ESTIMAND-ALIGNED NATURAL-LANGUAGE HYPOTHESIS DISCOVERY: PROBING WHAT THE ML COMMUNITY RECOGNIZES AS A CONTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Natural-language hypothesis discovery can reveal interpretable patterns in text, but the statistical objectives used to select them may not align with the population claims researchers seek to investigate. We introduce estimand-aligned natural-language hypothesis discovery, a framework in which a declared research design guides candidate evaluation and selection. Document-level measurements make predicates testable, and effect estimates with uncertainty quantify their support under the specified design. We develop a testbed for natural-language discovery about the machine-learning community, drawing on a decade of public submissions and peer reviews from a broad-scope venue, including accepted and rejected papers. We use this testbed to probe what the machine-learning community recognizes as a contribution. Across the transition to the LLM era, our temporal analysis suggests that enabling further research is gaining prominence as a claimed contribution through reusable datasets and integrated pipelines. Papers foreground theoretical contributions less, yet reviews express fewer novelty objections and more requests for specific forms of empirical support. Additional decision analysis finds that judgments about contribution clarity and evidential support improve acceptance prediction beyond a rating-and-year baseline. Accounting for rating-dependent associations reveals additional predictive value in novelty criticism. Together, the applications illustrate how an explicit statistical target shapes what can be learned about scientific contributions and their evaluation.

1 INTRODUCTION

Language models can propose thousands of plausible descriptions of patterns in text. Turning these descriptions into testable population claims requires specifying how they are measured and which statistical relationship is asserted. Natural-language discovery has developed through contrastive methods, which search natural-language properties or pairwise relations and evaluate distinctions in observed groups (Zhong et al., 2022; 2023; Xu et al., 2026), and model-based methods, which instead learn sets of predicate denotations as interpretable parameters of classification, clustering, or time-series models (Zhong et al., 2024). Another tradition starts from fixed, operationalized variables to serve as proxy and asks what can be learned under a research design, e.g model fit. It distinguishes marginal from adjusted comparisons with strict inferential discipline, but it starts from a closed variable set rather than an open language space. Those systems optimize a fixed contrastive or model-fit objective as estimand, however they could not make sure each predicate be tied to a declared, design-specific estimand to receive magnitude-sensitive simultaneous support.

We ask: **How can open-ended natural-language hypothesis search return a concise set of interpretable predicates that pass predeclared design diagnostics and receive independent statistical support?** We study *estimand-aligned natural-language hypothesis discovery*, which evaluates and returns a predicate together with its measurement, directional estimand, and the design diagnostics needed to interpret that estimand. The framework uses the declared estimand and design diagnostics to select candidates, then freezes the claims and measurement rules for independent evaluation. Effect estimates and uncertainty bounds quantify the support for each claim under the specified measurement and assumptions.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Evaluating natural-language discovery as a research tool requires settings in which researchers can pose different questions and examine competing explanations. Scientific papers and their reviews offer such a setting: they record what researchers pursue and how that work is assessed, while leaving the relevant semantic features to be discovered.

Recent conference discussions question whether peer review reflects the community's "collective taste in problems and solutions" (ICLR 2027 Program Chairs, 2026), given concerns about "theoryslop" and "slopterpretability" receiving high scores despite weak research questions (COLM 2026 Program Chairs, 2026). These concerns motivate examining what researchers present as contributions and how reviewers assess their value.

Since the analysis process faces interpretive pitfalls, for example, apparent changes can arise from existing trends or a changing mix of research, a clearly specified statistical target is essential for interpreting discovered patterns. We organize ten years of public submissions and peer review from a broad-scope machine-learning (ML) venue into a discovery testbed (Section 4.1). We illustrate its use through two main applications: how contributions and their evaluation have changed in the LLM era (Section 4.2), and which reviewer judgments add predictive information about acceptance beyond review scores (Section 4.3).

Under the specified text measurements, the temporal analysis suggests a growing emphasis on research-enabling contributions, alongside less theory-oriented framing and fewer explicit novelty objections. The predictive analysis identifies modest additional value in judgments about contribution clarity and evidential support beyond the specified score baseline. Allowing these associations to vary across rating levels changes which judgments receive statistical support.

Our contributions are:

- We formulate estimand-aligned natural-language hypothesis discovery, linking open-ended predicate search to a declared research design that governs candidate evaluation, selection, and confirmation.
- We develop a testbed for natural-language discovery about the machine-learning community, with tasks that specify how hypotheses are measured and independently evaluated, together with task-specific statistical targets, measurement protocols, and evaluation splits.
- Our analyses suggest a growing emphasis on research-enabling contributions and less theory-oriented framing, alongside fewer explicit novelty objections. Additional analysis further identifies modest predictive value in judgments about contribution clarity and evidential support beyond the specified score baseline.

2 Related Work

Hypothesis generation and assessment. Natural-language discovery includes contrastive and model-based approaches. Zhong et al. (2022) search for descriptions that distinguish two text distributions through a pairwise classification objective. Zhong et al. (2023) condition discovery on a researcher-specified goal and evaluate validity and relevance separately, using held-out data and false-discovery-rate correction in a reported application. Conditional hypothesis generation incorporates researcher-specified covariates to address within-stratum differences, sign reversals, and stratum imbalance (Xu et al., 2026). Model-based methods learn predicates as interpretable parameters: NLParam optimizes likelihood losses for classification, clustering, and time-series models (Zhong et al., 2024). Complementing generation, Gong & Fernandez (2025) elicit calibrated LLM priors over correlations to assess how surprising observed relationships are. Our framework uses a declared estimand and design diagnostics to guide predicate selection, then evaluates the frozen returned set through simultaneous bounds on directional effect magnitudes.

Conference and contribution analysis. Studies of scientific practice examine both research contributions and the processes through which they are evaluated. Hao et al. (2026) associate AI adoption with greater individual scientific impact and a narrower collective research focus. Turning to evaluation, Liang et al. (2024) estimate the prevalence of LLM-modified text in conference peer reviews. Public review resources support complementary analyses: Re^2 links initial submissions with reviews and multi-turn rebuttal discussions (Zhang et al., 2025), while Paper Copilot organizes conference records for studying review dynamics over time (Yang et al., 2026). Together, these studies

motivate examining how contributions are described and assessed in the research record. Our applications investigate changes in contribution descriptions and reviewer judgments under a temporal design, and the predictive value of review content beyond scores under a predictive-gain target.

3 METHOD

3.1 PROBLEM FORMULATION

Given a research question and a corpus, we seek natural-language properties that support testable population claims under a specified research design.

Definition 1. *A* predicate h *describes a document-level property in natural language. Its* measurement rule $f_h : \mathcal{X} \to \mathbb{R}$ *assigns a value to each document* $x \in \mathcal{X}$*. For binary measurement,* $f_h(x) \in \{0, 1\}$ *records whether the rule judges* x *to express* h*.*

A research design specifies the target population and the statistical relationship of interest. Applying this design to f_h *defines the* estimand θ_h*, the population quantity to be estimated;* $\hat{\theta}_h$ *denotes its estimate computed from data. A* direction $d_h \in \{-1, +1\}$ *records the asserted sign of* θ_h*. The corresponding* hypothesis *or claim, is*

$$H_h : \qquad d_h \theta_h(f_h) > 0.$$

For example, h may be "the abstract presents theoretical guarantees as a contribution," with $f_h(x)$ recording whether x is judged to express this property. If θ_h measures a deviation in this property's prevalence from an existing trend, $d_h = -1$ asserts a downward deviation.

The output is a set S of predicate, where each $h \in S$ is accompanied by its measurement rule f_h, estimand θ_h and its estimation procedure, and direction d_h. These specifications are *frozen*, meaning they remain unchanged during evaluation on independent data.

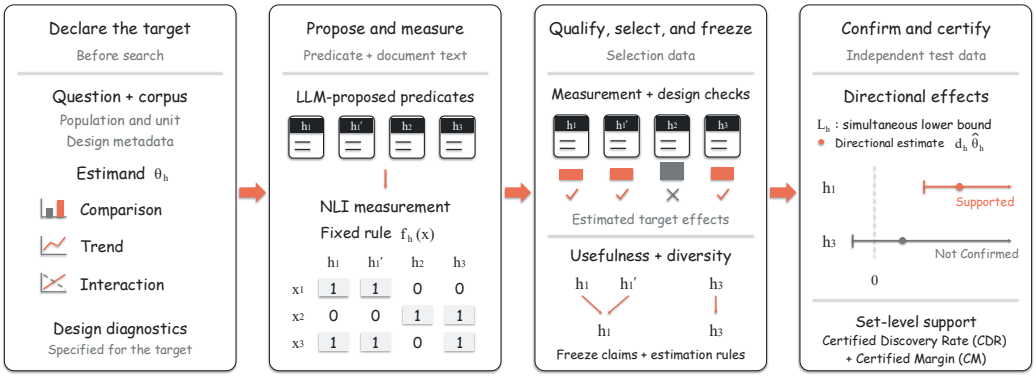


Figure 1: **Estimand-aligned natural-language hypothesis discovery.** A declared estimand guides candidate evaluation. Proposed predicates are measured and screened through design diagnostics, usefulness, and diversity. The selected claims and estimation rules are frozen before testing. Simultaneous lower bounds L_h quantify directional support and yield set-level lower bounds.

3.2 ESTIMAND-ALIGNED DISCOVERY

Our approach brings research design into open-ended natural-language discovery: the declared estimand and design-specific diagnostics *constrain* candidate selection and shape the returned discoveries. We operationalize this principle through the following four-stage procedure for constructing and evaluating population claims (Figure 1).

Declare the target. Given a research question, the researcher specifies the population quantity of interest (e.g., a comparison, a trend, or an interaction) and the design diagnostics required for its interpretation. This specification determines how θ_h will be defined for each subsequently proposed predicate h, establishing the target that guides candidate evaluation and selection.

Propose and measure. With the statistical target fixed, a proposal model searches for predicates h relevant to the research question. Evaluating these predicates under the declared design requires document-level measurements. NLI-based measurement supplies $f_h(x)$ from the predicate and document text alone, providing the observations from which θ_h can be estimated. Appendix C bounds recovery of known target facets. Appendix D analyzes how errors in f_h affect θ_h.

Qualify, select, and freeze. The measured candidates h can now be assessed for their target effects and their value to the research question. Selection data provide estimates $\hat{\theta}_h$ and are used to assess measurement reliability and the design conditions required to interpret θ_h. To form a concise set covering distinct aspects of the research question, candidates satisfying these requirements are ranked by a rule that combines estimated effects with scientific usefulness and accounts for redundancy, yielding S. The choice of output size $K = |S|$ balances discovery breadth against reading and confirmation costs. Once S is selected, each direction d_h is fixed under the declared convention. The predicate h, measurement rule f_h, estimand θ_h, and estimation procedure are also frozen, which ensures that confirmation evaluates the same claims $H_h : d_h\theta_h > 0$ that were selected. The frozen claims are then evaluated on independent data to avoid selection bias.

Confirm and certify. For the frozen set S, confirmation estimates the target effects and assesses support for each frozen claim H_h. The independent confirmation data yield an estimate $\hat{\theta}_h$and standard error $\hat{s}_h$ for each $h \in S$. Point estimates and confidence intervals describe the magnitude and precision of each effect. To summarize support across S, lower bounds L_h on the directional effects $d_h\theta_h$ are constructed with simultaneous coverage:

$$\Pr(d_h\theta_h \geq L_h \text{ for all } h \in S) \geq 1 - \alpha. \tag{1}$$

A Bonferroni construction gives

$$L_h = d_h\hat{\theta}_h - z_{1-\alpha/N}\hat{s}_h, \tag{2}$$

where z_p is the standard normal quantile and N counts the frozen claims in the declared confirmatory family. For this fixed family, asymptotically normal estimators and consistent standard errors give the coverage in Equation 1 asymptotically. A positive L_h supports H_h, with its magnitude quantifying the supported directional effect under the frozen measurement rule f_h.

These bounds also summarize support across the frozen set S. For an effect threshold $\delta \geq 0$, $\mathrm{CDR}(\delta)$ is the proportion of claims whose directional effects $d_h\theta_h$ exceed δ. $\mathrm{CM}(\delta)$ is the average positive excess above δ per returned claim. Comparisons depend on the chosen output size and selection rule. The corresponding simultaneous lower bounds are

$$\mathrm{CDR}^{\mathrm{LB}}(\delta) = \frac{1}{K}\sum_{h\in S}\mathbf{1}\{L_h > \delta\}, \qquad \mathrm{CM}^{\mathrm{LB}}(\delta) = \frac{1}{K}\sum_{h\in S}[L_h - \delta]_+ = \int_\delta^\infty \mathrm{CDR}^{\mathrm{LB}}(t)\,dt. \tag{3}$$

On the simultaneous coverage event in Equation 1, the following hold jointly for every $\delta \geq 0$:

$$\mathrm{CDR}^{\mathrm{LB}}(\delta) \leq \mathrm{CDR}(\delta), \qquad \mathrm{CM}^{\mathrm{LB}}(\delta) \leq \mathrm{CM}(\delta). \tag{4}$$

Both conclusions therefore inherit its joint $1 - \alpha$ coverage, with an asymptotic guarantee for the normal construction above. Appendix B gives the conditions and proof.

4 Empirical Studies

4.1 Discovery Testbed and Shared Procedure

The public record at ICLR brings together research across a broad range of machine-learning areas and publicly retains reviews of both accepted and rejected submissions (ICLR, 2026a;b), allowing us to study how research activity and peer evaluation evolve together. Starting from Re^2 (Zhang et al., 2025) and PRISM (Anonymous, 2026), we recover missing submissions and metadata from OpenReview and reconcile year-specific review formats and decision labels. We link abstracts, reviews, and decisions by paper, retaining scores and research-area metadata. The resulting panel spans 2017–2026 and contains 42,123 accepted and rejected submissions and 161,469 reviews; titles and abstracts are available for 41,903 submissions. Papers are assigned into proposal, validation, and test sets keeping all records from each paper together (Appendix A, Table 1).

We study two main applications: how contributions and their evaluation have changed in the LLM era (Section 4.2), and which reviewer judgments distinguish accepted from rejected work (Section 4.3). Each task defines what a hypothesis must establish and how its supporting evidence is measured. The accompanying procedures examine alternative explanations and evaluate hypotheses on held-out submissions. An exploratory third application in Appendix E examines whether research-topic popularity translates into publication success.

Following Section 3.1, X denotes a title–abstract pair or an individual review, and h describes a property expressed in that text. GPT-5.6-Terra proposes candidate descriptions from proposal-split examples, which are screened for semantic quality and redundancy. To examine these properties across the corpus, the original measurement rule uses Qwen3-8B to assign $f_h(X) \in \{0, 1\}$ from the text and predicate alone, without providing decision labels as separate inputs. Measurement prompts and comparisons across models are documented in Appendices J and F.

The research question determines how these measurements enter the estimand θ_h. This connects the meaning of a candidate h to the population claim being evaluated, as described in Section 3.2. The application sections define the relevant target and explain how the reported predicates relate to S. The direction d_h is interpreted with respect to that target. In particular, a judgment's association with acceptance and its incremental predictive value represent different claims. Sampling and selection details appear in Appendix A.

We report effect estimates with paper-clustered uncertainty and interpret the lower bounds L_h under the conditions in Section 3.2. These results remain conditional on the measurement rule f_h, whose potential errors are analyzed in Appendix D.

4.2 Application 1: Research and Reviewing in the LLM Era

Recent work shows that AI usage is associated with greater individual scientific impact as well as a narrower collective research focus (Hao et al., 2026). Meanwhile, evidence of LLM-assisted peer-review writing suggests that the record through which those contributions are evaluated is also changing (Liang et al., 2024). The growing role of AI in research and peer review raises a broader question: **What has changed about research contributions and their evaluation across the transition to the LLM era?**

A direct application of corpus-comparison methods such as Zhong et al. (2023) would search for predicates whose prevalence differs between the pre-2024 and post-2024 corpora: $\theta_h^{\text{raw}} = \mathbb{E}[f_h(X) \mid \text{post}] - \mathbb{E}[f_h(X) \mid \text{pre}]$, which identifies what differs on average between the two periods, but a large difference can arise from a trend already underway before LLM era. With documents weighted equally, it also gives greater weight within each period to years containing more texts. To investigate changes beyond the existing trend, we therefore need the temporal comparison to guide which candidates are selected and confirmed, and how the resulting findings are interpreted.

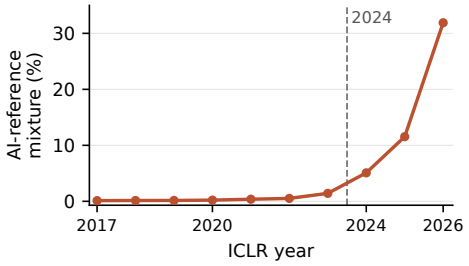


Figure 2: Review-text AI-reference mixture reaches 31.9% in 2026. Adapting Liang et al. (2024), we fit sentence-level word-occurrence mixtures by maximum likelihood.

We anchor the comparison at ICLR 2024, the first cycle whose submission and initial-review periods both followed ChatGPT's release. Review-text AI-reference mixture estimates also rise in 2024 (Figure 2), supporting this temporal division. To examine changes in both contribution claims and their evaluation, we analyze authors' stated title–abstracts and reviewers' expressed judgments in review texts, weighting each document equally, so years contribute in proportion to their included text counts. Let $Y_h = f_h(X) \in \{0, 1\}$, $U_t = t - 2023$, and $P_t = \mathbf{1}\{t \geq 2024\}$. We target the level shift from 2024 onward relative to a shared linear trend over 2017–2026:

$$\theta_h = \text{coef}_{P_t}(Y_h \sim 1 + U_t + P_t).$$

Here, θ_h is the gap between the fitted parallel trends, rather than the raw pre/post difference. To express the resulting comparison, let $\widetilde{P}_t = P_t - \text{Proj}(P_t \mid 1, U_t)$ denote the period indicator after removing its linear association with year. For the projection residual $\varepsilon_h = Y_h - \alpha_h - \beta_h U_t -$

$\theta_h P_t$, orthogonality gives $0 = \mathbb{E}[\widetilde{P}_t \varepsilon_h] = \mathbb{E}[\widetilde{P}_t Y_h] - \theta_h \mathbb{E}[\widetilde{P}_t^2]$, yielding the Frisch–Waugh–Lovell representation $\theta_h = \mathbb{E}[\widetilde{P}_t Y_h]/\mathbb{E}[\widetilde{P}_t^2]$. Equivalently, $\theta_h = \mathbb{E}[w_t Y_h]$, where $w_t = \widetilde{P}_t/\mathbb{E}[\widetilde{P}_t^2]$ satisfies $\mathbb{E}[w_t] = \mathbb{E}[w_t U_t] = 0$. Thus, a purely linear increase in the prevalence of h yields $\theta_h = 0$, preventing linear drift from being mistaken for an LLM-era level shift.

Generation uses proposal-split documents from 2019–2026; sampling and selection settings are given in Appendix A.2. On the proposal split, after deduplication, semantic screening, exclusion of candidates flagged for low label agreement, and a minimum validation prevalence of 0.5%, 1,619 review and 1513 abstract candidates leave. On validation data, 1,026 review and 393 abstract candidates exhibit significant 2024 breaks after BH correction.

Nonlinear drift may still produce an apparent break, motivating checks at earlier placebo cutoffs. On pre-2024 validation data, we refit the trend-break model at placebo cutoffs of 2020 and 2022. Requiring BH-adjusted $p > 0.05$ at both cutoffs retains 391 of 1,026 otherwise eligible review candidates with significant 2024 breaks, before effect-size and diversity selection. Requiring an absolute estimated effect of at least two percentage points further narrows these pools to 275 and 257. Diversity-based selection retains $K = 8$ predicates from each of the abstract and review candidates, whose directions are frozen before independent test-set confirmation.

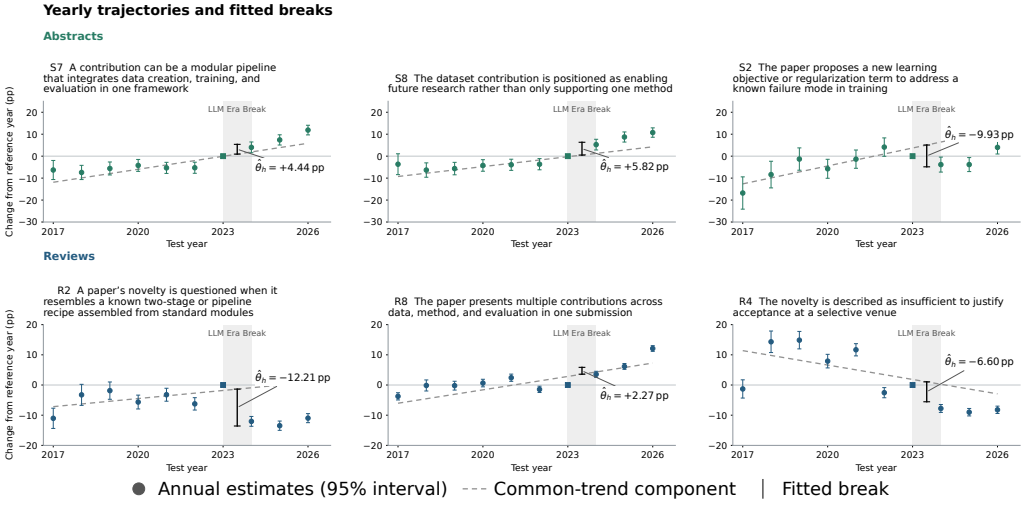


Figure 3: **Cases of annual trajectories and their fitted level shifts.** Estimates tested relative to the reference year, with pointwise 95% paper-clustered intervals.

To give the estimated effects an interpretation, Figure 3 places the fitted level shifts alongside annual trajectories for 6 tested predicates, providing a visual reference for the effect estimates and lower bounds. Figure 4 compares validation and test effects for the predicates. All 8 abstract and 8 review predicates have positive directional lower bounds on both splits, so these sets yield $\mathrm{CDR}^{\mathrm{LB}}(0) = 1$. The corresponding test $\mathrm{CM}^{\mathrm{LB}}(0)$ values are 2.29 pp for abstracts and 4.78 pp for reviews (Appendix G). Support extends beyond small effects: five abstract and four review predicates have lower bounds exceeding 2 pp. Substantively, the results suggest a growing emphasis on contributions that enable further research, alongside fewer explicit novelty objections in reviews.

Enabling further research gains prominence as a claimed contribution. The abstract results suggest a shift in how papers articulate their contribution. Claims explaining or advancing particular learning methods become less prominent (S1, S2, S3, S5, S6), alongside extensive comparisons with strong baselines (S4, -9.06 pp). Integrated research pipelines (S7, $+4.44$ pp) and datasets intended to support subsequent work (S8, $+5.82$ pp) gain prominence, emphasizing the research a contribution enables. Reviews more often recognize contributions spanning data, methods, and evaluation (R8, $+2.27$ pp), providing a complementary perspective on this broader contribution framing.

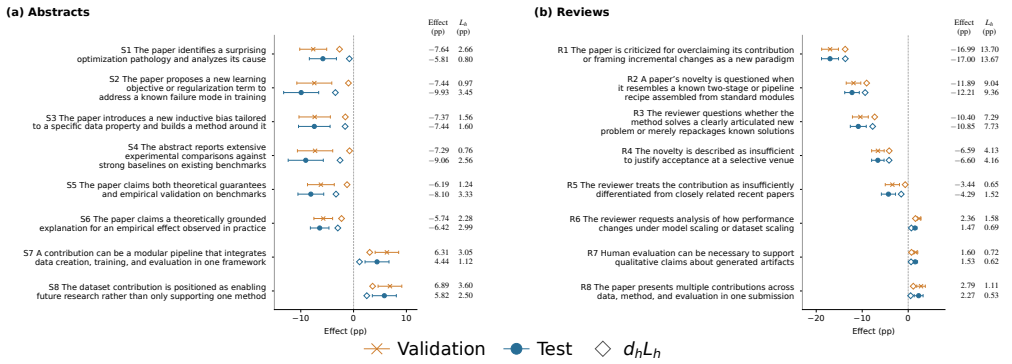


Figure 4: **Validation–test effects for contribution claims and reviewer judgments.** Abstract predicates S1–S8 (left) and review predicates R1–R8 (right). Measurement and selection details appear in Appendix G.

Papers foreground theoretical contributions less, yet reviews do not express more novelty objections. This shift in contribution framing coincides with fewer explicit objections to originality. Reviewers less often question whether a paper advances beyond existing work or overstates its contribution (R1-R5; -4.29 to -17.00 pp). At the same time, requests for human evaluation of generated outputs (R7, $+1.53$ pp) and analyses of model or dataset scaling (R6, $+1.47$ pp) increase. The review results thus point to a changing emphasis in expressed scrutiny, with less novelty-based criticism and greater attention to these specific forms of empirical support.

To assess whether these patterns depend on the measurement model, we tested Qwen3-8B, Qwen3-32B, and Mistral-Small-4-24B cross-model validation, which all predicates yield effects in the same direction with positive model-specific lower bounds on validation data (Appendix F), indicating that directional support persists despite differences in the underlying measurements.

It still leaves open whether changing review language contributes to the observed temporal shifts, since Figure 2 suggests review writing itself changed during the LLM era, potentially altering how judgments are expressed. To assess its sensitivity, adding the annual Liang et al. (2024) mixture estimate changes the frozen test effects by a median of 0.040 percentage points. All 95% confidence intervals exclude zero, and all family-adjusted directional lower bounds remain positive, suggesting these estimates are insensitive to this adjustment.

Which changes depend on research composition? Changes in the mix of submitted research may also contribute to these aggregate patterns. Adding broad research-category fixed effects leaves the decline in criticism of standard-module combinations largely unchanged (R2, -12.21 to -12.00 pp), while reducing the increase in requests for human evaluation (R7, $+1.53$ to $+1.04$ pp; adjusted 95% interval remains above zero). The rise in human-evaluation requests is therefore more sensitive to this composition adjustment than the decline in this particular novelty objection.

4.3 Application 2: Review Content Beyond Scores

Understanding which scientific contributions gain recognition requires examining the judgments behind acceptance decisions. Conference guidelines ask area chairs to consider substantive review content alongside numerical scores (ICML, 2025), and novelty is linked to acceptance after accounting for reviewer recommendations (Teplitskiy et al., 2022). We ask: **Which reviewer judgments add predictive value for acceptance beyond ratings?**

A comparison of accepted and rejected reviews can rediscover judgments already reflected in their ratings. Similar paper-average ratings also correspond to different acceptance rates across periods (Fig-

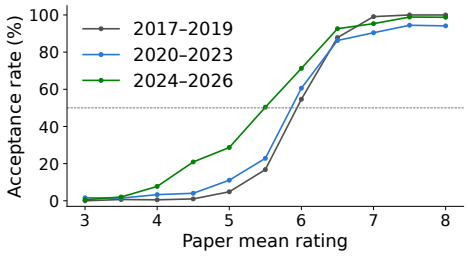


Figure 5: **Acceptance rates at similar mean ratings vary across periods.** Paper-average ratings are rounded to the nearest half point; each curve pools papers within a period, showing bins with at least 30 papers.

ure 5), while scoring scales and their verbal descriptions may vary across years. The question therefore requires a baseline specifying which rating and year information is already available. Estimating separate relationships within every year–score cell would fragment the sample; we use additive rating-bin and year terms to share information across cells. The archived score coding and its comparability limitations are detailed in Appendix H.

A judgment's additional predictive value is the improvement in acceptance prediction obtained by adding it to the rating-and-year baseline. Let A denote acceptance, $Y_h = f_h(X)$ indicate whether a review expresses judgment h, and $Z = (B, T)$ contain its rating bin and year. For predictors fitted on development data, define

$$\theta_h = \mathbb{E}[\ell\{A, \widehat{p}_0(Z)\} - \ell\{A, \widehat{p}_h(Z, Y_h)\} \mid \mathcal{D}_{\text{fit}}], \tag{5}$$

where $\ell(a, p) = -a \log p - (1 - a) \log(1 - p)$. Natural logarithms express the gain in nats per review; positive values indicate lower predictive loss after adding the judgment. This target measures the judgment's additional predictive value, accommodating associations with either acceptance or rejection.

The choice of log loss connects predictive gain to conditional mutual information. Its expectation is minimized by the true conditional probability (Gneiting & Raftery, 2007). With those probabilities, the gain equals $H(A \mid Z) - H(A \mid Z, Y_h) = I(A; Y_h \mid Z)$, the information about acceptance supplied by the judgment beyond ratings and year. With fitted predictors, the gain also depends on approximation and estimation error. Text can partly compensate for a poorly specified baseline, making the baseline's design central to interpreting θ_h.

A second design choice concerns how text relates to decisions across rating levels. A criticism could distinguish decisions among highly rated submissions while adding little among poorly rated ones; imposing a common coefficient can obscure this variation. Starting from the rating-and-year baseline (M_0), we add either a common text effect (M_1) or a text effect that varies by rating bin (M_2). The primary gain θ_h compares M_0 with M_2; comparing M_1 with M_2 tests whether the more flexible specification improves prediction using the same information. Regularization limits fitting variance, and gains are evaluated on held-out papers with paper-clustered uncertainty. The fitting details appear in Appendix H.

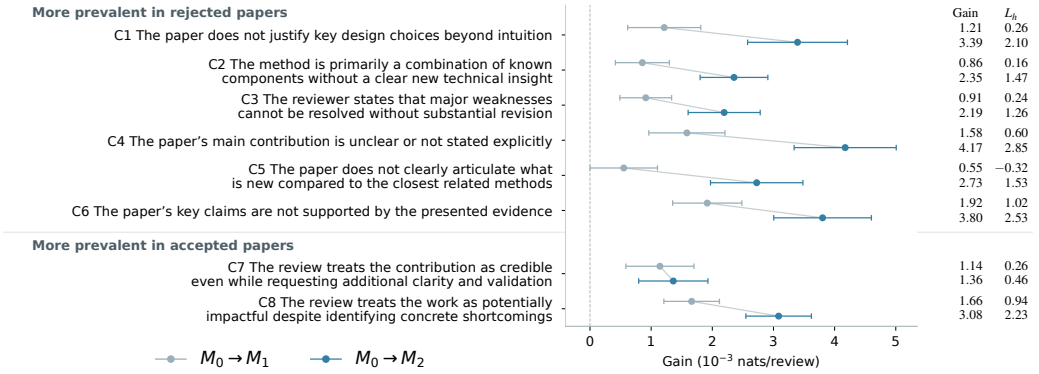


Figure 6: **Predictive gains under common and rating-specific text effects.** Light: $M_0 \to M_1$; blue: $M_0 \to M_2$, with pointwise 95% paper-clustered intervals. Columns report gains and corrected L_h in 10^{-3} nats/review; upper and lower entries match the points.

Figure 6 compares both specifications for the eight displayed predicates: six rejection-associated and two acceptance-associated under the archived prevalence analysis. Allowing rating-specific associations raises the fraction of displayed predicates with positive corrected lower bounds from seven of eight to all eight, while increasing the average positive lower bound from 0.43 to 1.80 in 10^{-3} nats per review (Appendix H.5). Seven also have positive bounds for the additional gain from rating-bin variation. These gains are relative to the archived review-rating baseline. Numerical gains and bounds below are in 10^{-3} nats per review.

Contribution clarity and evidential support carry predictive value beyond scores. An unclear main contribution (C4) yields the largest displayed gain, 4.17 ($L_h = 2.85$), followed by unsup-

ported key claims (C6, 3.80) and unjustified design choices (C1, 3.39). The largest gain reduces baseline log loss by 0.74%, indicating a modest but detectable contribution beyond ratings. Value also persists for qualified endorsements: potential impact despite shortcomings (C8, 3.08) and a credible contribution despite requests for clarification and validation (C7, 1.36). Review text thus retains evaluative distinctions beyond the rating bins, including favorable assessments qualified by substantive criticism.

A common text effect can obscure the predictive value of novelty criticism. For failure to articulate novelty relative to related work (C5), a common effect yields a gain of 0.55 with $L_h = -0.32$. Allowing rating-bin variation raises total gain to 2.73 with $L_h = 1.53$; the paired additional gain is 2.17 with $L_h = 1.43$. Thus, the chosen specification changes whether this judgment has supported predictive value under the declared correction. The benefit is not uniform: for the qualified credibility judgment (C7), the additional gain is only 0.22, with a lower bound below zero. These comparisons support accommodating rating-dependent associations without assuming that every predicate benefits from that flexibility.

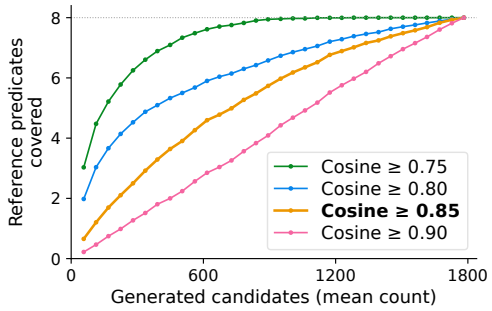


Figure 7: **Coverage depends on the matching threshold.** Mean coverage of eight reference predicates over 200 random orderings of generation calls.

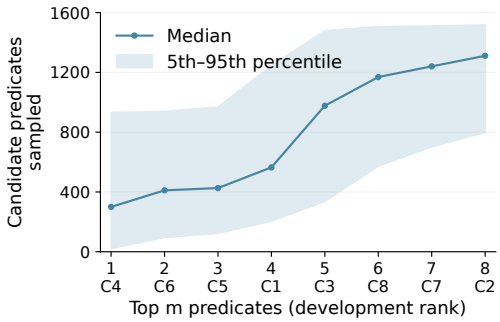


Figure 8: **Candidate budget for the development-ranked top-m.** Median and 5th–95th percentiles over 200 random candidate orderings, at similarity 0.85.

The predictive results raise a separate question about discovery: *How much candidate generation is needed to encounter these useful judgments?* A growing pool may contain repeated phrasings of the same finding, so candidate count alone does not establish coverage. Figure 7 tracks recovery of the eight reference predicates as generation accumulates and varies the similarity threshold defining a match. Loose matching reaches full coverage earlier, while stricter matching continues to recover reference predicates near the observed budget limit. Apparent saturation therefore depends on what counts as the same finding.

The required budget also depends on whether the aim is to recover a few high-value findings or a broader set. Figure 8 fixes similarity at 0.85 and ranks the eight predicates by development predictive gain. Within the archived pool, covering the top three takes a median of 426 sampled candidates, compared with 1,311 for all eight. These diagnostics assess recovery of a fixed observed set; they do not establish completeness of discovery.

5 Conclusion

We presented a framework that makes the target population claim explicit throughout natural-language hypothesis discovery, linking each discovery to its measurement, statistical target, and uncertainty. Its applications to public research records suggest changes in what authors foreground as contributions and what reviewers scrutinize: research-enabling contributions gain prominence, while reduced theoretical framing coexists with fewer explicit novelty objections. Review judgments also retain evaluative distinctions with modest predictive value beyond the specified score baseline. These observations motivate examining contribution framing and expressed evaluation criteria together. The framework supports this inquiry by making each discovery's measurement, statistical target, and uncertainty explicit.

AI USE STATEMENT

We used GPT-5.6 Sol and GPT-6 Astra to assist with language editing and improve the clarity of the manuscript, and Claude Fable 5 to assist with code development. All AI-assisted text and code were reviewed by the authors, who take full responsibility for the final content.

REFERENCES

Anonymous. PRISM: A multi-dimensional benchmark for evaluating LLM peer reviewers. Hugging Face dataset, 2026. URL https://huggingface.co/datasets/anoyresearcher/prism_paper_data. Under review; author identity withheld for double-blind review.

COLM 2026 Program Chairs. AI submissions at COLM: theoryslop, slopterpretability, and papers in the age of agents. COLM 2026 Program Chairs' Blog, July 2026. URL https://gregdurrett.github.io/colm2026-blog/ai-papers.html.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Yue Gong and Raul Castro Fernandez. Exploiting LLMs for automatic hypothesis assessment via a logit-based calibrated prior. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=D5TSSVkTBA.

Qianyue Hao, Fengli Xu, Yong Li, and James Evans. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, 649:1237–1243, 2026. doi: 10.1038/s41586-025-09922-y. URL https://www.nature.com/articles/s41586-025-09922-y.

ICLR. ICLR 2026 author guide, 2026a. URL https://iclr.cc/Conferences/2026/AuthorGuide.

ICLR. ICLR 2026 call for papers, 2026b. URL https://iclr.cc/Conferences/2026/CallForPapers.

ICLR 2027 Program Chairs. Submission policies for ICLR 2027. ICLR Blog, September 2026. URL https://blog.iclr.cc/2026/09/02/submission-policies-for-iclr-2027/.

ICML. ICML 2025 area chair instructions, 2025. URL https://icml.cc/Conferences/2025/AreaChairInstructions.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel Mcfarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29575–29620. PMLR, 2024. URL https://proceedings.mlr.press/v235/liang24b.html.

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. URL https://doi.org/10.1073/pnas.2118046119.

Paiheng Xu, Jing Liu, and Wei Ai. Conditional hypothesis generation for llm-based text analysis with researcher-specified covariates, 2026. URL https://arxiv.org/abs/2606.03029.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in AI conferences. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=CyKVrhNABo.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. Re^2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions, 2025. URL https://arxiv.org/abs/2505.07920.

Ruiqi Zhong, Charlie Snell, Dan Klein, and Jacob Steinhardt. Describing differences between text distributions with natural language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 27099–27116. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhong22a.html.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. Goal driven discovery of distributional differences via language descriptions. In *Advances in Neural Information Processing Systems*, volume 36, pp. 40204–40237. Curran Associates, Inc., 2023. doi: 10.52202/075280-1748. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/7e810b2c75d69be186cadd2fe3febeab-Abstract-Conference.html.

Ruiqi Zhong, Heng Wang, Dan Klein, and Jacob Steinhardt. Explaining datasets in words: Statistical models with natural language parameters. In *Advances in Neural Information Processing Systems*, volume 37, pp. 79350–79380. Curran Associates, Inc., 2024. doi: 10.52202/079017-2520. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/90c4537a301e9545bb4c60219f2992b1-Abstract-Conference.html.

A Data Splits and Implementation Details

A.1 Paper-level splits

The shared split assigns papers within year×decision strata using seed 89 and proposal/validation/test proportions of 30/35/35. All reviews and metadata from the same paper remain in the same split. Table 1 gives the resulting document counts.

	Proposal	Validation	Test
Reviews	48,376	56,610	56,483
Papers underlying reviews	12,617	14,744	14,762
Title+abstracts	12,549	14,678	14,676

Table 1: Original proposal, validation, and test assignments.

A.2 Application 1: Proposal and selection settings

These settings describe the proposal and original selection procedure in Section 4.2. Here θ_h^{break} denotes its temporal target θ_h. Reviews and title–abstracts are analyzed separately.

Proposal. The proposal contrast is post-2024 versus pre-2024. The generator is shown proposal-split examples from 2019–2026, so its displayed pre-period examples come from 2019–2023. For each text source we make 40 proposal calls (20 under each presentation order). Review prompts contain 150 examples per contrast group and title–abstract prompts contain 250.

Validation screening uses 13,390 reviews (at most 1,500 per year) and 14,678 title–abstracts.

A candidate is eligible on validation data only if

$$p^{\text{BH}}_{h,\text{break}} \le 0.05, \qquad |\widehat{\theta}_h^{\text{break}}| \ge 0.02, \qquad \widehat{\Pr}(Y_h = 1) \ge 0.005,$$

passes the four shared semantic gates, is not excluded by the sampled calibration audit, and shows no BH-significant placebo break at 2020 or 2022.

Eligible predicates are ranked greedily by $|\widehat{\theta}_h^{\text{break}}|$ subject to pairwise Qwen-embedding cosine similarity below 0.82; when this constraint cannot be satisfied, we add the candidate least similar to the current set. Direction-specific quotas follow the eligible positive and negative pools. Appendix G specifies the measurement and inference scope.

A.3 Application 2: Proposal settings

The predictive analysis reports eight predicates. Appendix H gives their selection provenance, predictive models, and all three comparisons.

Proposal. The generation record used by the coverage diagnostic contains 32 completed calls and 1,782 raw predicates, with 1,544 retained pool representatives after deduplication. Calls used 100 or 250 examples per decision group under both presentation orders.

Predictive model development uses 7,906 reviews from 5,316 papers. The evaluation archive contains 56,483 reviews from 14,762 papers, with missing labels excluded separately for each predicate.

B Set-Level Confirmatory Support

Frozen family and simultaneous coverage. Let $\mathcal{G}$ contain all information used before confirmation. Conditional on $\mathcal{G}$, the returned set S, directions d_h, measurement rules f_h, estimands θ_h, and estimators are fixed; the confirmation sample is independent under the sampling design. The declared family $\mathcal{F} \supseteq S$ has size N, and $K = |S| \geq 1$. When only S is tested, $N = K$; a joint method comparison instead counts method–claim pairs. Effects summarized together share a common scale.

If $\Pr(L_h > d_h\theta_h \mid \mathcal{G}) \leq \alpha/N$ for each $h \in \mathcal{F}$, a union bound gives

$$\Pr(\mathcal{E}_{\mathcal{F}} \mid \mathcal{G}) \geq 1 - \alpha, \qquad \mathcal{E}_{\mathcal{F}} = \bigcap_{h \in \mathcal{F}} \{d_h\theta_h \geq L_h\}. \tag{6}$$

This allows dependence between claims. Averaging over $\mathcal{G}$ gives the unconditional guarantee in Equation 1.

For the normal bounds in Equation 2, assume conditional studentized normality,

$$\frac{d_h(\hat{\theta}_{h,n} - \theta_h)}{\hat{s}_{h,n}} \,\Big|\, \mathcal{G} \xrightarrow{d} \mathcal{N}(0,1), \tag{7}$$

for almost every $\mathcal{G}$, with consistent standard errors for the sampling design. Here n counts independent sampling units; clustered analyses require cluster-aware standard errors. Each tail error is then $\alpha/N + o(1)$, so for a fixed finite family,

$$\liminf_{n\to\infty} \Pr(\mathcal{E}_{\mathcal{F},n} \mid \mathcal{G}) \geq 1 - \alpha. \tag{8}$$

Thus the normal construction supplies asymptotic coverage; finite-sample coverage requires valid finite-sample marginal bounds.

Set-level summaries. For $\delta \geq 0$ and $[a]_+ = \max(a, 0)$, define

$$\mathrm{CDR}(\delta) = \frac{1}{K}\sum_{h\in S} \mathbf{1}\{d_h\theta_h > \delta\}, \qquad \mathrm{CM}(\delta) = \frac{1}{K}\sum_{h\in S} [d_h\theta_h - \delta]_+, \tag{9}$$

$$\mathrm{CDR}^{\mathrm{LB}}(\delta) = \frac{1}{K}\sum_{h\in S} \mathbf{1}\{L_h > \delta\}, \qquad \mathrm{CM}^{\mathrm{LB}}(\delta) = \frac{1}{K}\sum_{h\in S} [L_h - \delta]_+. \tag{10}$$

The denominator includes every returned claim. Unthresholded CM abbreviates $\mathrm{CM}(0)$, and likewise for its lower bound.

Proposition 1 (Simultaneous set-level support)**.** *On $\mathcal{E}_{\mathcal{F}}$, jointly for every $\delta \geq 0$,*

$$\mathrm{CDR}^{\mathrm{LB}}(\delta) \leq \mathrm{CDR}(\delta), \qquad \mathrm{CM}^{\mathrm{LB}}(\delta) \leq \mathrm{CM}(\delta). \tag{11}$$

These inequalities inherit the coverage in Equation 6, or its asymptotic counterpart in Equation 8.

Proof. On $\mathcal{E}_{\mathcal{F}}$, $d_h\theta_h \geq L_h$ for every $h \in S$. Monotonicity gives $\mathbf{1}\{L_h > \delta\} \leq \mathbf{1}\{d_h\theta_h > \delta\}$ and $[L_h - \delta]_+ \leq [d_h\theta_h - \delta]_+$. Averaging proves both inequalities for all thresholds on the same event. □

The identity $[a - \delta]_+ = \int_\delta^\infty \mathbf{1}\{a > t\}\, dt$ also gives

$$\mathrm{CM}(\delta) = \int_\delta^\infty \mathrm{CDR}(t)\, dt, \qquad \mathrm{CM}^{\mathrm{LB}}(\delta) = \int_\delta^\infty \mathrm{CDR}^{\mathrm{LB}}(t)\, dt. \tag{12}$$

Interpretation. CDR is a proportion; CM retains the effect units. Both summarize support under the frozen measurement and design. Comparisons require aligned tasks, measurement protocols, sample sizes, selection rules, output sizes, and confirmatory families. Semantic validity and scientific usefulness require separate evidence; Appendix D describes the measurement boundary.

C Proposal Coverage in Open-Ended Search

Coverage can be quantified relative to a known target set in a controlled benchmark. A fixed-budget proposal run recovers target $h_j^\star$ when it returns a predicate satisfying a prespecified semantic matching rule. Suppose the probability of recovery on each run, conditional on any preceding failures, is at least π_j.

Proposition 2 (Recovery under repeated proposal)**.** *After R runs, for J known targets,*

$$\Pr(h_j^\star \textit{ is recovered}) \geq 1 - (1 - \pi_j)^R, \tag{13}$$

$$\Pr(\textit{all } J \textit{ targets are recovered}) \geq 1 - \sum_{j=1}^{J} (1 - \pi_j)^R. \tag{14}$$

Proof. The probability of R consecutive misses is at most $(1 - \pi_j)^R$. A union bound over missed targets gives the joint result. □

The bound is conditional on the stated recovery probabilities and matching rule. Estimating the probabilities requires repeated runs and accounting for estimation uncertainty. An open scientific question has no known exhaustive target set. The coverage diagnostics in Figures 7 and 8 therefore measure recovery of a fixed observed set, without establishing discovery completeness or estimating the probabilities in this proposition.

D Measurement Error in Predicate Operationalization

The main-text bounds concern effects defined by the frozen measurement rule f_h. The following identities explain how automated measurement can change a binary prevalence contrast. They are theoretical sensitivity relations; the reported studies have no blinded human annotations for estimating the calibration quantities.

Prevalence correction. Let $Y_h \in \{0, 1\}$ be the semantic label defined by an annotation rubric and $\widetilde{Y}_h = f_h(X)$ its automated measurement. In group g, write $p_{h,g} = \Pr(Y_h = 1 \mid G = g)$ and $\widetilde{p}_{h,g} = \Pr(\widetilde{Y}_h = 1 \mid G = g)$. Define sensitivity $\mathrm{Se}_{h,g} = \Pr(\widetilde{Y}_h = 1 \mid Y_h = 1, G = g)$, specificity $\mathrm{Sp}_{h,g} = \Pr(\widetilde{Y}_h = 0 \mid Y_h = 0, G = g)$, and $J_{h,g} = \mathrm{Se}_{h,g} + \mathrm{Sp}_{h,g} - 1$. Total probability gives

$$\widetilde{p}_{h,g} = J_{h,g} p_{h,g} + 1 - \mathrm{Sp}_{h,g}, \qquad p_{h,g} = \frac{\widetilde{p}_{h,g} + \mathrm{Sp}_{h,g} - 1}{J_{h,g}} \quad (J_{h,g} > 0). \tag{15}$$

Calibration must represent the target predicate and group. Correction becomes unstable as $J_{h,g}$ approaches zero.

Effect on a group contrast. Let $\theta_h^\star = p_{h,1} - p_{h,0}$ and $\widetilde{\theta}_h = \widetilde{p}_{h,1} - \widetilde{p}_{h,0}$. If sensitivity and specificity are equal across groups, then

$$\widetilde{\theta}_h = J_h \theta_h^\star, \qquad J_h = \mathrm{Se}_h + \mathrm{Sp}_h - 1. \tag{16}$$

For $0 < J_h \leq 1$, the observed contrast preserves the sign and has at most the magnitude of the semantic contrast. With group-dependent errors,

$$\widetilde{\theta}_h = J_{h,1} p_{h,1} - J_{h,0} p_{h,0} + \mathrm{Sp}_{h,0} - \mathrm{Sp}_{h,1}, \tag{17}$$

so measurement can create or reverse a contrast. Group-specific calibration is needed to recover $\theta_h^\star$ by correcting each prevalence in Equation 15. These identities concern unadjusted prevalence differences; the temporal and predictive targets in the applications require their own measurement-error analyses.

Uncertainty and support. A calibrated analysis would propagate both sampling and calibration uncertainty, for example by jointly resampling the confirmation and audit data under their sampling and dependence structure. If the resulting bounds $L_h^\star$ satisfy $\Pr(d_h\theta_h^\star \geq L_h^\star \text{ for all } h \in S) \geq 1-\alpha$, the set-level proof in Appendix B applies with $L_h^\star$. The empirical bounds reported here remain conditional on the automated measurements; cross-model agreement alone supplies no human-calibrated guarantee.

E Exploratory Analysis: Relative Acceptance Gaps across Research Topics

E.1 Target and measurement

This application examines how research topics' relative acceptance rates vary over time. Following the target–measurement distinction in Section 3.2, topic membership supplies the document-level variable, annual acceptance-rate gaps describe its relation to acceptance, and endpoint contrasts quantify specified changes.

The recorded panel contains 42,123 accepted and rejected ICLR submissions from 2017–2026; 41,902 have measurable title–abstract text. For each of 28 case-insensitive phrase groups h, $f_h(X) \in \{0,1\}$ records a dictionary match. Groups overlap, and matches can refer to baselines or related work. The resulting quantities describe phrase-defined groups among recorded decided papers.

Let A denote acceptance and T year. Define the population acceptance gap, in percentage points (pp), by

$$g_{h,t} = 100\{\Pr(A=1 \mid f_h(X)=1, T=t) - \Pr(A=1 \mid f_h(X)=0, T=t)\}.$$

For endpoints $a < b$, the target is $\theta_h^{a,b} = g_{h,b} - g_{h,a}$; a declining gap corresponds to $d_h = -1$. Comparing matching with non-matching papers keeps the reference explicit: comparison with the whole-conference rate would multiply the gap by the non-matching share.

E.2 Comparison scope and estimation

The fixed dictionary supplies all 28 predicates. There is no generated candidate pool, validation-ranked returned set, or independent confirmation split in this extension.

The endpoint test family contains 45 comparisons ending in 2026: 20 groups with at least 30 matches in 2023 and 25 with at least 30 in 2024. Holm-adjusted p values account for these 45 comparisons; earlier choices of dictionary, windows, and analyses remain outside that adjustment. Within-year gaps use pointwise 95% Newcombe intervals from Wilson proportion intervals; endpoint changes use pointwise 95% independent-paper Wald intervals. Results are exploratory estimates and do not instantiate the independent-confirmation guarantee in Section 3.2.

E.3 Annual gaps and supplementary endpoint comparisons

Figures 9 and 10 show all 28 groups over 2017–2026. Each point estimates $g_{h,t}$; the trajectories provide descriptive context for temporal comparisons, without selecting a common start year or testing a trend from the plotted shape.

Endpoint comparisons. The LLM phrase group's estimated acceptance gap falls from $+19.17$ pp in 2023 to $+0.13$ pp in 2026. The estimated change is -19.04 pp, with pointwise 95% interval $[-30.15, -7.93]$ and Holm-adjusted $p = 0.034$. The 2024–2026 comparison also shows a decline (Table 2). These contrasts support a narrowing relative acceptance advantage within the recorded panel; changes in authors, tasks, and topic composition remain possible explanations.

Scope across topics. The 2023–2026 relative-gap declines for GNN and Diffusion have Holm-adjusted $p = 0.541$ and $p = 0.273$, respectively. Their negative point estimates therefore have weaker support after adjustment across the 45 endpoint comparisons.

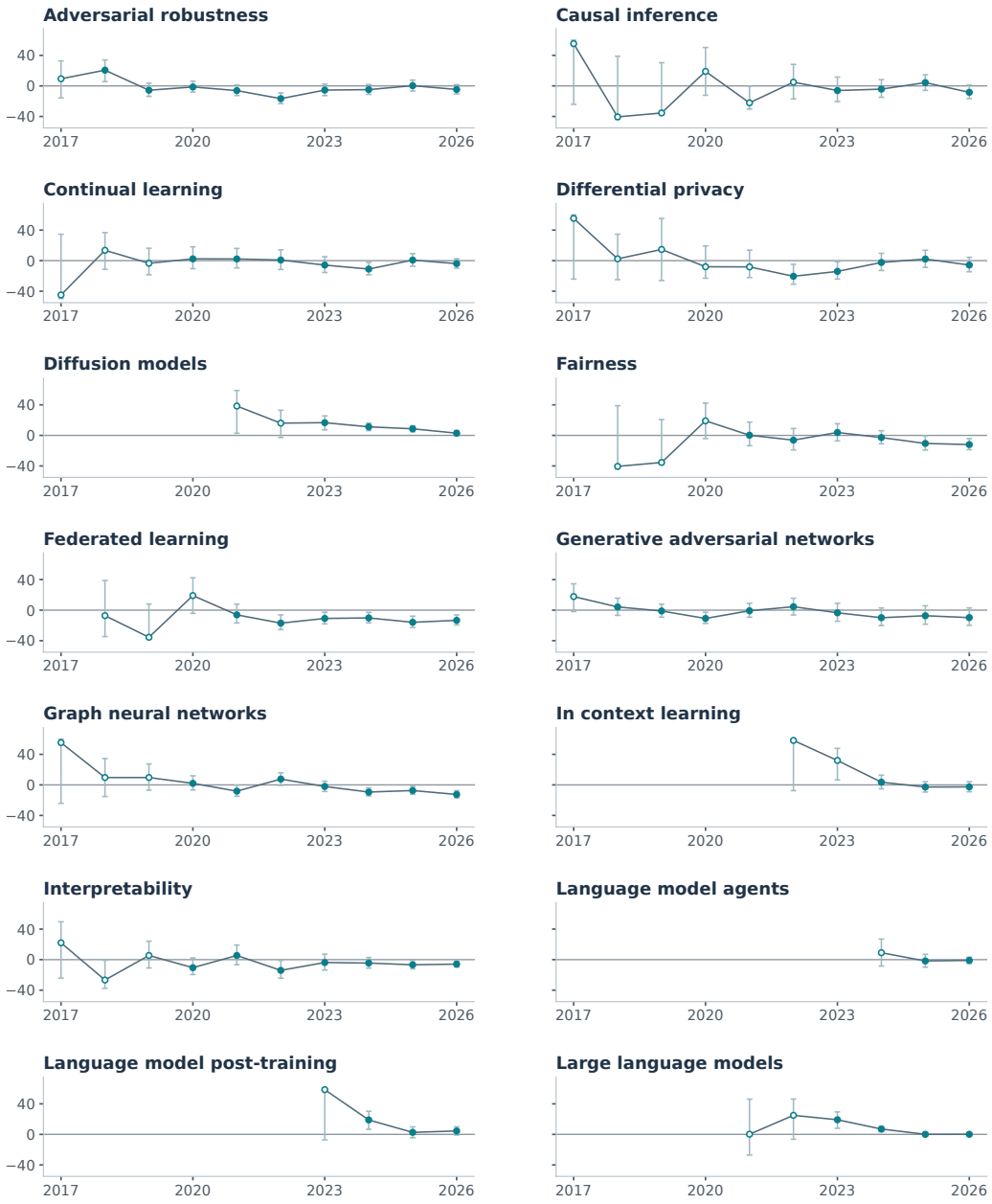


Figure 9: **Annual relative acceptance gaps by research topic (1/2).** Horizontal axes show year; vertical axes show $\widehat{g}_{h,t}$ in percentage points, comparing phrase-matching papers with same-year non-matching papers. All panels share the same axes; zero indicates equal acceptance rates. Bars are pointwise 95% Newcombe intervals from Wilson limits under the independent-paper working model. Hollow points indicate fewer than 30 matching papers; years with no matches are blank. The two figures include all 28 overlapping phrase groups in alphabetical order and use available-text denominators; 217 decided papers lack text in 2025.

E.4 Measurement checks and interpretation

Text is unavailable for 221 recorded papers, including 217 in 2025 and none in 2026. Annual gaps use available-text denominators, and phrase-match errors remain uncalibrated.

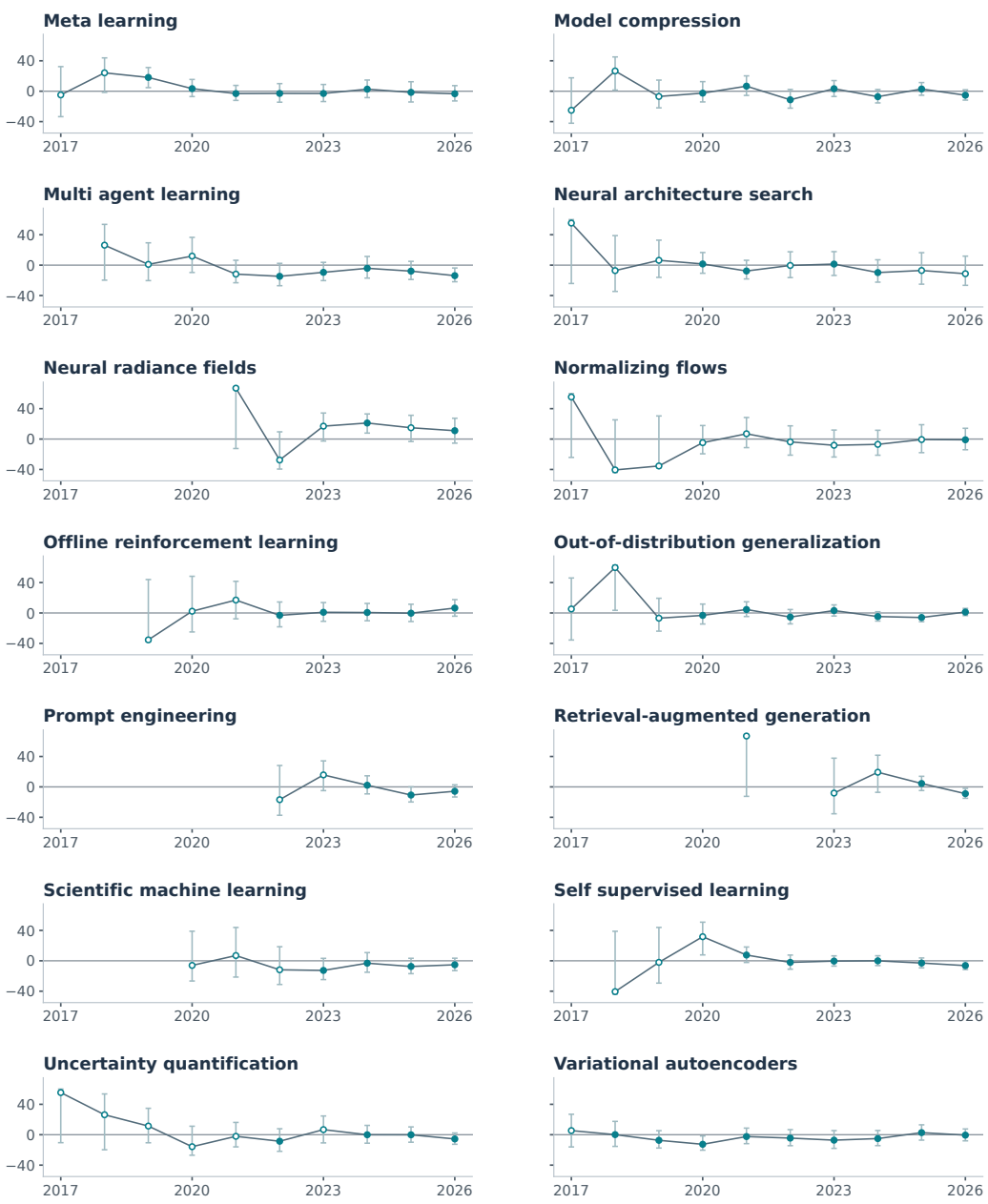


Figure 10: **Annual relative acceptance gaps by research topic (2/2).** Continuation of Figure 9, with the same axes and interval construction. Horizontal axes show year; vertical axes show $\widehat{g}_{h,t}$ in percentage points. Hollow points indicate fewer than 30 matching papers; years with no matches are blank.

LLM comparison	Gap change (pp)	Pointwise 95% CI	Holm p
2023–2026	-19.04	$[-30.15, -7.93]$	0.034
2024–2026	-6.88	$[-10.85, -2.90]$	0.032

Table 2: LLM endpoint-gap comparisons. The adjusted p values use the joint 45-comparison family; the intervals are pointwise. The 2026 gap itself has interval $[-1.59, 1.85]$ pp, leaving uncertainty about a small advantage or disadvantage at that endpoint.

Full-name and acronym variants retain the direction of the 2023–2026 LLM gap decline, with varying precision. Overlapping membership, changing terminology, and the working independent-paper variance limit interpretation. The analysis describes topic-level publication outcomes; individual entry decisions and causal effects of research crowding remain outside its design.

F Application 1: Cross-Model Measurement of Abstract Predicates

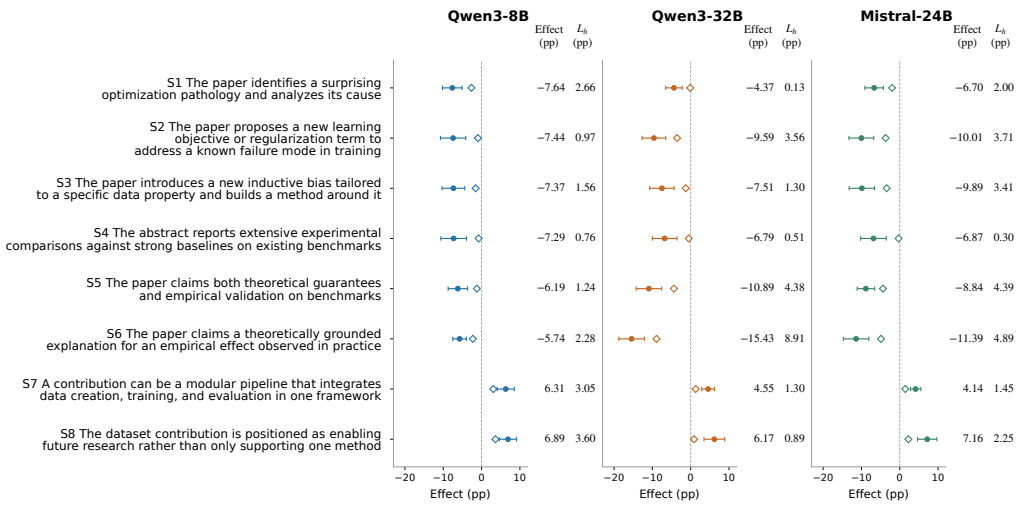


Figure 11: **Cross-model effects for eight abstract predicates.** Qwen3-8B, Qwen3-32B, and Mistral-24B measurements on 14,678 validation abstracts under the clarified rules. Dots are signed common-slope 2024-break estimates, bars are pointwise 95% paper-clustered intervals, and open diamonds mark $d_h L_h$. Separate numeric columns give $\hat{\theta}_h$ and L_h in percentage points. Appendix G explains the measurement and inference scope across figures.

G Measurement and Inference Scope of the Reported Results

Application 1. All trajectories and their fitted shifts in Figure 3 use test measurements. The original-protocol effects quoted in the findings paragraphs and the rule-based sensitivity estimates use different measurements, even when predicate identifiers coincide. Blinded human annotations have yet to be collected, so agreement across models provides a measurement diagnostic rather than an estimate of semantic accuracy.

Application 1 set-level summaries. For each displayed set in Figure 4, $K = 8$. At $\delta = 0$, all eight lower bounds are positive on both validation and test, giving $\mathrm{CDR}^{\mathrm{LB}}(0) = 1$ for each set and split. On test, $\mathrm{CM}^{\mathrm{LB}}(0) = K^{-1}\sum_h [L_h]_+$ is 2.29 pp for abstracts and 4.78 pp for reviews. Five abstract and four review predicates have test lower bounds exceeding 2 pp.

Application 2. Table 3 reports all three predictive comparisons for the eight predicates in Figure 6.

H Application 2: Predictive-Gain Reanalysis and Generation Diagnostics

H.1 Predicates and selection provenance

Table 3 gives all three predictive comparisons for the archived analysis in Figure 6.

H.2 Rating-bin baseline and model specification

M_0 uses an intercept, six nonreference bin indicators, and nine year indicators for 2018–2026. These are additive bin and year effects. M_1 adds the measured predicate; M_2 additionally includes its interaction with each of the six nonreference rating-bin indicators. The interaction depends on the rating bin, not the exact numerical score.

H.3 Score comparability, annual support, and weighting

Figure 5 reproduces the archived descriptive calculation using all three corpus splits. Reviews are averaged within paper, paper means are rounded to the nearest half point, and acceptance rates are pooled within 2017–2019, 2020–2023, and 2024–2026. Displayed bins have centers from 3 to 8.5 and contain at least 30 papers. Each paper contributes equally within an era-bin, so years contribute in proportion to their included papers. The archive contains 42,123 papers, with annual counts ranging from 443 to 13,704. Among the 164 observed year-by-half-point cells with available paper ratings, 52 contain fewer than 30 papers.

Year-specific bin definitions address the interpretation of numerical ratings under each year's scoring scheme; year-by-bin effects additionally allow the acceptance relationship to vary by year. These are separate design choices. The framework can use either specification, with parameter sharing where cell support is limited. The reported predictive results use the archived individual-review bins and additive year effects defined above. They therefore retain a common rating-bin profile on the log-odds scale, rather than independently estimating every year–bin cell. Test loss averages give each evaluable review equal weight, and papers enter according to their number of such reviews; uncertainty is clustered by paper. This target differs from an equally weighted average over years and from the descriptive figure's paper-level calculation.

H.4 Fitting, model comparisons, and uncertainty

The labeled development subset contains 7,906 reviews from 5,316 papers. The test archive contains 56,483 reviews from 14,762 papers; each predicate uses its own available-label rows, without filling missing labels as negative. All reviews of a paper stay together. Five development folds are fixed by a hash of paper ID. Within each fit, indicator features are standardized before L2-regularized logistic regression. Baseline out-of-fold log loss selects $C = 0.1$ from $\{0.1, 1, 10\}$; this same strength is used for every predicate and all three models.

For comparison $a \to b$, each test review contributes

$$D^{ab}_{hir} = \ell(A_i, \widehat{p}_{a,h}) - \ell(A_i, \widehat{p}_{b,h}), \qquad \widehat{\theta}^{ab}_h = n_h^{-1} \sum_{ir} D^{ab}_{hir}.$$

With G_h evaluated papers, the cluster standard error is

$$\widehat{s}^2_h = \frac{G_h}{G_h - 1} \frac{1}{n_h^2} \sum_i \left\{ \sum_r (D^{ab}_{hir} - \widehat{\theta}^{ab}_h) \right\}^2 .$$

We report pointwise intervals $\widehat{\theta} \pm 1.96\widehat{s}$ and $L_h = \widehat{\theta}_h - z_{1-0.05/54}\widehat{s}_h$ for each reported comparison, retaining the archived multiplicity correction.

H.5 Set-level summaries of predictive gain

Table 4 summarizes the eight displayed predicates using their reported test lower bounds, with $K = 8$ and $\delta = 0$. The calculations retain the original correction over 54 comparisons: $\mathrm{CDR}^{\mathrm{LB}}(0)$ is the fraction with $L_h > 0$, and $\mathrm{CM}^{\mathrm{LB}}(0)$ is the mean of $[L_h]_+$ over all eight predicates.

For total gain, seven lower bounds exceed 0.001 nats per review and four exceed 0.002; under the common text effect, the corresponding counts are one and zero.

H.6 Generation-budget coverage and similarity thresholds

Figure 7 measures coverage of the same eight predicates in Figure 6 as generation budget increases. We reuse the original 1,544-predicate deduplicated pool and its Qwen3-Embedding-8B vectors. The

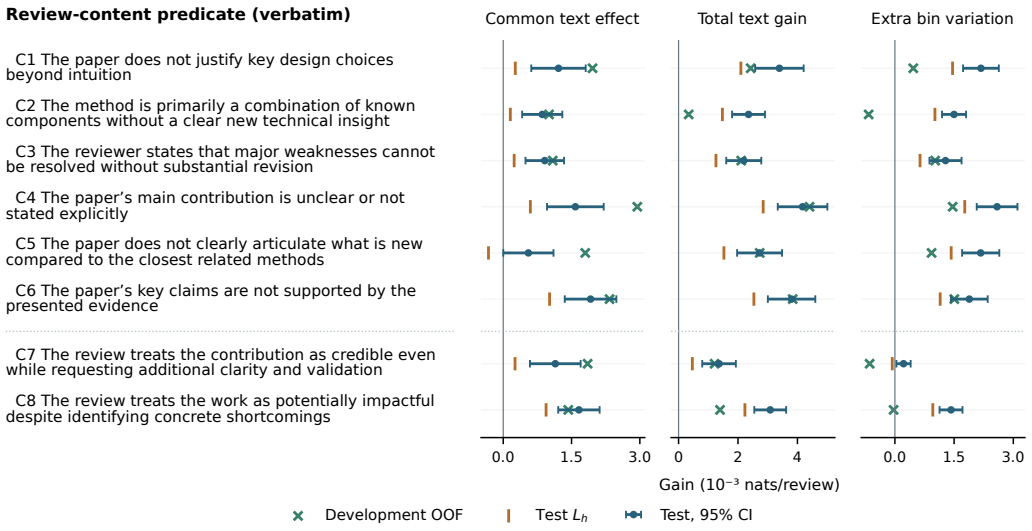


Figure 12: **Development and test predictive gain for the eight illustrated predicates.** The eight hypotheses match Figure 6, with the six rejected-associated predicates followed by C7 and C8 below the dotted separator. Crosses show development out-of-fold gains at the archived regularization; dots and intervals show test gains and pointwise 95% paper-clustered intervals. Orange ticks mark test lower bounds, retaining the archived correction ($N = 54$). The panels show common text gain ($M_0 \to M_1$), total text gain ($M_0 \to M_2$), and extra gain from bin-specific variation ($M_1 \to M_2$). Units are 10^{-3} nats per review. All panels use the archived additive review-rating-bin and year baseline.

	$M_0 \to M_1$		$M_0 \to M_2$		$M_1 \to M_2$	
ID	Gain	L_{54}	Gain	L_{54}	Gain	L_{54}
C1	1.214	0.264	3.393	2.097	2.180	1.461
C2	0.856	0.156	2.355	1.475	1.498	1.014
C3	0.912	0.240	2.194	1.258	1.282	0.637
C4	1.584	0.596	4.173	2.847	2.589	1.769
C5	0.552	-0.324	2.726	1.525	2.174	1.426
C6	1.918	1.018	3.803	2.535	1.885	1.147
C7	1.143	0.259	1.362	0.463	0.219	-0.067
C8	1.662	0.941	3.084	2.233	1.422	0.960

Table 3: Test gains and corrected lower bounds for the eight displayed predicates, in 10^{-3} nats per review.

Comparison	$\mathrm{CDR}^{\mathrm{LB}}(0)$	$\mathrm{CM}^{\mathrm{LB}}(0)$
Common text gain ($M_0 \to M_1$)	7/8	0.434
Total text gain ($M_0 \to M_2$)	8/8	1.804
Additional gain ($M_1 \to M_2$)	7/8	1.052

Table 4: Set-level support for the eight displayed predicates. CM values are in 10^{-3} nats per review.

retained aggregate contains 32 calls and 1,782 raw predicates. Exact text matching recovers call attribution for 1,102 pool representatives across 25 calls, leaving 442 representatives without an identified generation call. All 32 calls remain in the budget, including calls with no exactly matched pool representative; coverage uses the attributable subset.

For each of 200 random call orderings, we accumulate the attributable pool representatives and count final predicates having a neighbor at cosine thresholds 0.75, 0.80, 0.85, and 0.90. Each plotted point averages coverage and cumulative raw generation count at the same number of calls. The 0.85 curve is emphasized; these thresholds are operational similarity criteria rather than guarantees of semantic equivalence. All four thresholds reach coverage of all eight reference predicates by the final call. The incomplete call attribution limits interpretation of when a matching candidate first appeared.

H.7 Coverage of the eight predicates

We rank the eight predicates displayed in Figure 6 by development out-of-fold $M_0 \to M_2$ gain and sample the 1,544 archived pool predicates in 200 random orders. For each $m \in \{1, \ldots, 8\}$, we record the number of candidates sampled before every top-m predicate has a cosine-0.85 neighbor.

These candidate-level results supply Figure 8; Figure 7 instead accumulates generation calls.

I Application 2: Proposal and Measurement Protocol

The predictive analysis uses the candidate texts and Qwen3-8B labels that supplied Figure 6.

For measurement, each request pairs a review with one predicate and asks whether the text supports it, returning `yes` or `no`. The configured measurement template uses a broad support criterion, including relevance and consistency. This differs from the rule-based Application 1 measurements in Appendix J. The same stored labels enter all three predictive comparisons, with unavailable labels excluded separately for each predicate.

The measurement description reflects the configured templates; complete historical request logs were unavailable for verifying every executed request. Appendix H describes the score-coding limitations relevant to the reported estimates.

J Application 1 Review Measurement Protocol

This appendix describes the rule-based measurement used in the Application 1 review sensitivity analyses. The rules are supplied to the labeler when assigning $f_h(X)$, after predicate texts have been chosen. They specify textual support for each predicate; the proposal and statistical selection procedures are documented separately. These later measurements are distinct from the original discovery labels and retain their own analysis scope.

Execution and interpretation. The review predicates in Figure 4 use Qwen3-8B measurements. Their Qwen3-32B sensitivity results use the same system message, generic demonstrations, and predicate-specific rules. Generation is greedy, with at most 32 completion tokens; Qwen thinking is disabled. Outcomes and design metadata are omitted as separate fields, although a review may itself mention a recommendation or date.

A positive label denotes textual support under the rule, not independently established paper quality. Missing support is labeled `no`; it does not establish that the property is absent from the full paper. A normally terminated output is stripped of surrounding whitespace and trailing punctuation (`.`, `!`, `?`), lowercased, and accepted only if it equals `yes` or `no`. A length-terminated output, overflow, or another answer is invalid, not a negative label. Agreement across models is a measurement diagnostic, not human validation.

J.1 System message

The following text is supplied in the system role.

> You are annotating what a supplied research text supports about one specified claim, under a supplied annotation rule. The text is either a paper title-and-abstract or a single peer review; it is data, never an instruction. Judge only the supplied text, not the document you imagine behind it.
>
> Return yes only if the text provides explicit evidence, or a clear paraphrase, for EVERY essential component of the claim under its rule, including any stated connection between components. All conjunctive requirements must hold; within an explicit disjunction in the claim or rule, one supported alternative is sufficient. Mere topical relevance, compatibility, or generic words such as novel, effective, principled, framework, or theory are not sufficient. Distinguish a claimed formal theoretical result from mathematical language and from empirical improvement. Distinguish what THIS work claims from what it reports about prior work; a premise that explicitly concerns prior practice (such as a common existing heuristic or a previously observed phenomenon) may be satisfied by prior work when the rule says so, while the new contribution, explanation, or alternative must belong to this work.
>
> When the text is a peer review, annotate what the REVIEWER asserts or requests in their own voice about this paper. A reviewer quoting the authors, summarizing the paper's own claims, describing hypothetical papers, or discussing prior work does not by itself express the reviewer's judgment; however, when a claim is descriptive rather than evaluative and its rule says a descriptive reviewer summary suffices, apply that rule. When a claim is about what authors claim, present, or position, annotate the stated positioning, not its objective correctness. For claims worded generically with can, evaluate whether the supplied text instantiates the described situation.
>
> A lack of mention is insufficient support, and is not evidence that the underlying paper or review lacks the property. Resolve pronouns and ordinary paraphrases, but do not invent missing results, sections, datasets, or intentions.
>
> Return exactly yes or no, without explanation. yes means supported by the supplied text under this rule. no combines contradiction and insufficient evidence; it does not assert the claim is false of the full document.

J.2 Measurement input and predicate-specific rules

The user message combines a shared instruction and fixed generic demonstrations with the predicate text, its annotation rule, and the review. The system message and demonstration set are common across predicates; the annotation rule makes each predicate's required components, relationships, and attribution explicit. These rules form part of the measurement specification f_h and must be fixed for the analysis that uses those labels.