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

Odysseus Kaloidas
UC Berkeley

Abstract

When AI systems use reviewer scores to rank scientific work, which scores carry information about the downstream outcome? We compare Technical Novelty and Significance (TNS) and Empirical Novelty and Significance (ENS) against later citation uptake in 3,443 conservatively linked ICLR 2022–2023 submissions. After conditioning on year and acceptance, TNS’s citation rank association falls from 0.200 to 0.009 whereas ENS retains 0.188. The contrast persists in fixed-horizon and reviewer-composition sensitivity analyses. In a retrospective selection exercise with a 10% reading budget, ENS selects about ten more high-uptake papers per 100 read than TNS among retained 2023 submissions. We then derive sharp bounds on selection differences under missing citation outcomes in a larger score-observed cache. At a fixed threshold of 25 citations within 24 months, ENS’s all-submission advantage remains positive in both years under every assignment of the missing binary outcomes. These bounds apply within the score-observed cache, not to the full venue. Differences between all-submission and accepted-only selection gains remain unresolved. Together, these findings distinguish criterion association, fixed-budget yield and robustness to missing outcomes as separate requirements for validating research-selection targets. Citations measure uptake rather than scientific value, and the observational analysis establishes neither a causal policy benefit nor improved AI-selector performance.

1 Introduction

AI research agents can generate, evaluate and refine scientific ideas at increasing scale [Yamada et al., 2025, Weng et al., 2025, Gottweis et al., 2025, Si et al., 2025]. They must also help allocate scarce execution, reading and review effort. Consider an assistant that ranks reviewed papers for a reader with time to examine only a small fraction. Training it to imitate a reviewer score can reproduce the reviewers’ ordering. Whether that ordering identifies papers that later receive scientific attention is a separate question. The score chosen as a training target therefore needs validation against the outcome it is meant to support.

We study this selection-target problem through two ICLR review fields. *Technical Novelty and Significance* (TNS) concerns the technical contribution; *Empirical Novelty and Significance* (ENS) concerns the empirical contribution. Both bundle novelty with significance and are scored by reviewers of the same submissions. We ask whether they carry similar information about later citation uptake, and whether the answer changes when acceptance is known. Citations operationalize scientific attention and reuse imperfectly; they are not ground truth for novelty, technical correctness or scientific value.

The reader’s available information changes the comparison. Across all submissions, a score can predict citations partly because it predicts acceptance, which changes visibility and dissemination. Once acceptance is known, we ask whether the score adds information beyond that decision and publication year. We measure this remaining association by removing the year and acceptance associations from both score ranks and citation ranks. We separately compare the high-uptake papers

reached by equally sized reading lists, either across submissions or within a known decision pool. Association across a population need not imply an advantage among the few papers selected.

Earlier NeurIPS work distinguishes quality, impact and confidence, and earlier ICLR work shows attenuation within decision strata [Cortes and Lawrence, 2021, Tran et al., 2021]. Our contribution links a matched comparison of TNS and ENS to the yield of a fixed reading budget and its sensitivity to missing citations. The primary analysis uses papers whose citation records pass a title-and-author matching rule. We then include additional papers with observed reviewer scores but unknown or unverified citations and bound the selection difference over every assignment of their high- or low-uptake outcomes. The original export was filtered for citation metadata and cited-reference abstracts, with unequal retention by decision status, so neither population is a complete ICLR submission census.

TNS has little residual citation-rank association after year and acceptance adjustment, while ENS remains positive. This contrast persists with fixed 24-month outcomes and submission-area conditioning. ENS also has higher retrospective selection yield across retained 2023 submissions, although differences between reading pools remain unresolved. Its all-submission yield advantage survives every assignment of missing binary outcomes in the larger score-observed cache in both years. A secondary analysis of abstract profiles examines additional uptake information while retaining human-agreement results and adverse scorer-portability evidence.

2 Related Work

Review criteria and downstream validity. PeerRead and subsequent work support review prediction and analysis [Kang et al., 2018, Gao et al., 2019, Wang et al., 2020, Jung et al., 2025]. Tran et al. [2021] report that overall ICLR score–citation associations weaken within decision strata. Cortes and Lawrence [2021] find little accepted-paper citation association for quality, but positive associations for impact and reviewer confidence. Thus criterion-specific validity and decision-stratum attenuation are established antecedents. We extend these studies by comparing TNS and ENS on the same papers and examining their fixed-budget yield and sensitivity to missing outcomes. ENS and Recommendation are internal positive controls: their associations show that the analysis can retain signal for other review criteria, but cannot rule out criterion-specific selection bias.

Automated scientific evaluation. LLM review studies examine agreement and informativeness [Liang et al., 2024, Akella et al., 2025, Baumann et al., 2026]; NovBench and RINoBench evaluate novelty judgments against expert labels [Wu et al., 2026, Schopf and Färber, 2026]. Reviewer-derived targets also appear in paper-quality and novelty predictors [Zhao et al., 2026a, Wu et al., 2025a,b, Wang et al., 2024]. Citation-aligned judges choose a different criterion directly [Rasool et al., 2026, Tong et al., 2026]. These designs make target choice consequential, but our retrospective associations do not establish that citation optimization produces better science. Reference-grounded novelty assessment addresses comparison with prior art [Zhang et al., 2026, Afzal et al., 2026]; it is a distinct function from forecasting uptake.

Construct and temporal boundaries. Metric audits show why validity needs more than aggregate agreement: Saakyan et al. [2026] test lexical novelty against expert creativity judgments, while Elangovan et al. [2025] expose sensitivity to human-label uncertainty. Our criterion is later uptake; these studies do not validate citations as scientific value. Zhao et al. [2026b] distinguish theory, method and results novelty on published Nature Communications papers. Our contrast concerns reviewer sub-scores and includes nonaccepted submissions. Cross-topic score calibration is another distinct failure mode [Xu et al., 2026]. Studies of realized research execution show that favorable idea-stage judgments need not persist after execution [Si et al., 2026]; bibliometric work distinguishes unconventionality, reviewer responses and later impact [Uzzi et al., 2013, Teplitskiy et al., 2022]. These findings motivate outcome- and stage-specific validation, rather than a universal hierarchy of novelty criteria.

Table 1: Submission coverage by year and decision. Frame is the saved OpenReview submission list; Export is the subset with citation metadata and at least one cited-reference abstract. Retained is Export/Frame. Linked additionally requires matching title and author metadata; Joint also requires TNS, ENS and Recommendation. Unavailable intermediate counts are unknown, not zero.

Year	Status	Frame	Export	Retained	Linked	Joint
2022	Accepted	1,094	843	77.1%	827	819
2022	Nonaccepted	1,523	756	49.6%	623	616
2023	Accepted	1,573	1,229	78.1%	1,205	1,184
2023	Nonaccepted	2,219	1,058	47.7%	828	824

3 Population and Outcomes

3.1 Source frame, retention and linkage

The saved OpenReview submission lists contain 2,617 ICLR 2022 and 3,792 ICLR 2023 papers. Each submission has a unique forum ID. An existing Semantic Scholar export covers 3,886 of these submissions, with target abstracts, citation counts, reviewer averages and at least one retrieved cited-reference abstract. This last requirement supported reference-aware modeling rather than the reviewer-score question and may favor papers with more readily indexed literature. The original matching and filtering logs are unavailable, so we cannot separate failed citation matches from reference-related or other exclusions. The export is therefore a selected subset, not a complete submission census.

Table 1 exposes retention by year and decision. The frame’s venue metadata separates accepted papers from those marked submitted; the export labels the latter rejected. We use *nonaccepted* because this cache does not independently resolve withdrawal histories. Exact forum IDs link the two local sources, and accepted status agrees for every retained forum. Differential retention remains a threat even after conditioning on acceptance: selection can vary with score and outcome within each decision stratum.

The citation-linkage audit compares OpenReview titles and authors with cached Semantic Scholar metadata. Fifteen links have different titles and incompatible author lists; examples include distinct works sharing the phrases “Self-Contrastive Learning” and “Replicable Bandits.” We exclude them. To avoid deciding the remaining renamed cases using their outcomes, the main population requires exact titles after case, accent and punctuation normalization and at least one compatible author identity (full normalized name, or surname plus first initial). This conservative rule retains 3,483 papers. It sacrifices genuine renamed versions and is a metadata-agreement check, not exhaustive manual verification. No duplicate Semantic Scholar identifiers occur in the original export.

There are 3,443 papers with joint TNS, ENS and Recommendation values in the retained set, and 3,440 with valid target-only profiles as well. The reviewer-label analysis requires no LLM score. For the missing-outcome analysis, an expanded review cache adds 1,024 score-complete forums to the original 3,843 joint-label export. The resulting 4,867-paper population has 1,424 unknown or unverified citation outcomes. The remaining 1,542 source forums lack complete scores. Missing values are never set to zero; each analysis states its population.

3.2 Scores and citation criteria

Each paper’s TNS and ENS values are means of numeric reviewer ratings on a 1–4 scale. TNS concerns the technical contribution; ENS concerns the empirical contribution. Both fields combine novelty and significance: higher response categories require substantive contributions and departure from prior work. They do not separately measure originality, empirical maturity or significance. ENS also permits a nonnumeric not-applicable response, which is excluded from its mean. A reviewer can therefore contribute to TNS but not ENS. Appendix C gives the observed response anchors and provenance. Recommendation is the overall recommendation score and uses its own scale. The earlier reviewer-level analysis found higher reliability for the mean TNS score than for ENS, making uniquely poor TNS reliability an inadequate explanation on its own.

Our cumulative outcome is $c_i = \log(1 + \text{citations}_i)$ from the retained Semantic Scholar snapshot. Timestamps are not available for every original record, so cumulative follow-up is imperfectly documented. The fixed-window sensitivity counts citation events in the 24 months after January 24, 2022 or January 21, 2023, respectively. Exact publication dates determine inclusion; an event with only a year is counted only when its entire calendar year lies inside the window. This is a lower-bound count. Known self-citations are removed when citing and cited papers share a Semantic Scholar author ID. Missing author identities and dates can leave residual error. The logarithm does not change rank correlations but defines the outcome for fitted regressions.

4 Questions and Analysis

4.1 Association at different decision stages

We first ask whether higher-scored papers receive more citations across the retained submissions, using raw Spearman correlation. We then ask whether that association remains after accounting for year and acceptance. To do this, we rank the scores and citation outcomes, regress each set of ranks on year and acceptance, and correlate the two sets of residuals. For example, a positive score residual means that a paper ranks higher than this regression predicts from its year and acceptance status; a positive citation residual has the analogous meaning. The question is whether those two deviations tend to occur together.

Formally, for a score x , let R_x, R_c be rank vectors with average ranks for ties. Let Z contain an intercept, year and binary acceptance, and let Z^+ be its Moore–Penrose pseudoinverse. The residuals are $e_x = R_x - ZZ^+R_x$ and $e_c = R_c - ZZ^+R_c$. Their Pearson correlation is

$$\rho_{\text{partial}}(x, c; Z) = \text{corr}(e_x, e_c). \quad (1)$$

This whole-population association does not directly tell a reader how many high-uptake papers a short reading list would contain. We test that separate question below.

Acceptance can change visibility and can also depend on several properties related to scores and citations. Conditioning on it can therefore remove mediated associations or introduce collider bias. Equation 1 describes residual association, not an intrinsic score effect or a causal policy benefit. A small conditional TNS coefficient leaves its unconditional information and other roles in peer review open. Comparing ENS under the same population, outcome and controls isolates a criterion contrast without eliminating criterion-specific selection bias.

4.2 Fixed attention budget

For each year, a reading pool consists of all retained submissions, accepted papers or nonaccepted papers. Each score ranks the same n papers and selects $k = \text{round}(0.10n)$ of them. For example, a 100-paper pool gives a ten-paper reading list. We call a paper high uptake when its fixed-window citation count reaches the pool’s 75th percentile; ties can make this group larger than 25%.

Tied scores are handled by expected inclusion rather than arbitrary ordering. Papers above the cutoff have weight one. If four papers tie for two remaining places, each has weight one-half; generally m papers tied for b places receive weight b/m . Let w_i be this weight and $h_i = 1$ for a high-uptake paper and zero otherwise. Then

$$\text{precision@}k = \frac{\sum_i w_i h_i}{k}, \quad \text{recall@}k = \frac{\sum_i w_i h_i}{\sum_i h_i}. \quad (2)$$

Precision is the expected fraction of the reading list with high uptake; recall is the fraction of all high-uptake papers reached by that list. Fractional inclusion averages over uniform tie-breaking and is invariant to row order. Random selection has expected precision equal to the pool’s high-uptake fraction. We measure overlap between score-based lists as $\sum_i \min(w_i^{\text{TNS}}, w_i^{\text{ENS}})/k$, the shared fraction of their inclusion weights, rather than Jaccard similarity between deterministic sets.

The exercise fixes available reviewer information and budget; it does not simulate selecting unreviewed ideas. The 10% budget was fixed after an earlier multi-budget analysis had been inspected, so results are exploratory. To test pool differences directly, we compare the ENS–TNS gains in all and accepted submissions, sharing bootstrap draws across the overlapping pools. Pool-relative

thresholds define high uptake separately in each pool. A second comparison uses the all-submission upper-quartile threshold for every pool in that year, so high uptake has the same meaning across them. The complete 5%, 10% and 20% budget family is retained, including unfavorable results, alongside Recommendation–ENS contrasts.

4.3 What missing outcomes can change

Could missing citations reverse the preference between two scores? We can still rank papers whose reviewer scores are observed, even when we cannot classify their uptake. In this larger cache, each score selects 10%. We hold the high-uptake threshold at the all-submission upper quartile of the matched population: 25 citations within 24 months in both years. This threshold stays fixed even as papers with unknown outcomes enter the ranking; we do not estimate a new quartile from their unknown citations. Only the unknown high/low outcomes vary; scores, selection weights and threshold remain fixed.

For rankings A, B , define $d_i = (w_i^A - w_i^B)/k$, each paper’s contribution to their precision difference if it has high uptake. Let O and M be the sets of papers with known and missing binary outcomes. The known contribution is $g_O = \sum_{i \in O} d_i h_i$. The complete precision difference must lie in

$$\left[g_O + \sum_{i \in M} \min(0, d_i), \quad g_O + \sum_{i \in M} \max(0, d_i) \right]. \quad (3)$$

To see the lower bound, suppose two ten-paper lists differ only in one paper each, and both outcomes are unknown. Giving the A -only paper low uptake and the B -only paper high uptake contributes -0.1 to $\text{precision}(A) - \text{precision}(B)$; reversing those assignments gives $+0.1$. An equally weighted shared paper cancels. Equation 3 applies this reasoning to every unknown paper, including fractional weights. Both endpoints are attainable, so the bounds are sharp. A positive lower endpoint guarantees A ’s higher yield for every allowed assignment. These are identification bounds, not sampling intervals, and exclude unknown scores or a changed outcome threshold.

4.4 Uncertainty and additional controls

The primary association and selection intervals use 2,000 paired paper resamples (seed 42), reranking and refitting controls or recomputing thresholds and selections. For direct comparisons between overlapping reading pools, we resample groups of papers with the same first author (2,000 draws, seed 260928 plus year), using each draw for both pools. We also report a paper-bootstrap sensitivity. A maximum-deviation bootstrap band covers all 12 pool contrasts across both years, three budgets and two threshold definitions. These intervals describe resampling uncertainty in the observed corpus; the missing-outcome bounds instead vary unknown outcomes with the observed inputs fixed. Appendix B gives the full comparisons.

We report estimates and 95% intervals without treating the exploratory comparisons as a family of confirmatory discoveries. In particular, an interval crossing zero is unresolved evidence, not equivalence; a near-zero coefficient is interpreted with its interval and the positive-control contrast. ICLR 2023 submission metadata supplies 13 author-selected primary areas. The area sensitivity adds exact category indicators to acceptance and uses the fixed-window outcome. It tests measured area composition, not every possible topical confound.

5 Results: Reviewer Criteria and Selection

5.1 The matched TNS/ENS contrast

TNS’s cumulative raw association is 0.200, while its conditional estimate is 0.009 with interval $[-0.024, 0.041]$. ENS’s conditional estimate is 0.188 ($[0.156, 0.221]$); the paired ENS–TNS difference is 0.179 ($[0.144, 0.216]$). Recommendation also retains residual association. These comparisons locate the divergence within the measured criteria, rather than establishing a general failure of reviewer judgment.

The fixed-window result preserves the contrast (Table 2). Separate year estimates appear in Appendix A; they use acceptance adjustment within year. With 2023 submission areas included, TNS is

Table 2: Year-plus-acceptance partial Spearman correlations and paired ENS–TNS contrast on the same 3,443 metadata-agreement papers. Fixed-window counts remove known self-citations. Intervals use 2,000 paired paper resamples, reranking and refitting controls.

Criterion	Cumulative	95% CI	Fixed 24 months	95% CI
TNS	0.009	[−0.024, 0.041]	0.001	[−0.032, 0.032]
ENS	0.188	[0.156, 0.221]	0.187	[0.156, 0.220]
Recommendation	0.177	[0.144, 0.208]	0.169	[0.136, 0.200]
ENS–TNS difference	0.179	[0.144, 0.216]	0.186	[0.151, 0.222]

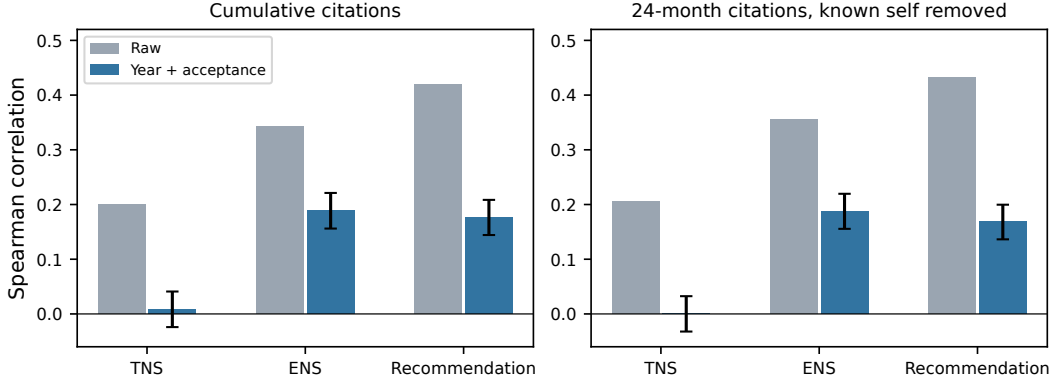


Figure 1: Citation-rank associations for papers passing the title-and-author match with all reviewer scores available ($n = 3,443$). Raw correlations use ranks directly; conditional correlations remove year and acceptance associations first. Whiskers show paired-paper 95% bootstrap intervals for conditional estimates. The second panel uses lower-bound 24-month citations with known self-citations removed. TNS attenuates toward zero while ENS retains a positive association.

0.022 and ENS is 0.171 on 2,000 papers, with intervals $[-0.020, 0.067]$ and $[0.128, 0.216]$, respectively. Persistence under these measured controls strengthens the local descriptive result. It does not resolve differential inclusion in the original export or the causal role of acceptance.

First-author cluster resampling across 3,103 identities preserves the cumulative TNS interval $[-.023, .041]$, ENS interval $[.156, .221]$ and their difference $[.142, .216]$. This addresses repeated first-author submissions, not every coauthor dependency. Among 2,973 complete author-history papers, controlling team size and maximum publication count through the last full pre-submission calendar year gives fixed-window TNS .007 $[-.031, .042]$ and ENS .165 $[.130, .201]$.

Because ENS permits not-applicable responses, the two paper means can average different reviewers. Holding those reviewers fixed tests whether this difference accounts for the contrast. Averaging both criteria over exactly the same reviewers on the 1,554 papers with cached joint review fields preserves the gap: 0.188 $[0.131, 0.246]$. Appendix C details this selected-subset test and illustrates score agreement and disagreement with papers chosen without using their citations.

5.2 Selection yield and the limits of stage comparisons

Across all retained 2023 submissions, ENS precision is 0.424 compared with 0.323 for TNS (Table 3). The paired difference is 0.101 $[0.040, 0.167]$: about ten additional high-uptake papers per 100 read at the same 10% budget. Recommendation provides an overall-review baseline, so the practical choice is not restricted to novelty fields. Appendix A reports the 2022 results and overlap between lists.

Within accepted 2023 papers, the ENS–TNS precision difference is 0.065 $[-0.006, 0.155]$. To ask whether restricting the pool changes ENS’s advantage, we subtract this accepted-pool gain from the all-submission gain, rather than compare their significance labels. The resulting difference is 0.036 $[-0.040, 0.101]$ with pool-relative thresholds and 0.030 $[-0.026, 0.091]$ with a common threshold, using first-author cluster resampling. Both intervals and all 12 simultaneous pool-contrast bands

Table 3: Fraction of selected papers with high uptake at a 10% reading budget: n is the pool size and k the selected count. High uptake means reaching the pool-specific 75th percentile of 24-month citations. Random is the expected precision of uniform selection; Rec. is Recommendation. Ties receive fractional inclusion. Paired exploratory intervals recompute selections and thresholds in each paper resample. Appendix A gives both years and all pools.

2023 pool	n	k	Random	TNS	ENS	Rec.	ENS–TNS [95% CI]
All	2008	201	0.251	0.323	0.424	0.482	0.101 [0.040, 0.167]
Accepted	1184	118	0.253	0.267	0.332	0.383	0.065 [−0.006, 0.155]
Nonaccepted	824	82	0.274	0.277	0.368	0.402	0.090 [−0.011, 0.180]

Table 4: Possible precision differences at a 10% budget among papers with all reviewer scores observed; Rec. is Recommendation. Unknown counts include absent and unverified citation links. High uptake means at least 25 citations within the fixed window in either year. Bounds cover every high/low assignment to unknown outcomes, not sampling uncertainty. A positive lower endpoint preserves the first score’s advantage throughout those assignments. Papers without complete scores are excluded.

Year	Pool	n	Unknown	ENS–TNS	Rec.–ENS
2022	All	2163	728	[0.015, 0.237]	[−0.096, 0.064]
2022	Accepted	974	155	[0.027, 0.230]	[−0.135, 0.031]
2022	Nonaccepted	1189	573	[−0.213, 0.473]	[−0.368, 0.257]
2023	All	2704	696	[0.015, 0.173]	[−0.044, 0.076]
2023	Accepted	1312	128	[0.004, 0.107]	[0.032, 0.094]
2023	Nonaccepted	1392	568	[−0.169, 0.266]	[−0.250, 0.293]

include zero. Thus the data leave the difference between pools unresolved; conditional association alone does not guarantee a particular selection gain. Recommendation’s 2023 all-submission gain over ENS is also uncertain: .057 [−.012, .136]. These retrospective comparisons do not estimate a causal policy benefit.

5.3 Selection with unobserved citation outcomes

Table 4 expands the population without imputing missing outcomes. Among all score-complete submissions, the ENS–TNS precision bounds are [.015, .237] in 2022 and [.015, .173] in 2023. Thus ENS retains at least about 1.5 additional high-uptake papers per 100 read even under the least favorable assignment of missing outcomes. This is a finite-cache guarantee conditional on the observed scores and 25-citation threshold, not a confidence statement about future papers.

The accepted-pool bounds are also positive. They do not resolve the earlier accepted-pool sampling interval: here the population is larger, the outcome threshold is fixed across pools, and the bounds vary missing outcomes rather than resample papers. Nonaccepted-pool bounds include both signs in each year. Recommendation–ENS bounds also include both signs across all submissions. Missing outcomes therefore leave some choices unresolved while preserving the ENS–TNS all-submission ordering. The result does not extend to the 1,542 forums without complete scores.

6 A Secondary Diagnostic of Omitted Information

6.1 Abstract-presented dimensions

The reviewer-score contrast does not identify which properties of a paper carry the remaining uptake information. As a secondary diagnostic, a target-only (TO) prompt receives the title and first 1,500 abstract characters, excluding reviewer labels, decisions and citations. It scores seven dimensions from 1–5: Problem, Method, Empirical and Recombination novelty, plus Significance, Evidence and Clarity. These names refer to properties presented in the abstract. Novelty dimensions concern claimed contributions rather than verified departure from prior work. Evidence measures specificity of reported support, not verified correctness; Significance measures apparent reach, not true scientific

Table 5: Secondary diagnostic on the same 3,440 papers with both profiles and reviewer scores. Partial correlations adjust for year and acceptance. AP is tie-aware average precision for the cumulative-citation top quartile (873 positives; base rate 0.254). Accepted is binary acceptance status; its partial correlation is undefined because acceptance is itself a control. Entries are descriptive estimates, not familywise tests. Table 2 gives primary intervals; text reports the profile interval after pilot exclusion.

Predictor	Raw ρ	AP	Partial ρ
TNS	0.200	0.291	0.009
ENS	0.343	0.360	0.188
Recommendation	0.419	0.374	0.176
mean(TNS,ENS)	0.311	0.337	0.117
Accepted	0.408	0.325	—
Significance	0.437	0.425	0.398
Evidence	0.351	0.359	0.304
Clarity	0.370	0.342	0.314
$\mathcal{C}_3$	0.475	0.485	0.421
$\mathcal{C}_4$	0.483	0.499	0.441

value. TO7 denotes the seven scores. $\mathcal{C}_3$ is the arithmetic mean of Significance, Evidence and Clarity; $\mathcal{C}_4$ adds Empirical novelty to that mean.

The original scorer was `claude-sonnet-4-6` through Vertex AI, using structured JSON, an 800-token cap and no temperature/top- p override. Recognition through pretraining remains possible despite withholding labels. Reference-grounded scoring adds up to five cited abstracts and two comparison dimensions; it is secondary to the target-only analysis. Exact prompts and anchors are retained in Appendix D.

The composites were defined before the citation-linkage correction, without external preregistration. $\mathcal{C}_3$ groups significance, reported evidence and clarity. $\mathcal{C}_4$ was selected on a 574-forum pilot using citation association and discrimination of top-quartile papers. We retain these definitions and test them on eligible papers outside the entire pilot-source union. This separation avoids reselecting dimensions on the evaluation papers; it does not make either the original selection or the present analysis prospective.

The profile comparison uses a common 3,440-paper population. Table 5 shows raw association, tie-aware average precision (AP) for the cumulative-citation top quartile, and conditional association. AP integrates precision over score thresholds and is distinct from the fixed-budget precision in Table 3. On 2,932 corrected papers outside the pilot, $\mathcal{C}_3$ retains partial association 0.418 ([0.385, 0.449]). This comparison supplies a diagnostic of residual uptake information, not a validation of deployment thresholds.

We also fit Ridge regressions to predict log citations from either the seven TO scores or TNS and ENS. Each paper is predicted by a model fitted, including standardization, on the other papers: leave-one-out (LOO) prediction. The correlation between predicted and observed citation ranks is higher for TO7 than for the TNS+ENS predictor by 0.177 ([0.139, 0.215]). The interval resamples fixed predictions and omits training-set uncertainty. Appendix C.3 gives this comparison and a 2022-to-2023 holdout against TF-IDF text features and SciBERT and SPECTER2 embeddings. These fixed-configuration baselines test additional predictive information, not the benefit of structured elicitation or prospective validity.

6.2 Human agreement and adverse portability evidence

Two researchers with ML experience independently scored 100 blinded abstracts on Significance, Evidence and Clarity. Neither reported prior knowledge of an assigned paper or a conflict of interest. For each rater, the triad is the mean of those three scores per paper. Human–human agreement correlates those 100 paper means; human–Sonnet agreement correlates the two-rater mean with the model’s triad. The 300 dimension cells are not treated as independent papers. Six cells differed by at least two points. Paper-bootstrap intervals accompany both agreement statistics in Table 6.

Human agreement remains evaluated on all 100 abstracts because it does not use citation links. The exploratory outcome analysis uses only the 90 items passing the same metadata rule as the main

Table 6: The first two rows give Spearman agreement between per-paper triad means on all 100 rated abstracts. The next two give year-plus-acceptance partial citation correlations on the 90 metadata-agreement papers; the last gives their paired difference. All intervals resample papers, not the 300 dimension cells.

Statistic (papers)	Estimate	95% CI
Human–human triad (100)	0.723	[0.610, 0.807]
Human–Sonnet triad (100)	0.598	[0.452, 0.718]
Human Significance partial (90)	0.454	[0.257, 0.621]
TNS partial (same 90)	0.129	[−0.079, 0.319]
Human Significance–TNS	0.325	[0.070, 0.579]

study. Human Significance and TNS are compared on exactly these papers with year and acceptance adjustment. Agreement establishes some reproducibility of judgments about the abstracts; it does not validate true significance, full-paper evidence quality or technical correctness.

Changing the scorer does not preserve the fitted profile’s advantage. On the 508 title-and-author-matched papers from the original 574-paper pilot, the same LOO prediction protocol gives GPT-4o-mini TO7 $\rho = .135$, TNS+ENS .242 and Sonnet TO7 .518. The paired GPT-minus-label difference is $-.107 [-.229, .010]$, while Sonnet-minus-label is .277 [.173, .372]. Thus this comparison does not establish a GPT profile advantage.

Holding the scorer fixed asks a different question: are the scores stable when the prompt changes? A separate 60-paper, 240-call Sonnet study tested whether reordering the requested dimensions, reordering the JSON schema or paraphrasing the prompt preserved score rankings. Dimension and schema reordering passed the frozen correlation threshold of .85. Under paraphrase, C_3 fell below it (.815, versus .855 for C_4). These comparisons concern score stability, not citations, and do not depend on citation linkage.

A separate temporal check uses 295 June–July 2024 arXiv cs.LG preprints sampled without consulting citation outcomes. They were scored by pinned gpt-4o-2024-05-13, whose stated October 2023 cutoff precedes the preprints [OpenAI, 2024, 2026]. Against exact-dated 18-month citations, C_3 has correlation .159 [.051, .267] and C_4 .210 [.099, .319]. Event dates cover 92.7% of citations. These positive unconditional associations concern a different corpus, scorer and follow-up period. The cohort has neither TNS/ENS nor an acceptance analog, so it does not replicate the conditional reviewer-score contrast.

7 Discussion and Limitations

Two adjacent reviewer novelty fields need not be interchangeable targets for later uptake. Their conditional associations differ on the same papers, and ENS’s all-submission selection advantage persists under arbitrary missing binary outcomes in the larger score-observed cache. The unresolved pool contrasts also show why conditional association, selection yield and missing-outcome robustness must be assessed separately: evidence for one does not settle the others.

For an AI system evaluated before a decision is available, unconditional and prospective ranking performance are relevant; conditioning away acceptance answers a different question. For postdecision retrieval, comparisons within the known decision pool are relevant. For a reward trained to reproduce reviewer scores, our result motivates testing the downstream criterion explicitly, but does not estimate how changing the reward would change generated research. A prospective test would freeze eligible papers, available inputs, budget and outcome horizon before any outcomes are observed and compare the candidate target with Recommendation and simple text baselines.

The principal remaining limitation is population coverage. The reference filter and unequal retention prevent a full-venue interpretation. The bounds address arbitrary missing binary outcomes within the score-observed cache, but cannot recover unknown score rankings, validate renamed citation links or reconstruct withdrawal histories. Recovering those records and a compatible independent review cycle would test broader transport. Citation-intent research already distinguishes method use from background and result comparison [Cohan et al., 2019]; this study does not independently verify

those forms of use. Establishing which scientific consequences TNS and ENS capture requires such evidence.

Additional limits are observational conditioning, incomplete citation dates and author identities, dependence beyond first-author clusters, and the small human sample. Fixed-alpha comparisons do not exhaust representation-specific tuning. Historical profile selection and pretraining exposure preclude prospective claims. The post-cutoff cohort and prompt replays cannot substitute for reviewer-score replication. Citations measure uptake that can reflect visibility, controversy or fashion; no result supports optimizing them as scientific value.

8 Conclusion

Technical and empirical novelty/significance scores carry different residual information about later citation uptake in the verified-link ICLR population. At a fixed reading budget, ENS’s all-submission advantage also survives arbitrary missing binary outcomes in a larger score-observed cache. Direct stage comparisons and some rival rankings remain unresolved. Together these findings distinguish criterion association, selection yield and missing-outcome robustness: evidence for one does not automatically establish the others.

Ethics and AI assistance. The profiles are retrospective research instruments and require prospective validation before triage use. Generative AI assisted writing, code and analysis. The authors retain responsibility for the scientific claims. Public manuscript metadata and reviewer aggregates are used for population-level analysis; the work does not evaluate named reviewers.

References

- Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. Beyond “Not Novel Enough”: Enriching scholarly critique with LLM-assisted feedback. In *EACL*, 2026.
- Akhil Pandey Akella, Harish Varma Siravuri, and Shaurya Rohatgi. Pre-review to peer review: Pitfalls of automating reviews using large language models. *arXiv preprint arXiv:2512.22145*, 2025.
- Joachim Baumann, Jian Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. In *International Conference on Machine Learning (ICML)*, 2026.
- Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. Structural scaffolds for citation intent classification in scientific publications. In *Proceedings of NAACL-HLT*, pages 3586–3596, 2019. doi: 10.18653/v1/N19-1361. URL <https://aclanthology.org/N19-1361/>.
- Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv preprint arXiv:2109.09774*, 2021.
- Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Bodapati, and Dan Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *ICLR*, 2025. URL <https://openreview.net/forum?id=E8gYIrbP00>.
- Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In *NAACL-HLT*, 2019.
- Juraj Gottweis et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.
- Sangkeun Jung, Goun Pyeon, Inbum Heo, and Hyungjin Ahn. What drives paper acceptance? A process-centric analysis of modern peer review. *arXiv preprint arXiv:2509.25701*, 2025. Version 2, 29 January 2026.
- Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *NAACL*, 2018.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, et al. Can large language models provide useful feedback on research papers? A large-scale empirical analysis. *NEJM AI*, 1(8), 2024.

- OpenAI. GPT-4o system card. Technical report, OpenAI, 2024. URL <https://openai.com/index/gpt-4o-system-card/>.
- OpenAI. GPT-4o Model. OpenAI API Documentation, 2026. URL <https://developers.openai.com/api/docs/models/gpt-4o>. Accessed 2026-07-22.
- Abdur Rasool, Xiaohui Huang, Yanqing Hu, and Linyi Yang. ReviewGuard: Aligning LLM-assisted peer review with long-term scientific impact. *arXiv preprint arXiv:2606.24892*, 2026.
- Arkadiy Saakyan, Najoung Kim, Smaranda Muresan, and Tuhin Chakrabarty. Death of the novel(ty): Beyond n-gram novelty as a metric for textual creativity. In *ICLR*, 2026. URL <https://openreview.net/forum?id=z2idLjqzBe>.
- Tim Schopf and Michael Färber. Is this idea novel? an automated benchmark for judgment of research ideas. *arXiv preprint arXiv:2603.10303*, 2026.
- Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *ICLR*, 2025. URL <https://openreview.net/forum?id=M23dTGWCZy>.
- Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation–execution gap: Execution outcomes of LLM-generated versus human research ideas. In *ICLR*, 2026. URL <https://openreview.net/forum?id=Flp8l6Puy>.
- Misha Teplitskiy, Hao Peng, Andrea Blasco, and Karim R. Lakhani. Is novel research worth doing? Evidence from peer review at 49 journals. *Proceedings of the National Academy of Sciences*, 119(47):e2118046119, 2022. doi: 10.1073/pnas.2118046119.
- Jingqi Tong et al. AI can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.
- David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of OpenReview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2021.
- Brian Uzzi, Satyam Mukherjee, Michael Stringer, and Ben Jones. Atypical combinations and scientific impact. *Science*, 342(6157):468–472, 2013.
- Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Small, Sujian Li, and Heng Ji. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *INLG*, 2020.
- Zhongyi Wang, Haoxuan Zhang, Jiangping Chen, and Haihua Chen. An effective framework for measuring the novelty of scientific articles through integrated topic modeling and cloud model. *Journal of Informetrics*, 18(4):101587, 2024. doi: 10.1016/j.joi.2024.101587.
- Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. CycleResearcher: Improving automated research via automated review. In *ICLR*, 2025.
- Wenqing Wu, Chengzhi Zhang, Tong Bao, and Yi Zhao. SC4ANM: Identifying optimal section combinations for automated novelty prediction in academic papers. *Expert Systems with Applications*, 273:126778, 2025a.
- Wenqing Wu, Chengzhi Zhang, and Yi Zhao. Automated novelty evaluation of academic paper: A collaborative approach integrating human and large language model knowledge. *Journal of the Association for Information Science and Technology*, 76(11):1452–1469, 2025b. doi: 10.1002/asi.70005.
- Wenqing Wu, Yi Zhao, Yuzhuo Wang, Siyou Li, Juexi Shao, Yunfei Long, and Chengzhi Zhang. NovBench: Evaluating large language models on academic paper novelty assessment. In *Findings of ACL*, 2026.
- Binyan Xu, Xilin Dai, Fan Yang, and Kehuan Zhang. Reviewer scores are not comparable across research areas in ML peer review. *arXiv preprint arXiv:2607.27209*, 2026. doi: 10.48550/arXiv.2607.27209.

- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025.
- Ming Zhang, Kexin Tan, Yueyuan Huang, Yujiong Shen, Chunchun Ma, Li Ju, Xinran Zhang, and Yuhui Wang. OpenNovelty: Agentic novelty assessment via claim extraction and prior-work comparison. *arXiv preprint arXiv:2601.01576*, 2026.
- Penghai Zhao, Jinyu Tian, Qinghua Xing, Xin Zhang, Zheng Li, Jianjun Qian, Ming-Ming Cheng, and Xiang Li. NAIPv2: Debaised pairwise learning for efficient paper quality estimation. In *ICLR*, 2026a. URL <https://openreview.net/forum?id=rNl8XiSHiJ>.
- Yi Zhao, Chenggang Yang, Yuzhuo Wang, Tong Bao, Heng Zhang, and Chengzhi Zhang. Beyond single-dimension novelty: How combinations of theory, method, and results-based novelty shape scientific impact. *arXiv preprint arXiv:2604.12471*, 2026b.

A Complete Results and Population Rules

The current main-paper population is defined by metadata agreement before analysis: normalize titles to lowercase ASCII alphanumerics after Unicode decomposition; require equal normalized titles and at least one shared normalized author name or surname–first-initial pair. This rule leaves 3,483 of 3,886 exported forums, with 3,443 joint-label records. Fifteen excluded links also have incompatible author lists. The remaining exclusions include genuine renamed papers and are not all alleged mismatches. Forty retained forums lack at least one joint reviewer value; three joint-label forums lack usable profiles. These exclusions are separate from the 2,523 source-frame forums absent from the original export.

The matching algorithm and intermediate exclusion logs preceding that export are unavailable. Consequently, the difference between 6,409 cached forums and 3,886 exported forums cannot be partitioned into failed matches, reference filtering and other causes. Nonaccepted source status does not resolve withdrawals. No missing count or outcome has been replaced by zero. Raw data, corrected inclusion decisions and deterministic scripts are kept separately.

A.1 Both years and all decision pools

The following estimates complete the main-paper comparison without choosing pools by significance. Pool-specific top-quartile thresholds differ. Sampling intervals are exploratory marginal intervals. There is no claim of a common effect across years or budgets.

Table 7: Within-year acceptance-adjusted correlations for fixed 24-month citations after known-self removal.

Year	n	Criterion	Partial ρ	95% CI
2022	1435	TNS	0.014	[−0.038, 0.063]
2022	1435	ENS	0.195	[0.144, 0.247]
2022	1435	Recommendation	0.180	[0.130, 0.230]
2022	1435	ENS–TNS	0.180	[0.121, 0.242]
2023	2008	TNS	−0.009	[−0.050, 0.034]
2023	2008	ENS	0.182	[0.138, 0.224]
2023	2008	Recommendation	0.159	[0.116, 0.201]
2023	2008	ENS–TNS	0.191	[0.142, 0.237]

Table 8: 2022 complete 10%-budget analysis. Outcome thresholds and rankings are recomputed within each bootstrap resample.

Pool	n	Score	Precision	95% CI	Difference vs. TNS [95% CI]
All	1435	TNS	0.289	[0.227, 0.356]	—
All	1435	ENS	0.440	[0.366, 0.511]	0.150 [0.065, 0.233]
All	1435	Recommendation	0.432	[0.354, 0.501]	0.143 [0.057, 0.220]
Accepted	819	TNS	0.271	[0.177, 0.360]	—
Accepted	819	ENS	0.388	[0.279, 0.472]	0.117 [−0.006, 0.225]
Accepted	819	Recommendation	0.340	[0.248, 0.442]	0.070 [−0.026, 0.177]
Nonaccepted	616	TNS	0.300	[0.191, 0.405]	—
Nonaccepted	616	ENS	0.431	[0.318, 0.555]	0.131 [0.001, 0.283]
Nonaccepted	616	Recommendation	0.334	[0.228, 0.445]	0.034 [−0.112, 0.171]

Table 9: 2023 complete 10%-budget analysis. Outcome thresholds and rankings are recomputed within each bootstrap resample.

Pool	n	Score	Precision	95% CI	Difference vs. TNS [95% CI]
All	2008	TNS	0.323	[0.265, 0.384]	—
All	2008	ENS	0.424	[0.368, 0.486]	0.101 [0.040, 0.167]
All	2008	Recommendation	0.482	[0.417, 0.556]	0.159 [0.091, 0.233]
Accepted	1184	TNS	0.267	[0.193, 0.334]	—
Accepted	1184	ENS	0.332	[0.260, 0.413]	0.065 [−0.006, 0.155]
Accepted	1184	Recommendation	0.383	[0.306, 0.461]	0.116 [0.038, 0.207]
Nonaccepted	824	TNS	0.277	[0.184, 0.348]	—
Nonaccepted	824	ENS	0.368	[0.264, 0.436]	0.090 [−0.011, 0.180]
Nonaccepted	824	Recommendation	0.402	[0.301, 0.489]	0.124 [0.013, 0.242]

Shared inclusion mass for ENS/TNS and Recommendation/TNS, respectively: 2022 all 0.372/0.352; 2022 accepted 0.299/0.297; 2022 nonaccepted 0.335/0.312; 2023 all 0.418/0.365; 2023 accepted 0.391/0.345; 2023 nonaccepted 0.394/0.254. These measure common fractional budget mass.

A.2 Dependence and pre-submission author controls

First-author cluster resampling draws OpenReview first-author identities with replacement and retains all associated papers; it does not model the entire coauthor network. Author-history controls use complete cached Semantic Scholar identities and publication histories, team size and the maximum focal-author publication count through the last full year before submission (2020 or 2021). Continuous author-history controls are ranked, as in the historical specification. On 3,443 papers in 3,103 first-author clusters, 2,000 resamples (seed 873) give TNS, ENS and ENS–TNS cumulative partial-correlation intervals [−0.023, 0.041], [0.156, 0.221] and [0.142, 0.216]. On 2,973 complete author-history papers, the fixed-window adjusted correlations are TNS 0.007 (−0.031, 0.042) and ENS 0.165 (0.130, 0.201). These observed controls do not identify a causal score effect.

A.3 Interpretation of secondary and historical evidence

All primary correlations, selection-budget results, current profile comparisons, fitted baselines, cross-scorer comparisons and human outcome comparisons in this manuscript use the corrected population stated in their captions or text. Human agreement uses all 100 abstracts because it is independent of citation linkage. The separate post-cutoff arXiv study and prompt-order/paraphrase replays do not depend on the ICLR citation matching filter. The GPT/Sonnet comparison uses 508 metadata-agreement papers from the original scorer pilot, with the same fold-local regression and reviewer-label baseline. Its intervals resample fixed out-of-fold predictions and do not include nested refitting.

The original 574-forum pilot selected the four-dimension composite using a positive lower confidence bound for its dimension’s correlation advantage over TNS and an AUROC greater than the original acceptance baseline of .663. This rule and its outcomes preceded the current linkage correction. The present analysis preserves the selected membership rather than rerunning the gate, and removes the entire pilot-source union for its disjoint comparison. That distinction prevents the corrected full corpus from silently becoming a new development set.

The original human assessment measured 100 independently rated abstracts with three dimensions each. Human–human and human–model triad correlations use paper means, not pooled cells. The ten papers removed from the current human outcome comparison are removed only because of the same metadata rule, not because of disagreement or their citation values.

B Direct Selection Contrasts and Missing Outcomes

To compare selection gains across overlapping reading pools, we retain the main-paper title-and-author-matched population and citation events and use 2,000 resamples (seed 260928 plus year). First-author names are normalized to alphanumeric Unicode text; papers lacking a name receive unique forum-based cluster IDs. There are 1,363 first-author clusters in 2022 and 1,931 in 2023. Draws are independent across years; the family bands do not model cross-year author dependence or

Table 10: Direct all-submission minus accepted-pool difference in the ENS–TNS precision gain. A single first-author cluster draw supplies both overlapping pools. Common uses the year-wide threshold; Pool uses each pool’s threshold. Family bands use the centered-bootstrap maximum absolute deviation across the 12 displayed contrasts.

Year	Threshold	Budget	Difference	Marginal 95% CI	Family 95% band
2022	Pool	5%	0.054	[−0.081, 0.190]	[−0.091, 0.198]
2022	Pool	10%	0.033	[−0.050, 0.129]	[−0.111, 0.177]
2022	Pool	20%	0.003	[−0.050, 0.059]	[−0.141, 0.147]
2022	Common	5%	0.027	[−0.070, 0.131]	[−0.117, 0.171]
2022	Common	10%	−0.010	[−0.078, 0.075]	[−0.154, 0.134]
2022	Common	20%	−0.019	[−0.063, 0.030]	[−0.163, 0.125]
2023	Pool	5%	−0.017	[−0.129, 0.085]	[−0.161, 0.127]
2023	Pool	10%	0.036	[−0.040, 0.101]	[−0.108, 0.180]
2023	Pool	20%	−0.007	[−0.051, 0.038]	[−0.151, 0.137]
2023	Common	5%	−0.051	[−0.121, 0.021]	[−0.195, 0.093]
2023	Common	10%	0.030	[−0.026, 0.091]	[−0.115, 0.174]
2023	Common	20%	−0.019	[−0.055, 0.020]	[−0.163, 0.125]

Table 11: 2022 complete selection sensitivity on the strict population. Intervals resample first-author clusters, preserving pool overlap and recomputing thresholds and score selections. All entries are exploratory marginal intervals.

Threshold	Budget	Pool	ENS–TNS [95% CI]	Rec.–ENS [95% CI]
Pool	5%	All	0.118 [−0.015, 0.259]	−0.031 [−0.160, 0.102]
Pool	5%	Accepted	0.065 [−0.094, 0.231]	0.009 [−0.130, 0.180]
Pool	5%	Nonaccepted	0.003 [−0.154, 0.215]	−0.090 [−0.278, 0.066]
Pool	10%	All	0.150 [0.066, 0.233]	−0.007 [−0.090, 0.069]
Pool	10%	Accepted	0.117 [−0.001, 0.230]	−0.048 [−0.148, 0.083]
Pool	10%	Nonaccepted	0.131 [−0.008, 0.279]	−0.097 [−0.235, 0.028]
Pool	20%	All	0.124 [0.083, 0.168]	0.002 [−0.048, 0.049]
Pool	20%	Accepted	0.121 [0.053, 0.189]	−0.023 [−0.090, 0.047]
Pool	20%	Nonaccepted	0.088 [0.012, 0.191]	−0.062 [−0.163, 0.012]
Common	5%	All	0.118 [−0.015, 0.259]	−0.031 [−0.160, 0.102]
Common	5%	Accepted	0.091 [−0.083, 0.266]	−0.016 [−0.174, 0.146]
Common	5%	Nonaccepted	0.135 [−0.017, 0.287]	−0.096 [−0.241, 0.032]
Common	10%	All	0.150 [0.066, 0.233]	−0.007 [−0.090, 0.069]
Common	10%	Accepted	0.161 [0.032, 0.277]	−0.059 [−0.178, 0.067]
Common	10%	Nonaccepted	0.182 [0.073, 0.296]	−0.116 [−0.217, −0.009]
Common	20%	All	0.124 [0.083, 0.168]	0.002 [−0.048, 0.049]
Common	20%	Accepted	0.143 [0.066, 0.216]	−0.027 [−0.096, 0.047]
Common	20%	Nonaccepted	0.104 [0.042, 0.173]	−0.073 [−0.136, −0.009]

the complete coauthor network. Each draw supplies every pool, threshold and budget within its year. The common threshold is recomputed in the all-submission draw and shared with the accepted and nonaccepted subsets. This keeps its meaning common within a draw without treating the empirical threshold as externally fixed.

The focal comparison subtracts the accepted-pool ENS–TNS gain from the all-submission gain in 2023, using a common threshold and a 10% budget. A positive value would mean that ENS’s advantage over TNS is greater in the all-submission pool. Its cluster interval is [−0.026, 0.091]; the paper-bootstrap interval is [−.026, .094]. The corresponding simultaneous cluster band is [−0.115, 0.174]. These intervals neither establish a pool difference nor establish similarity within the retrospective reference of five additional high-uptake papers per 100 read. The reference is an interpretive benchmark, not an elicited user preference. All 12 simultaneous bands contain zero.

Table 12: 2023 complete selection sensitivity on the strict population. Intervals resample first-author clusters, preserving pool overlap and recomputing thresholds and score selections. All entries are exploratory marginal intervals.

Threshold	Budget	Pool	ENS—TNS [95% CI]	Rec.—ENS [95% CI]
Pool	5%	All	0.088 [−0.008, 0.188]	0.060 [−0.042, 0.184]
Pool	5%	Accepted	0.105 [−0.019, 0.244]	−0.019 [−0.154, 0.090]
Pool	5%	Nonaccepted	0.053 [−0.083, 0.197]	0.002 [−0.141, 0.164]
Pool	10%	All	0.101 [0.044, 0.168]	0.057 [−0.012, 0.136]
Pool	10%	Accepted	0.065 [−0.009, 0.155]	0.051 [−0.041, 0.138]
Pool	10%	Nonaccepted	0.090 [−0.011, 0.183]	0.034 [−0.066, 0.155]
Pool	20%	All	0.085 [0.053, 0.125]	0.024 [−0.019, 0.064]
Pool	20%	Accepted	0.092 [0.040, 0.149]	−0.007 [−0.062, 0.058]
Pool	20%	Nonaccepted	0.109 [0.036, 0.168]	0.000 [−0.062, 0.083]
Common	5%	All	0.088 [−0.008, 0.188]	0.060 [−0.042, 0.184]
Common	5%	Accepted	0.138 [0.003, 0.273]	−0.027 [−0.151, 0.120]
Common	5%	Nonaccepted	0.097 [−0.023, 0.219]	−0.088 [−0.215, 0.032]
Common	10%	All	0.101 [0.044, 0.168]	0.057 [−0.012, 0.136]
Common	10%	Accepted	0.071 [−0.016, 0.160]	0.061 [−0.032, 0.170]
Common	10%	Nonaccepted	0.085 [0.006, 0.167]	−0.002 [−0.095, 0.086]
Common	20%	All	0.085 [0.053, 0.125]	0.024 [−0.019, 0.064]
Common	20%	Accepted	0.105 [0.046, 0.168]	0.005 [−0.052, 0.077]
Common	20%	Nonaccepted	0.062 [0.016, 0.113]	0.024 [−0.038, 0.078]

For the population bounds, the original joint-label export supplies 3,843 score vectors; the expanded review cache adds 1,024 joint-label forums from the same source frame. On overlapping score-complete forums, all three cached aggregates agree to numerical tolerance. No new citation outcome is recovered. The union has 4,867 papers, with 3,443 verified-link outcomes and 1,424 unknown outcomes. The other 1,542 source forums lack complete score vectors in this union. The known outcome threshold stays fixed at the strict year-wide upper quartile; rankings and budget sizes are recomputed on the enlarged score-observed pool. Consequently these bounds and Table 3’s sampling intervals concern different populations and uncertainty questions.

Equation 3 is sharp conditional on these inputs: set each unknown binary outcome to one when its coefficient is negative to attain the lower endpoint, and to one when it is positive to attain the upper endpoint. This construction assumes no restrictions on unknown outcomes. The standard linear bound is an explanatory tool, not a claimed new theorem. Independent descending-score-group selection and linear-programming optimization reproduce all 12 interval endpoints. A synthetic enumeration verifies the expected precision under random boundary tie-breaking. Neither these checks nor positive lower endpoints validate unknown score records or identify a full-venue ordering.

C Reviewer Anchors and Secondary Evidence

The retained official-review responses use the same ordinal anchors for TNS and ENS: the lowest category denies both novelty and significance, the next denotes marginal novelty or significance, the third denotes significant contributions with some prior-work overlap, and the highest denotes significant contributions absent from prior work. ENS additionally allows a nonnumeric not-applicable response. Among the 7,181 review rows retained inside the export, 443 use that ENS response (245 in 2022; 198 in 2023). These are observed response texts, not a reconstructed complete historical invitation schema. Missing and not-applicable entries are excluded from numeric averages, so TNS and ENS can average different reviewers. The invitation-level instructions were not recovered. The main comparison is between the observed bundled fields and does not separately identify originality and significance.

The secondary tests address different questions: the 100-paper human study tests reproducibility of abstract judgments; the 90-paper corrected outcome subset explores citation association; pilot exclusion addresses composite-selection overlap; the 2022-to-2023 holdout tests temporal prediction within this selected venue; scorer and prompt replays test portability; the 295-paper arXiv cohort supplies a separate unconditional profile association. None is an independent replication of the conditional TNS/ENS contrast or a verification of scientific reuse.

Table 13: Outcome-blind illustrative selection, including concordant controls. Parentheses give the number of numeric reviews contributing to each mean. Descriptions characterize the text and do not establish why a reviewer assigned a score.

Paper	TNS / ENS	Claimed contribution in the abstract
On the Convergence of Certified Robust Training with Interval Bound Propagation	3.33 / 1.00 (3 / 1 reviews)	Analyzes convergence of interval-bound-propagation training for wide two-layer networks under a small perturbation radius.
Open-vocabulary Object Detection via Vision and Language Knowledge Distillation	2.67 / 3.67 (3 / 3 reviews)	Distills a vision-language teacher into an object detector and reports held-out-category and cross-dataset performance.
PAC-NeRF: Physics Augmented Continuum Neural Radiance Fields for Geometry-Agnostic System Identification	3.25 / 3.25 (4 / 4 reviews)	Combines radiance fields with continuum mechanics to infer geometry and material parameters from videos.
Learning Time-dependent PDE Solver using Message Passing Graph Neural Networks	2.00 / 1.50 (4 / 2 reviews)	Uses message-passing graph networks to learn time-dependent PDE solvers and transfer across domain geometries.

C.1 Holding contributing reviewers fixed

On 1,554 strict-linked papers with cached joint review fields, we average TNS, ENS and Recommendation over exactly the same numeric reviews (5,653 reviews; 1,460 first-author clusters). The original contributing counts differ on 262 papers. For fixed-window uptake, original TNS and ENS partial correlations are -0.034 and 0.161 ; their same-reviewer counterparts are -0.027 and 0.161 . The paired ENS-TNS gap is 0.188 ($[0.131, 0.246]$), compared with 0.194 originally. The change is -0.007 ($[-0.023, 0.009]$). Intervals use 2,000 global first-author-cluster draws, reranking and refitting both aggregations on each shared draw (seed 260929). This weakens the reviewer-composition explanation within the available-review subset; it does not remove its selective coverage or establish construct validity.

C.2 Illustrative score disagreement

The four examples are selected without outcomes: among strict-linked papers with cached reviews, take the smallest seeded forum hash in each of $\text{TNS} - \text{ENS} \geq 1$, $\text{ENS} - \text{TNS} \geq 1$, both scores ≥ 3 , and both scores ≤ 2 . The eligible group sizes are 48, 26, 293 and 67. Titles and descriptions below come from the cached abstracts; these are claimed contributions, not independently verified originality or reuse. No citation outcome determines their inclusion. The first example makes the reviewer-denominator issue concrete: two reviewers marked ENS not applicable, leaving one numeric ENS rating.

C.3 Fitted representations

Leave-one-out Ridge compares TO7 with TNS+ENS at fixed $\alpha = 10$, fitting standardization and an intercept within each training fold. Out-of-fold correlations are 0.498 and 0.321 ; their paired difference is 0.177 ($[0.139, 0.215]$). The interval resamples fixed out-of-fold predictions without nested refitting, so it omits full training-set uncertainty.

The temporal comparison trains on 2022 and tests on 2023. TF-IDF uses full titles and abstracts, at most 5,000 unigram/bigram features, minimum document frequency three and sublinear term frequency, with training-only vocabulary and IDF. Frozen SciBERT and SPECTER2 proximity-adaptor CLS embeddings use a 512-token limit. Every regression has an intercept and fixed $\alpha = 10$; TO7 standardization uses training data only. No corrected test outcome selects a model or hyperparameter. The comparison holds configuration fixed rather than optimizing each representation. Stronger retrospective prediction cannot distinguish content from writing style, visibility or memorized associations.

The historical holistic-novelty versus structured-significance comparison changed both the requested construct and its elicitation format. It therefore cannot isolate rubric structure. A crossed construct-by-format experiment would be needed for that attribution; this paper claims no format advantage.

Table 14: Corrected temporal holdout: 1,434 training papers from 2022, 2,006 test papers from 2023. All fits use Ridge $\alpha = 10$ and an intercept. TO7 standardization and TF-IDF vocabulary use training data only. Frozen embeddings are refitted through their regression heads. Paired intervals resample the fixed test predictions.

Representation	ρ	AP	TO7—row $\Delta\rho$ [95% CI]
SciBERT	0.289	0.420	0.228 [0.182, 0.278]
SPECTER2	0.398	0.493	0.119 [0.073, 0.165]
TF-IDF	0.444	0.514	0.073 [0.027, 0.116]
TO7	0.517	0.600	—

D Scoring Prompts and Dimension Definitions

Table 15 defines all nine scoring dimensions. Before prompt construction, Python string slicing limits the target abstract to 1,500 characters and each selected reference abstract to 600 characters. Table 16 reproduces both scoring templates verbatim from the scoring code, including the strict JSON response schema; the historical templates use the internal term “structured novelty” for what the paper calls the scientific-uptake profile. Immutable per-call endpoint-version manifests were not recorded for the historical run.

Table 15: Structured Scientific-Uptake Profile dimensions. Cond. marks scoring condition. *Both*: TO scores from title+abstract only; RG additionally compares to supplied reference abstracts. *RG only*: require references.

#	Dimension	Cond.	Definition
1	Problem Novelty	Both	TO: new task/question/domain claims in the abstract; RG: same relative to supplied references
2	Method Novelty	Both	TO: new algorithm/architecture/objective claims in the abstract; RG: same relative to references
3	Empirical Novelty	Both	TO: new datasets/benchmarks/findings claims in the abstract; RG: same relative to references
4	Recombination	Both	TO: non-obvious combination of ideas claimed in the abstract; RG: same relative to references
5	Reference Departure	RG only	Departs substantially from the closest supplied reference
6	Reference Dependence	RG only	Mainly depends on one or a few references (5=highly dependent)
7	Significance Potential	Both	Contribution appears broadly reusable or impactful from the available text
8	Evidence Strength	Both	Specificity of evidence reported in the available text for the stated claims
9	Clarity of Contribution	Both	Available text specifically and clearly states the claimed contribution

Table 16: Exact scoring templates reproduced from the scoring code. Left: target-only (TO) template used for all primary profile analyses. Right: reference-grounded (RG) template used for the secondary condition. Braces denote runtime substitution fields (the target abstract is pre-truncated to 1,500 characters and each reference abstract to 600); the JSON blocks are the required strict response schemas. The historical templates use the internal term “structured novelty” for the scientific-uptake profile.

You are evaluating a scientific paper for structured novelty. You must not decide acceptance. You must score only what is supported by the title and abstract.

Return strict JSON only.

Paper ID: {paper_id}
Title: {title}
Abstract: {abstract}

Score each dimension from 1 to 5:
1 = low, 3 = moderate, 5 = high.

Dimensions:

- problem_novelty: new task/question/domain/problem formulation
- method_novelty: new algorithm/architecture/objective/mechanism
- empirical_novelty: new datasets/benchmarks/experiments/findings
- recombination_novelty: combines distinct ideas or areas
- significance_potential: likely usefulness beyond this paper
- evidence_strength: how much evidence supports the novelty/significance claim
- clarity_of_contribution: how clearly the paper states its contribution

Return JSON:

```
{
  "paper_id": "...",
  "scores": {
    "problem_novelty": 1-5,
    "method_novelty": 1-5,
    "empirical_novelty": 1-5,
    "recombination_novelty": 1-5,
    "significance_potential": 1-5,
    "evidence_strength": 1-5,
    "clarity_of_contribution": 1-5
  },
  "short_rationale": "2-4 sentences",
  "uncertainty": "low|medium|high"
}
```

You are evaluating the novelty of a scientific paper relative to selected cited prior works.

Do not decide acceptance.

Do not reward broad claims unless the target text supports them.

Compare the target paper to the references, then score structured dimensions.

Return strict JSON only.

Paper ID: {paper_id}
Target title: {title}
Target abstract: {abstract}

Selected references:
{references_json}

Score each dimension from 1 to 5:
1 = low, 3 = moderate, 5 = high.

Definitions:

- problem_novelty: target addresses a new task/question/domain/problem formulation compared with references
- method_novelty: target introduces a new algorithm/architecture/objective/mechanism compared with references
- empirical_novelty: target provides new datasets/benchmarks/experiments/findings compared with references
- recombination_novelty: target combines distinct ideas or literatures in a non-obvious way
- reference_departure: target departs from the closest selected reference
- reference_dependence: target mainly depends on one/few references rather than standing independently
- significance_potential: target appears broadly reusable or important
- evidence_strength: target provides enough evidence to support its claimed novelty/significance
- clarity_of_contribution: target clearly states what is new

Also identify:

- closest_reference_ids: references the target most directly builds on
- most_distinct_reference_ids: references from which the target differs most

Return JSON:

```
{
  "paper_id": "...",
  "scores": {
    "problem_novelty": 1-5,
    "method_novelty": 1-5,
    "empirical_novelty": 1-5,
    "recombination_novelty": 1-5,
    "reference_departure": 1-5,
    "reference_dependence": 1-5,
    "significance_potential": 1-5,
    "evidence_strength": 1-5,
    "clarity_of_contribution": 1-5
  },
  "closest_reference_ids": [...],
  "most_distinct_reference_ids": [...],
  "short_rationale": "2-4 sentences grounded in target/reference comparison",
  "uncertainty": "low|medium|high"
}
```