THE PREDICTABLE CORE OF PEER REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Peer reviewers often assign different scores to the same paper. Across 25,579 ICLR papers from 2020–2025, the Pearson correlation between one reviewer's score and the mean of the remaining reviewers ranges from 0.435 to 0.505. We show that collective ratings contain patterns that a small text model can learn. Our model combines word and character TF-IDF features with ridge regression over identity-filtered scientific text. On all 1,028 public NAIDv2 test papers, it reaches Spearman correlation 0.522, surpassing the dedicated NAIPv2-8B scorer's 0.439 and reducing calibrated mean-squared error by 7.03%. The comparison uses the same papers, outcomes, and calibration protocol, with broad manuscript text for the sparse model and NAIPv2's native title-and-abstract interface. Annual forward evaluations, presentation controls, and fixed-budget reading experiments trace the advantage to information beyond the opening text. Reading more of the manuscript reduces NAIPv2's added benefit for mean-score prediction by 34.8%. Human judgments supply further information after the manuscript model. Sparse learning thus reveals a predictable core in collective ratings while individual reviews contribute additional signal.

1 INTRODUCTION

Reviewers can read the same paper and assign substantially different scores. Across 25,579 ICLR papers from six conference years, a reviewer's score correlates with the mean of the other reviewers at 0.435–0.505. This limited agreement raises a basic question for automated scientific assessment: *how much of a paper's collective rating can be learned from the manuscript itself?*

A learner trained across many papers can recover associations that recur across individual review panels. Individual judgments vary around a collective response to the manuscript; repeated patterns in the text can still make that response predictable. We call this manuscript-predictable component the *predictable core* of peer review.

We learn this component with a compact manuscript scorer. The model represents identity-filtered scientific text with word and character TF-IDF features and fits a ridge regressor. Its input spans the paper's claims, methods, evidence, and positioning. The scoring function is a linear combination of at most 90,000 sparse features, with vocabulary and coefficients learned from training papers. Inference requires manuscript text alone.

This simple model outperforms the dedicated NAIPv2-8B scorer (Zhao et al., 2026). Across all 1,028 papers in the public NAIDv2 test release, Spearman correlation with reviewer means rises from 0.439 to 0.522, and calibrated mean-squared error falls by 7.03%. We compare the systems on the same papers and outcomes with a common calibration procedure. The sparse model reads up to 64,000 characters of scientific text; NAIPv2 uses its released 512-token title-and-abstract interface.

We then ask where that strength comes from. Across annual forward evaluations, broad-text scores retain predictive information after accounting for opening-text scores. On the public benchmark, the increment survives controls for 17 presentation statistics and the NAIPv2 score. Under a fixed reading budget, allocating text to claim, method, and evidence sections improves over uniformly distributed windows. As the sparse model reads more of the manuscript, adding NAIPv2's score produces 34.8% less additional MSE reduction. Broader text gives the small model access to some of the useful signal supplied by the 8B scorer.

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.

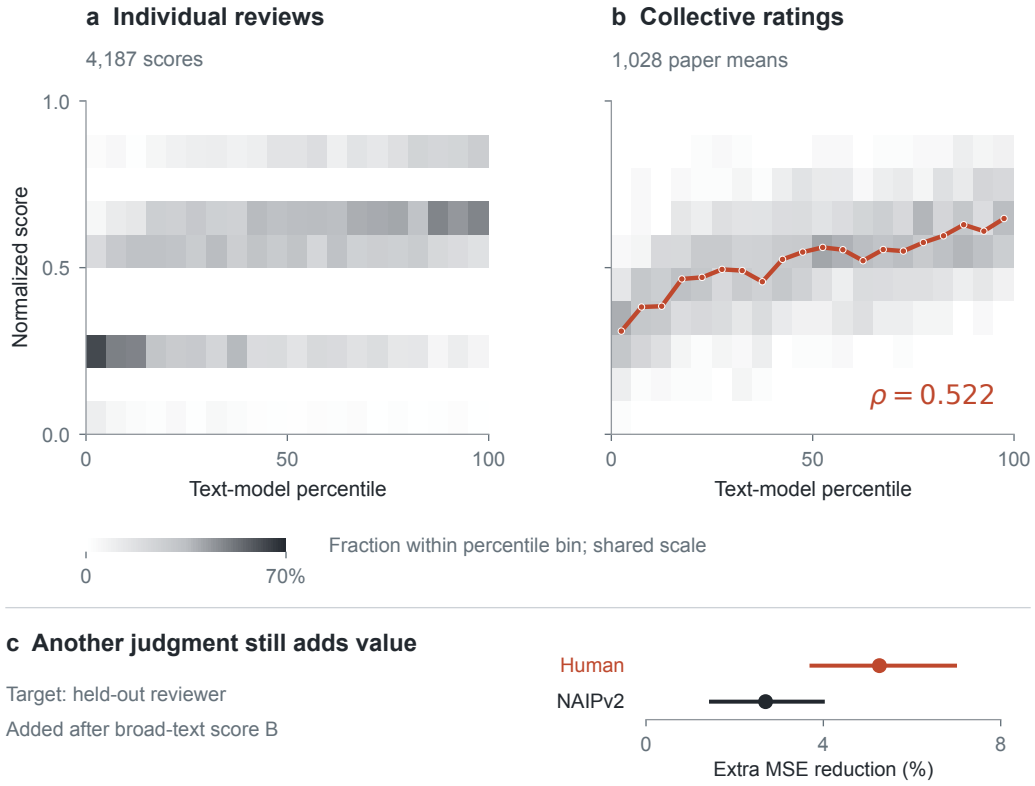

Figure 1: **Individual disagreement and collective predictability.** Papers have the same text-model ordering in both maps. (a) Distribution of 4,187 individual scores, with equal total weight per paper. (b) Distribution of 1,028 paper means; the line joins their means in 20 prediction bins. Color gives within-bin fractions on a shared scale. Correlations use individual papers: Spearman 0.522 for broad text and 0.439 for NAIPv2. (c) Additional MSE reductions from another reviewer or NAIPv2 when predicting a held-out review. Whiskers are paired 95% paper-bootstrap intervals divided by the text-model baseline MSE; additions are evaluated separately.

The manuscript model also provides a reference for studying individual judgment. Once its prediction is available, another reviewer's score still improves prediction of a distinct held-out reviewer. The learned collective response and the information supplied by a new judgment can therefore be measured separately.

Sparse prediction of review outcomes has a foundation in PeerRead (Kang et al., 2018). Our contribution is a broad-text scoring model that surpasses a dedicated 8B scorer on a common public benchmark, together with an empirical account of why it succeeds.

2 Reviewer Disagreement and the Prediction Target

2.1 Measuring Agreement with the Rest of the Panel

For paper i, let $r_{i1}, \ldots, r_{im_i}$ be its review scores and a_i their arithmetic mean. We measure agreement between a focal reviewer and the remaining panel using

$$r_{ij} \qquad \text{and} \qquad a_{i,-j} = \frac{1}{m_i - 1} \sum_{k \neq j} r_{ik}. \tag{1}$$

Excluding the focal score removes self-inclusion from the comparison. Each paper receives equal total weight, divided equally among its reviewers. The analysis covers papers with an explicit accept/reject outcome and at least three usable official reviews in the OpenReview snapshot. Scores

Table 1: **Individual-to-panel agreement across six ICLR years.** Pearson correlation compares one official score with the mean of the other reviews, giving every paper equal total weight.

Year	Papers	Reviews per paper	Reviewer–panel Pearson
2020	2,130	3.08	0.450
2021	2,616	3.86	0.452
2022	2,677	3.90	0.505
2023	3,859	3.78	0.480
2024	5,684	3.87	0.435
2025	8,613	4.05	0.442

are the public values at collection, so the quantity measures cross-reviewer agreement at that review stage.

Agreement remains moderate across all six years (Table 1). An individual score follows the other reviewers' assessment imperfectly. Learning across manuscripts pools recurring patterns across these variable panels, motivating the collective-rating prediction task below. The human-only cohorts and the public model benchmark serve these separate purposes.

2.2 Learning the collective response

Let X denote the permitted manuscript information. Draw a paper with equal probability, then one of its reviewers uniformly, and call the resulting score R. The manuscript-predictable response is $g(X) = \mathbb{E}[R \mid X]$. Squared-error risk separates as

$$\mathbb{E}[(f(X) - R)^2] = \mathbb{E}[(f(X) - g(X))^2] + \mathbb{E}[\mathrm{Var}(R \mid X)]. \tag{2}$$

This forecasting identity (Gneiting, 2011) accommodates both a learnable mean response and variation within the review process. Conditional variation includes differences in expertise and criteria, shared influences, and information outside the text representation. We evaluate the learned scorer against paper means, then measure the information supplied by another judgment through a distinct-reviewer prediction task.

3 A sparse manuscript model outperforms NAIPv2-8B

3.1 Data and comparison design

Our comparison combines **complete public-test coverage, full-manuscript identity filtering, and a common paper-level evaluation protocol**. We retain all 1,028 NAIDv2 test records. The official 23,246-record training partition supplies 20,921 fitting records and a separate 2,325-record reserve; public test IDs are excluded from base fitting. We build the text inputs from complete Markdown manuscripts, apply identity filtering before section routing and truncation, and retain source and version records. The resulting broad view covers up to 64,000 characters across claim, method, evidence, and related-work text.

Training uses the released Review Tendency Signal (RTS). Evaluation reconstructs the arithmetic reviewer mean from each test record's individual normalized ratings, giving every system the same target. Reviewer comments, meta-reviews, and decisions are excluded from the manuscript feature map. Appendix A documents the processing rules and provenance of the publicly available versions. Annual forward evaluation and distinct-reviewer prediction provide complementary tests in Sections 4.1 and 5.

3.2 Representation, comparator, and evaluation

Sparse scorer. The feature map $\phi(x)$ concatenates word unigram/bigram TF-IDF features, capped at 40,000, and character-boundary 3–5-gram features, capped at 50,000. Vocabulary and inverse document frequencies are fitted on training papers. Ridge regression estimates

$$(\hat{b}, \hat{\beta}) = \arg\min_{b,\beta} \sum_{i \in \mathcal{T}} \left(t_i - b - \phi(x_i)^\top \beta\right)^2 + 10\|\beta\|_2^2. \tag{3}$$

Table 2: **The broad-text sparse model surpasses NAIPv2-8B.** All rows use the same 1,028 test papers and normalized reviewer means. Spearman uses raw predictions; MSE uses common cross-fitted calibration.

Scorer	Manuscript input	Spearman	MSE
TF-IDF/Ridge	Title and abstract	0.3427	0.018603
NAIPv2-8B	Title/abstract, $\leq$512 tokens	0.4388	0.017103
TF-IDF/Ridge	Broad text, $\leq$64k characters	**0.5215**	**0.015900**

The title-and-abstract and broad-text variants share this feature family and regularization strength. Feature thresholds and solver settings appear in Appendix B.

Released 8B comparator. We run the released NAIPv2-8B checkpoint with its native title-and-abstract prompt and 512-token truncation. Table 2 compares the resulting predictions against our manuscript scorers on identical papers and reviewer means, with each system's reading scope listed alongside its results.

Common metrics and calibration. Spearman correlation measures the ranking of raw model scores against reviewer means. MSE additionally assesses the numerical predictions. Each scorer receives the same five-fold, paper-level affine calibration, including acquisition source as a common control. Every calibrated prediction comes from a head fitted on the other four folds. The base scorer's inference input remains manuscript text; source indicators enter only these common evaluation heads. Paired paper-bootstrap intervals retain within-paper dependence. The public and annual analyses are retrospective; Appendix A records their split roles and uncertainty scope.

3.3 Main result: higher correlation and lower prediction error

The broad-text model leads on both metrics. Its Spearman correlation exceeds NAIPv2-8B by 0.0827, and its MSE is **7.03% lower**. The absolute MSE reduction is 0.001203, with a paired 95% interval of [0.000191, 0.002210] from 2,000 paper-bootstrap resamples of the out-of-fold predictions.

Manuscript access changes the comparison decisively within the sparse model family: title-and-abstract prediction reaches 0.343, whereas broad-text prediction reaches 0.522. The same feature family and learning rule cross the 8B system's performance level as their reading scope expands. We next examine which aspects of manuscript access account for this gain.

4 Where the predictive advantage comes from

4.1 Broad text contributes across annual forward evaluations

The first test asks whether broad text adds information beyond the opening under temporal transfer. For 2023–2025, vocabulary and model fitting use only preceding years, with training pools of 6,518, 8,518, and 10,497 papers. The short view contains the first 2,000 characters; the broad view combines retained claim and body text, averaging approximately 32k–37k characters. Targets are the corresponding stored current-rating means.

For each annual cohort, we compute the partial Pearson correlation between the broad-text score and reviewer mean after accounting for the short-text score. The reverse calculation tests what short text adds after broad text. These within-year associations are invariant to score shifts and positive rescaling, allowing the analysis to follow predictive information across changing annual score levels.

Broad text retains a positive conditional association in every panel, and the association is consistently larger than the reverse contribution (Table 3). The 2026 diagnostic extends the directional test to 2,218 papers with available review-window-version text, using title/abstract and bounded 40k-character predictions. Across these settings, useful scoring patterns repeatedly extend beyond the information captured by the short-text model.

Table 3: **Broad text retains additional signal in every annual panel.** Partial Pearson correlations condition on the other text score. Intervals are 95% paper-bootstrap intervals. The 2026 subset uses the separate text views described below.

Test year	Papers	Broad given short	95% interval	Short given broad
2023	1,998	0.373	[0.333, 0.415]	0.016
2024	1,977	0.312	[0.272, 0.352]	0.055
2025	1,981	0.292	[0.250, 0.336]	0.080
2026 subset	2,218	0.170	[0.129, 0.211]	0.060

Table 4: **Text selection at a fixed 16k-character budget.** Training records, features, regularization, and evaluation match Table 2.

Selector	Spearman	Mean-score MSE
Stored-text prefix	0.4521	0.017523
Uniformly spaced windows	0.4256	0.017811
Section-priority rule	**0.4691**	**0.017061**

4.2 The increment survives measured presentation controls

Reading more text exposes both scientific content and presentation cues. We therefore test broad text against a reference containing NAIPv2, the common source control, and 17 surface statistics. These include length, figure and table indices, equation and citation references, recent-citation share, numerical density, code-release language, and writing and claim-frequency measures. Evaluation again uses paper-level cross-fitting.

Adding the broad-text score reduces mean-score MSE by **5.14%**, an absolute reduction of 0.000820 [0.000408, 0.001257]. Its advantage survives the fitted combination of the 8B score and measured presentation statistics. The result narrows the explanation from broad document appearance to additional patterns in the retained scientific text.

Exploratory channel probes further localize the signal. Content-word and numeric/reference-marker predictors reduce error by 2.26% and 2.63% over the same reference, with nominal intervals above zero. Function-word style and citation-year channels have smaller increments whose intervals include zero. These probes indicate useful content and reporting signals; Appendix B specifies the controls and inference scope.

4.3 At a fixed budget, where the model reads matters

We hold training papers, feature family, ridge penalty, and character budget fixed. Three selectors retain at most 16,000 characters: the prefix of the stored sequence; sixteen approximately evenly spaced windows; or a section-priority allocation of up to 2k characters to claims, 6k to methods, and 8k to evidence-related sections. Unused text fills any short allocation. All selectors average 15,982.4 training characters and use 16,000 test characters per paper.

Section priority lowers MSE by **4.21%** relative to uniform sampling, with an absolute-gain interval of [0.000273, 0.001254]. It yields the highest observed correlation and lowest MSE in Table 4. Its 2.64% MSE improvement over the prefix has an interval spanning zero. Uniform windows have lower observed correlation and higher MSE than the prefix. Appendix B provides both remaining intervals.

These comparisons distinguish broader manuscript access from wider spacing of a fixed amount of text. Distributing windows across the document alone yields no established improvement; allocating that budget to recognized scientific sections improves on uniform sampling. Across Tables 2 and 4, the broad 64k view remains the strongest representation.

Table 5: **Additional information for a held-out reviewer.** Positive reductions indicate better predictions. Intervals use 1,000 paired paper-bootstrap resamples of fixed out-of-fold losses. Relative gains use each row's reference MSE.

Available → added signal	ΔMSE	95% interval	Reduction
Broad text → NAIPv2	0.001013	[0.000551, 0.001499]	2.70%
NAIPv2 → broad text	0.002220	[0.001538, 0.002887]	5.72%
Broad text → one reviewer	0.001979	[0.001403, 0.002620]	5.26%
Broad text + reviewer → NAIPv2	0.000616	[0.000325, 0.000921]	1.73%
Broad text + reviewer → reviewer	0.001255	[0.000972, 0.001574]	3.52%

4.4 Broader reading reduces the added benefit of the 8B score

We add NAIPv2's score to each sparse predictor and measure the further reduction in mean-score error. The sparse predictors differ in how much manuscript text they read; papers, targets, source controls, and cross-fitting remain fixed. NAIPv2 reduces mean-score MSE by 0.001558 after title-and-abstract prediction and by 0.001015 after broad-text prediction. Its added gain falls by **34.8%** [1.5%, 58.8%], an absolute decrease of 0.000542 [0.000015, 0.001086]. A separate comparison using a 2k-character prefix as the short reference yields a 46.2% decrease [28.2%, 63.3%].

NAIPv2 contributes less additional predictive signal once the sparse model reads beyond the title and abstract. Broader manuscript access lets the small model recover some of the signal that the 8B scorer supplies after short-text prediction.

5 Information supplied by another reviewer

5.1 A distinct-reviewer prediction task

To measure what a human judgment adds after manuscript-based prediction, we use one reviewer's score to predict a different reviewer on the same paper. The target changes from the paper mean to an individual score, keeping the supplied judgment separate from the quantity being predicted. Let B be the broad-text score, L the NAIPv2 score, and $H = r_{ik}$ the supplied reviewer's score. The target is $Y = r_{ij}$ with $k \neq j$. When two human scores enter the predictor, both input positions and the target position are distinct.

For available signals Z, define the added predictive value of U as

$$\Delta_Y(U \mid Z) = \widehat{\mathcal{R}}_Y(Z) - \widehat{\mathcal{R}}_Y(Z, U), \tag{4}$$

where $\widehat{\mathcal{R}}_Y$ is paper-weighted MSE from an out-of-fold linear prediction head. Five paper-level folds keep all combinations from one paper together, and every paper receives equal total weight. Enumerating eligible ordered reviewer triples, capped at 128 per paper, gives 30,396 weighted comparisons over the 1,028 papers. All heads include the common source controls.

5.2 Human judgments remain informative after manuscript prediction

Among single-signal predictors, broad text has the lowest observed MSE: 0.03759, compared with 0.03880 for NAIPv2, 0.03843 for one other reviewer, and 0.04302 for the source-controlled prior. Adding another judgment improves prediction further (Table 5).

One human judgment reduces the broad-text reference error by **5.26%**; NAIPv2 reduces it by 2.70%. Broad text is also informative after the 8B scorer, reducing error by 5.72%. After broad text and one human judgment are available, a second reviewer supplies a further 3.52% reduction. Each increment is conditional on the existing signals; the percentages use different reference errors and describe overlapping information.

For comparison, the mean of two other reviewers has MSE 0.03608; broad text combined with one reviewer has MSE 0.03561, with an interval for their difference spanning zero. The consistent positive conditional gains establish the central result: a manuscript learner captures collective scoring patterns, and individual judgments supply further predictive information.

6 Related work

Learning paper scores. PeerRead established manuscript, review, and decision datasets and sparse outcome-prediction baselines (Kang et al., 2018). NAIPv2 uses domain/year-conditioned pairwise learning to train an efficient pointwise scorer (Zhao et al., 2026). Comparison-native evaluation studies collaborative ranking from comparative judgments (Zheng et al., 2026). We extend the sparse-learning comparison to broad manuscript inputs and connect its predictive strength to reviewer disagreement, information scope, and conditional judgment value.

Generated reviews and numerical predictions. DeepReview studies review generation through a human-like deep-thinking process (Zhu et al., 2025); Beyond Rating evaluates the properties of generated critiques (Li et al., 2026). Our target is numerical agreement with recorded reviewer assessments. It supports a direct study of the shared scoring patterns learned from text and the information added by a separate judgment.

Manuscript access and human information. Thelwall & Ali (2026) compare ChatGPT ratings from titles/abstracts and PDFs on small journal-assessment samples and report mixed effects of richer input. Our annual and fixed-budget experiments study information scope within a common sparse model family. Tan et al. (2026) formulate the allocation of costly human information in AI-assisted decisions, including peer review. Our conditional-risk analysis measures the value of a distinct review score after an informative manuscript reference, and varies that reference's reading scope.

7 Discussion

A learnable collective response. Moderate individual-to-panel agreement coexists with strong manuscript-based prediction of collective scores. A compact sparse model learns enough of this response to surpass a dedicated 8B scorer. Its advantage extends beyond opening-text predictions and measured presentation statistics, and broader reading captures some information that the 8B model supplies after a short-text reference. This makes manuscript access a central experimental variable in automated scoring.

Complementary measurements. Predicting collective ratings and contributing a new expert judgment measure different capabilities. Human scores remain informative after the manuscript prediction. Application-specific objectives also matter: at a top-20% retrieval budget, broad text recovers 39.5% of the highest-scored quintile and NAIPv2 recovers 43.3%. Global rank correlation, numerical error, and top-tail retrieval characterize different aspects of a scorer.

Scope. The target is recorded reviewer response, whose associations can reflect topic, reporting practices, community expectations, and scientific substance. Our tests locate predictive information; identifying its causal sources requires interventions. The public benchmark uses available manuscript versions and a same-year-inclusive official split; annual models use preceding-year training. These are retrospective analyses, with version, overlap, and uncertainty details in Appendix A. Extracted text provides partial access to visual and mathematical content. Cross-year deployment requires destination-regime calibration. Additional 2026 and matched-text ChatGPT results appear in Appendix C.

8 Conclusion

Human review scores combine limited cross-reviewer agreement with a learnable collective pattern. A small TF-IDF/Ridge model captures this pattern from broad scientific text and outperforms the dedicated NAIPv2-8B scorer on the public benchmark, reaching Spearman 0.522 and reducing calibrated MSE by 7.03%. Annual evaluations, presentation controls, and reading experiments connect the advantage to information across the manuscript. Human judgments provide additional signal after this prediction. These findings reveal a predictable core in peer review and establish broad-text sparse learning as a strong reference for automated paper scoring.

AI USE STATEMENT

We used generative AI tools to clean and reformat the dataset. We did not use them for any other research tasks requiring disclosure (e.g., ideation, experimental design, theory, method implementation, or result interpretation). Additionally, we used AI tools to help debug code and to polish the writing. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for this work.

ETHICS STATEMENT

This study analyzes publicly available manuscripts and review scores in aggregate. Manuscript inputs are filtered to remove author-identifying material. The learned outcome reflects historical reviewing behavior, including community and presentation effects. Scores should be interpreted as estimates of that behavior; decisions affecting authors require contextual expert judgment. The analysis reports no reviewer-level rankings or identification of anonymous reviewers.

REPRODUCIBILITY STATEMENT

Section 3.1 specifies manuscript processing and target construction. Sections 3–5 define the compared inputs, fitting and evaluation splits, and conditional-prediction tasks. Appendix A records data provenance and analysis roles; Appendix B provides representation parameters, surface controls, uncertainty procedures, and computing details.

REFERENCES

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. URL https://arxiv.org/abs/0912.0902.

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 *Proceedings of NAACL-HLT*, 2018. URL https://aclanthology.org/N18-1149/.

Bowen Li, Haochen Ma, Yuxin Wang, Jie Yang, Yining Zheng, Xinchi Chen, Xuanjing Huang, and Xipeng Qiu. Beyond rating: A comprehensive evaluation and benchmark for AI reviews. *arXiv preprint arXiv:2604.19502v2*, 2026. URL https://arxiv.org/abs/2604.19502v2.

Lezhi Tan, Naomi Sagan, Lihua Lei, and Jose Blanchet. When should humans step in? optimal human dispatching in AI-assisted decisions. *arXiv preprint arXiv:2603.13688*, 2026. URL https://arxiv.org/abs/2603.13688.

Mike Thelwall and Parveen Ali. Is ChatGPT as reliable as individual reviewers assessing the quality of published journal articles from PDFs or titles and abstracts? *arXiv preprint arXiv:2607.25965*, 2026. URL https://arxiv.org/abs/2607.25965.

Penghai Zhao, Jinyu Tian, Qinghua Xing, Xin Zhang, Zheng Li, Jianjun Qian, Ming-Ming Cheng, and Xiang Li. NAIPv2: Debiased pairwise learning for efficient paper quality estimation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/e953fd35e34a5214cbc8030cc6a8ba90-Abstract-Conference.html.

Pujun Zheng, Jiacheng Yao, Jinquan Zheng, Chenyang Gu, Guoxiu He, Jiawei Liu, Yong Huang, Tianrui Guo, and Wei Lu. From isolated scoring to collaborative ranking: A comparison-native framework for LLM-based paper evaluation. In *Findings of ACL*, 2026. URL https://aclanthology.org/2026.findings-acl.1195/.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of ACL*, 2025. URL https://aclanthology.org/2025.acl-long.1420/.

A DATA, TARGETS, AND ANALYSIS ROLES

Input processing. The public benchmark is built from complete Markdown representations drawn from historical manuscript archives and public full-text mirrors. Cleaning precedes section routing and every fixed-length view, so identity-bearing material is handled at its original document boundary. Front matter, author-role and affiliation footnotes, acknowledgments, funding passages, publication markers, contact strings, literal URLs, and bibliography sections are filtered. Boundary handling distinguishes identity-bearing passages from nearby scientific sentences. Examples include author-contribution declarations embedded in footnotes and extraction artifacts that split section titles across lines. The same cleaning rules apply across the benchmark; the cleaner accepts manuscript content and does not receive outcome, year, or acquisition-source labels. Scientific prose is assigned to claim, method, evidence, related-work, and other zones, and the broad view takes the first 64,000 characters of their stored sequence. The title-and-abstract view and longer text views use the resulting manuscript records. Reviewer comments, meta-reviews, and decision labels are excluded from the manuscript feature map. Per-record source and version fields remain separate from this text representation.

Source composition and versions. All 1,028 official test records are retained: 751 map to the historical manuscript archive and 277 to public-mirror manuscripts whose original-submission version remains unverified. The experiment evaluates these available manuscript versions, with identity filtering for specified non-scientific cues and common acquisition-source controls in evaluation. Original-submission provenance remains incomplete across the historical and mirror collection. The official record split also contains eight previously flagged candidate train/test title overlaps under normalized-title matching; those pairs remain unresolved at the document-family level.

Split roles. The public release contains 23,246 training records and 1,028 test records. Of the training records, 20,921 supply base-model fitting and 2,325 form a separate reserve. The test records are from 2025, and the official training partition also contains 2025 papers. Base fitting excludes the public test record IDs. Five-fold evaluation heads are cross-fitted within the analysis cohort, so calibrated MSE measures within-cohort prediction rather than an uncalibrated deployment score. The public partition and annual cohorts were examined during development; reported analyses are retrospective. The fixed-16k selectors were specified before fitting the two additional selectors on the previously examined split.

Target roles. Public-benchmark sparse models learn the released RTS. Tables 2 and 4 and the conditional-reading analysis evaluate arithmetic means reconstructed from individual normalized ratings. Table 5 predicts a single reviewer using distinct human input and target positions. Table 1 compares current official scores with leave-one-reviewer-out panel means. Its human-only cohorts and the model benchmark have separate targets, populations, and correlation definitions. Rating snapshots describe the public stage at collection; changes by the same reviewer require aligned score histories.

Annual cohorts. The 2023–2025 predictors fit vocabulary and regression coefficients using preceding years and stored current-rating means. Their short input is a 2k-character prefix, and their broad input combines claim and body text. The 2026 diagnostic uses the existing 2,218-paper review-window-version subset, title/abstract scores, and bounded 40k-character predictions. This row is a directional extension under a separate text and cohort definition.

Observed leave-one-out geometry. Let $v_i = m_i^{-1} \sum_j (r_{ij} - a_i)^2$. Under equal paper weight and uniform focal-reviewer sampling, write $R = r_{ij}$ and $T = a_{i,-j}$. Since $T - a_i = -(R - a_i)/(m_i - 1)$,

$$\mathrm{Cov}(R, T) = \mathrm{Var}(a) - \mathbb{E}[v/(m-1)], \tag{5}$$

$$\mathrm{Var}(R) = \mathrm{Var}(a) + \mathbb{E}[v], \tag{6}$$

$$\mathrm{Var}(T) = \mathrm{Var}(a) + \mathbb{E}[v/(m-1)^2]. \tag{7}$$

These finite-panel identities accommodate varying reviewer counts and yield the correlations in Table 1 directly from the observed ratings.

B IMPLEMENTATION AND STATISTICAL DETAILS

Sparse representation. Word features use unigram/bigram TF-IDF, at most 40,000 features, minimum document frequency three, and maximum document frequency 99.5%. Character-boundary 3–5-grams use at most 50,000 features and minimum document frequency four. Both use sublinear term frequency. Ridge uses $\alpha = 10$, the LSQR solver, and tolerance 10^{-4}. Public fits weight papers equally. Each text variant fits its own vocabulary on its fitting documents.

Surface controls. The 17 features are log length; maximum figure and table numbers; equation references per 1,000 characters; appendix mentions; citations per 1,000 characters; cited-year lag; recent-citation share; numeric and percentage density; code-release language; first-person frequency; mean sentence length; type–token ratio; and frequencies of theory terms, state-of-the-art claims, and ablation mentions. Head normalizations are estimated within the fitting folds. Conditional effects concern the fitted linear prediction heads. Channel probes are exploratory, with nominal intervals unadjusted for multiple comparisons.

Uncertainty. Conditional-human and annual partial-correlation analyses use 1,000 paper-level bootstrap resamples; the fixed-budget and main system comparisons use 2,000. Resampling operates on fixed predictions or per-paper losses and retains within-paper dependence. Intervals condition on the fitted models and analyzed comparisons; they exclude model-refitting and selection uncertainty. Relative MSE reductions divide an absolute reduction by the corresponding reference MSE.

Fixed-budget contrasts. Relative to the stored-text prefix, uniform windows have a -1.65% MSE improvement, with an absolute-difference interval of $[-0.000841, 0.000259]$. Section priority improves on the prefix by 2.64%, with an absolute-gain interval of $[-0.000056, 0.001012]$. Against uniform windows, section priority improves by 4.21%, with an absolute-gain interval of $[0.000273, 0.001254]$. The prefix follows the stored section order. Priority selection uses recognized headings and fills missing allocations from unused text.

Computing. Sparse fitting runs locally on an Apple M5 with 24 GB memory. The recorded 64k fitting-and-evaluation run took approximately 1,725 seconds end to end. The NAIPv2 checkpoint was evaluated locally in FP16 on MPS with its released weights. These timings describe the recorded implementations and include their respective pipeline stages; isolated inference latency requires a separate timing protocol.

C ADDITIONAL EVALUATION SETTINGS

2026 conditional gain. On the 2,218-paper 2026 subset, adding NAIPv2 after the available broad-text predictor reduces error by 0.14%, with an absolute-gain interval spanning zero. The text views and version coverage are specified in Appendix A.

Matched-text ChatGPT comparisons. Existing small panels use matched text for sparse-model and ChatGPT scoring. Their Spearman correlations are 0.408/0.386 on 96 papers from 2025 and 0.119/0.308 on 94 papers from 2026. These previously examined panels show setting-dependent model ordering. They are separate from the 1,028-paper released-NAIPv2 comparison in Section 3.