



2 Versions ▾

The Selection-Target Problem in AI-Assisted Science: Reviewer Novelty and Later Scientific Uptake



*Odysseus Kaloidas*¹

¹ University of California, Berkeley (berkeley.edu)

29 Aug 2026 (modified: 14 Sept 2026)
 NeurIPS 2026 Workshop AI4MetaScience Submission

Edit ▾

AI4MetaScience, Reviewers, Authors Revisions CC BY 4.0

Keywords: AI-assisted science, meta-science, peer review, research agents, scientific evaluation, criterion validity, novelty evaluation.

TL;DR: We introduce decision-aware validation for AI research-selection targets, revealing that ICLR technical novelty loses citation signal after controls while empirical novelty retains it.

Abstract:

As AI systems generate more scientific ideas, selecting which ones merit attention becomes a central design problem. Reviewer-derived scores are attractive targets for this step. Yet reproducing a human score is useful only if it carries information about the downstream criterion the system is intended to serve. We instantiate a decision-conditioned target-validation test using Technical Novelty and Significance (TNS) and Empirical Novelty and Significance (ENS) from more than 3,800 ICLR 2022–2023 submissions. After conditioning on publication year and acceptance, TNS citation rank association falls from $p=0.214$ to 0.017 while ENS retains $+0.192$. This contrast persists across fixed citation horizons, identity controls and reviewer-level analyses. To examine what these reviewer scores may omit, a label-blind abstract-derived profile separately rates novelty claims, significance, evidence and clarity. Uptake-oriented dimensions retain association under the same controls and show stronger retrospective ranking association than reviewer novelty scores. Two raters agree on how the abstracts present significance, evidence and clarity. A post-cutoff cohort preserves the positive profile–citation direction under its unconditional design. Reviewer novelty is therefore not a unitary selection target: adjacent sub-scores can have sharply different relationships with later uptake. More broadly, AI-assisted science requires outcome-specific and decision-aware target validation before optimization. Because citations measure uptake rather than scientific value, the profile remains a diagnostic rather than a deployment metric.

Track: Technical Track

Submission Type: Novel work (double-blind)

AI Usage Disclosure: Generative AI assisted drafting, coding and analysis; the authors verified all scientific content and retain responsibility.

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: 48

Filter by reply type...

Filter by author...

Search keywords...

Sort: Newest First

-

=

Everyone

Program Chairs

Submission48 Reviewers

5 / 5 replies shown

Submission48 Authors

Submission48...

Submission48...

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)



Evaluating Reviewer Novelty Scores as Targets for AI-Assisted Scientific Selection

Official Review

 by Reviewer eTG8 26 Sept 2026, 01:58 (modified: 30 Sept 2026, 12:06)
 Program Chairs, Reviewers, Authors Revisions

Review:

Summary

This paper investigates whether reviewer novelty scores provide useful targets for AI-assisted scientific idea selection. Analyzing more than 3,800 ICLR 2022–2023 submissions, the authors find that technical novelty and significance scores retain little association with later citation uptake after controlling for publication year and acceptance, while empirical novelty and significance scores maintain a positive association.

Strengths

The paper addresses an important methodological question in AI meta-science: whether reproducing human reviewer scores provides a meaningful objective for scientific idea selection. Its analysis distinguishes different reviewer criteria and includes statistical controls and robustness checks. The structured assessment of significance, evidence, and clarity provides additional insight into the information associated with later scientific uptake.

Weaknesses

Citation uptake is an imperfect measure of scientific value, limiting the interpretation of the findings beyond citation-oriented selection. The analysis focuses primarily on ICLR submissions, leaving generalizability across other scientific disciplines and review systems uncertain. Additionally, the LLM-generated assessment is retrospective, and its effectiveness for prospective research selection remains unestablished.

Questions for the Authors

1. Do the findings persist when scientific impact is assessed using outcomes other than citation counts?

2. How consistently do the differences between technical and empirical novelty scores appear across other scientific disciplines?

3. What prospective validation would be needed before using the proposed assessment in AI-assisted research-selection systems?

Rating: 2: Accept (oral 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: AI assistance was used to help draft and format this review.

Strong, well-controlled TNS/ENS dissociation; the secondary LLM-profile result is more scorer-dependent than the main text conveys

Official Review by Reviewer ByTM  25 Sept 2026, 18:41 (modified: 30 Sept 2026, 12:06)

 Program Chairs, Reviewers, Authors  Revisions

Review:

Summary

The paper studies whether ICLR 2022–2023 novelty sub-scores predict later citations after controlling for year and acceptance. TNS drops from $\rho=0.214$ to a partial of 0.017, while ENS keeps 0.192. A secondary LLM profile scored from title and abstract finds stronger residual associations for significance, evidence and clarity.

Strengths

The TNS/ENS result is convincing and does not depend on the LLM profile. It holds within accepted and rejected papers, in each year, and with influential citations as the outcome. ENS and Recommendation serve as useful positive controls, and because TNS has higher observed-mean reliability than ENS, measurement noise does not explain the gap.

Main concern

The LLM-profile result is much more scorer-dependent than the main text conveys. On the same $n=84$ subset, Sonnet's Significance reaches $p=0.576$, while GPT-4o-mini and Gemini give 0.056 and 0.068 (Table 9); the four-dimension cluster transfers better but at about half the Sonnet value. The post-cutoff cohort gives C3 $p=0.159$, compared with 0.470 on the ICLR corpus. This does not prove contamination, and the human ratings (Table 3) are real counter-evidence, but the appendix's "primary-scorer" caveat should appear in the main text and abstract. A recognition or title-masked sensitivity check would help.

Minor

The pilot is stratified by decision and TNS bin, which nearly decorrelates TNS from acceptance, so pilot-based lifts over TNS+ENS (Table 22, Appendix N) overstate the full-data ratio of about 1.4 $\times$.

Overall assessment

The primary TNS/ENS finding is strong and stands on its own. My main concern affects the secondary LLM-profile contribution rather than the core result, and I would raise my score if a recognition-based check showed the profile association holding for papers the scorer cannot identify.

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.

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. **Table 15** (Page 23) shows that Empirical Novelty has a low exact agreement (53.5%) between the Target-Only (TO) and Reference-Grounded (RG) conditions, with the RG condition significantly shifting the mean score (+0.42). Given that assessing novelty conceptually requires comparison against prior work, what was the primary rationale for deploying the TO condition rather than the RG condition for the full-corpus ($n = 3,882$) profiling?
2. **Section 6.3** (Page 7) reports that the full 7-dimensional profile scored by GPT-4o-mini yields a citation correlation of $\rho \approx 0.130$, which is substantially lower than the primary Sonnet model's correlation of ≈ 0.559 . Does this performance drop suggest that the structured profile is highly sensitive to the specific instruction-following priors of the Claude model family, or do the authors attribute the gap primarily to a general reasoning capacity difference between model tiers?

Disclaimer: *This is not a peer review, but a tool to help clarify potential points of confusion for first-time readers. Authors are encouraged to stay true to their own work, but can use this opportunity to clarify any genuine misconceptions without feeling pressured to answer every question. Your responses will be made available to the reviewers starting from **September 15th**.*

Author Response to Questions by Authors

Author Response to Questions

 by Authors (

Odysseus Kaloidas

)

14 Sept 2026, 22:25 (modified: 14 Sept 2026, 22:42)

Program Chairs, Reviewers Submitted, Authors, Reviewers AI Treatment

Revisions

Author Response:

Regarding the choice of Target-Only profiling, TO matches the full-corpus diagnostic's purpose: measuring how a paper presents its contribution in a standardized title-and-abstract input and testing whether those signals relate to later uptake. Its novelty dimensions characterize claims presented in the text, while Significance, Evidence and Clarity describe apparent usefulness, specificity of reported evidence and articulation of the contribution. Establishing originality against prior literature is a complementary task. RG addresses that task more directly by supplying selected cited abstracts and an explicitly comparative rubric.

The Empirical Novelty result in Table 15 shows why this distinction matters. The 53.5% exact agreement and +0.42 mean shift indicate substantial sensitivity to the scoring condition. We therefore distinguish TO novelty-claim scores from reference-relative novelty judgments rather than treating them as interchangeable. The uptake-oriented findings are more stable across these conditions. On the 572 shared papers, Significance has 81.6% exact agreement and the same citation rank correlation under TO and RG, $\rho=0.489$. RG Evidence and Clarity also retain positive citation associations, at $\rho=0.407$ and 0.419 respectively. These findings support TO as a consistent abstract-level diagnostic across the corpus, with RG providing complementary evidence about reference-context sensitivity. They do not establish that reference comparison is unnecessary or that either condition supplies independent ground truth for novelty.

Regarding the GPT-4o-mini comparison, the observed gap establishes scorer dependence but does not distinguish model-family priors from general capability differences. The reported values, approximately $\rho=0.130$ for GPT-4o-mini and 0.559 for Sonnet, are citation correlations of leave-one-out Ridge predictions using each scorer's seven-dimensional features on the same 574-paper pilot. The scoring prompt is identical, but the experiment does not independently vary family and capability. Attributing the gap primarily to either would therefore go beyond the evidence.

The simpler aggregate diagnostic shows partial transfer. GPT-4o-mini's C4 composite, comprising Significance, Evidence, Clarity and Empirical Novelty, retains a positive citation association after conditioning on year and acceptance: partial $\rho=0.229$ compared with 0.030 for TNS. This supports the presence of uptake-related information under another scorer without implying that Sonnet's full predictive performance transfers. The composite association and fitted seven-dimensional prediction are distinct analyses.

Together, these results reinforce the paper's central implication: a proposed selection target requires validation for its particular scorer, information setting and downstream criterion. The primary TNS-ENS contrast is computed directly from original reviewer scores and later citations, independently of either profiling model. The profile provides a complementary retrospective diagnostic whose predictive magnitude is scorer-dependent, rather than a universally transferable deployment metric.

[About OpenReview](#)

[FAQ](#)

[Sponsors](#)

[All Venues](#)

[Contact](#)

[Hosting a Venue](#)

[Terms of Use / Privacy Policy](#)

[Donate](#)