published after the cutoffs or to test properly for data contamination.

MINOR COMMENTS

- All models in the panel are quite small and, by current AI standards, already rather old. This should be mentioned as a limitation.
- The reference list is not sorted.
- The title on OpenReview is entered incorrectly: the "I" in "Is" is missing.

Rating: 1: Accept (poster quality)

Confidence: 3: The reviewer is fairly confident that the evaluation is correct

Full Text Attestation: I confirm that I have read the full main text of this submission, and I take full responsibility for this review, including any parts produced with AI assistance.

AI Assistance Disclosure: improving wording, translation

Great paper showing why current AI tests are broken

Official Review by Reviewer JbZm  25 Sept 2026, 01:01 (modified: 30 Sept 2026, 12:06)

 Program Chairs, Reviewers, Authors  [Revisions](#)

Review:

This paper is a massive, highly actionable wake-up call for the AI community. It clearly demonstrates that current tests reward AIs for acting like search engines rather than scientists, and it provides a solid set of rules for fixing how we evaluate them

Rating: 2: Accept (oral quality)

Confidence: 4: The reviewer is confident but not absolutely certain that the evaluation is correct

Full Text Attestation: I confirm that I have read the full main text of this submission, and I take full responsibility for this review, including any parts produced with AI assistance.

AI Assistance Disclosure: NA

A instrumented but ultimately convenience-sample-bound study of discovery benchmark validity

Official Review by Reviewer Gge9  15 Sept 2026, 04:30 (modified: 30 Sept 2026, 12:06)

 Program Chairs, Reviewers, Authors  [Revisions](#)

Review:

Summary

This paper argues that benchmarks scoring AI systems against a single historical paper endpoint can fail in two directions: over-determination (the endpoint is already recoverable from the input) and under-determination (the historical endpoint is not actually unique among plausible candidates). The authors build an 86-paper disclosure proxy, run length-nearest evidence-identity interventions, and transfer the same test to AI Idea Bench, FIRE-Bench, and a small ResearchBench pilot, with bootstrap intervals and a blinded reader study.

Review

The paper is methodologically careful and honest about its own scope, and the dual-axis framing is a genuinely useful way to think about discovery benchmarks. That said, the empirical weight of the paper rests almost entirely on one author-built dataset and a lexical similarity measure, and the validation steps meant to ground that measure (human agreement, cross-model generalization) turn out to be weaker than the headline numbers suggest. Given that the paper's main contribution is a measurement claim about validity, these gaps feel more central than incidental, which is why I lean toward the reject side rather than treating them as minor limitations.

Weaknesses

- The primary population, the 86-paper disclosure proxy, is a convenience sample built by the authors themselves, not an existing or independently validated benchmark. Nearly all the headline numbers (0.017 to 0.235, $\Delta = 0.136$) come from this one construction, and the paper does not show the results are robust to alternative choices in the disclosure ladder or the target-extraction rule, so it is hard to know how much of the effect is an artifact of this particular pipeline.
- The core measure R is lexical overlap, and the human study meant to validate that this tracks something meaningful is fairly weak: the two readers only clearly agree with the R ordering in the LOW-HIGH stratum, while MID-HIGH is near chance (55%). This suggests the measure may only be picking up coarse, obvious cases, which undercuts using it as the primary evidence for the paper's central claim.
- The score-R association and the 65% winner-flip result, which are used to argue that benchmark rankings are unreliable, come only from four small compact models (3B to 7B scale). Since the motivating concern is about leaderboard-grade discovery agents, it is not clear this result transfers to the systems the paper is actually worried about, and no evidence is offered either way.

Questions

- Have the authors tried the disclosure pipeline on an independent or officially released dataset (even a small one) to check whether the core numbers hold outside the author-built proxy?
- Given the near-chance MID-HIGH reader agreement, would the authors consider the R measure validated for anything beyond coarse LOW vs HIGH separation?
- Is there any indication, even preliminary, of whether the score-R tracking and winner-flip results would hold with larger or frontier-scale models rather than the four compact models tested?
- Typo: ("s the Endpoint..." vs "Is the Endpoint...").

Rating: 0: Reject (detailed justification required in the review text)

Confidence: 2: The reviewer is willing to defend the evaluation, but it is quite likely that the reviewer did not understand central parts of the paper

Full Text Attestation: I confirm that I have read the full main text of this submission, and I take full responsibility for this review, including any parts produced with AI assistance.

Questions by AI Reader



Questions by AI Reader 09 Sept 2026, 13:08 (modified: 14 Sept 2026, 16:13)

Program Chairs, Authors, Reviewers Submitted, Reviewers AI Treatment Revisions

Questions For Authors:

1. **Section 3** (Page 3) defines the identifiability metric (R_{id}) by ranking the historical gold endpoint y^* against a candidate pool $y_{j=1}^C$, and **Table 4** (Page 7) reports high identifiability (e.g., gold@1 of 31/35) for FIRE instructions. Does the candidate pool for this identifiability evaluation consist exclusively of the $n = 35$ target endpoints within the FIRE dataset, and if so, how might the identifiability conclusions change if the inputs were evaluated against a larger, more realistic pool of distractor endpoints?
2. **Figure 1** and **Section 5.1** (Pages 3-4) show that lexical recoverability (R_{avail}) rises monotonically across all progressive disclosure stages up to S4, while **Section 5.3** (Page 4) notes that semantic recoverability (R_{embed}) plateaus once the abstract enters the window (S1). If semantic recoverability saturates early in the disclosure ladder, does this suggest that the scientific meaning of the historical endpoint is already fully constrained by the abstract, and that the later disclosure stages primarily supply specific vocabulary rather than novel conceptual information?



Author Response to Questions by Authors

Author Response to Questions by Authors (👁️ Odysseus Kaloidas)

📅 14 Sept 2026, 22:31 (modified: 14 Sept 2026, 22:41)

👁️ Program Chairs, Reviewers Submitted, Authors, Reviewers AI Treatment

📄 Revisions

Author Response:

Yes. The candidate pool consists of the 35 historical conclusion endpoints in our evaluated FIRE subset. Each instruction is compared with its corresponding endpoint and the other 34 endpoints using a fixed bag-of-words cosine scorer. The reported gold@1 of 31/35 and MRR of 0.93 therefore quantify how readily the historical endpoint can be distinguished within this pool. They do not establish uniqueness among all scientifically plausible conclusions. With the inputs and scorer held fixed, adding distractors can leave the gold rank unchanged or worsen it; the effect depends particularly on whether the additional endpoints are close competitors. ResearchBench makes a related distinction in its negative-inspiration construction (§3.3), separating closely related, same-discipline, and cross-discipline alternatives. Accordingly, a broader identifiability evaluation would vary both pool size and distractor relevance, rather than size alone. Our current finding is that moderate endpoint availability coexists with high within-pool identifiability. A harder pool would test the extent of that identifiability, while the separately measured availability and evidence-identity effects would remain unchanged for the same inputs and targets. This is why specifying the candidate pool is part of the measurement contract: the two axes answer complementary questions about what the input reveals and how well it distinguishes the historical endpoint from specified alternatives.

The semantic curve indicates substantial early alignment with the historical endpoint, but it does not establish that the abstract fully determines its scientific meaning. More precisely, mean embedding similarity rises from 0.239 at S0 to 0.682 at S1, increases further to approximately 0.747 at S2, and then remains within 0.745–0.753 through S4. The near-plateau is therefore most evident from S2 onward. This pattern is consistent with abstracts already conveying much of a completed study's broad conclusion. However, whole-text embedding similarity does not distinguish all the conditions, qualifications, mechanisms, and supporting evidence needed to substantiate that conclusion. Later disclosure can add such information without substantially changing the overall embedding. ProjectionBench's claim-level evaluation (§3) illustrates the complementary value of assessing individual claims rather than relying solely on an aggregate similarity measure. Our observed plateau is also specific to the frozen encoder and 6,000-character input window used in this analysis; it is not a demonstrated ceiling on the semantic information in later sections. The continuing lexical increase supports additional endpoint-specific coverage, but does not establish that the additional content is merely vocabulary. The supported interpretation is that broad alignment emerges early while detailed recoverability continues to increase. Reporting both measurements helps distinguish similarity to a historical finding from the information required to derive and justify it.

About OpenReview

FAQ

Sponsors

All Venues

Contact

Hosting a Venue

Terms of Use / Privacy Policy

Donate