AI-LIKENESS IN THE PEER-REVIEW ECOLOGY: EVALUATION–RESPONSE ASYMMETRY

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-like language has been linked to favorable peer-review evaluations, but whether that association extends to the author–reviewer exchange remains unclear. We analyze 150,635 ICLR reviews from 2024–2026, linking reviews to author responses and reviewer follow-up. Using EditLens and CoCoDet as complementary continuous proxies, we benchmark their separation of matched human and directly generated reviews and compare downstream associations with earlier ICLR cohorts. In the primary models, higher review-side AI-likeness is associated with higher ratings but less subsequent positive reviewer updating under both measurements. Author participation shows no shared direction: EditLens is associated with less rebuttal submission in 2025–2026, whereas CoCoDet is associated with more only in 2026. Pooled analyses generally link higher AI-likeness scores to more tabular response organization, with the magnitude depending on the measurement and text side. We call this pattern *evaluation–response asymmetry*: favorable evaluation does not extend uniformly to participation, organization, and reconsideration. These findings show that ratings alone do not characterize how AI-like language relates to peer review as an exchange.

1 INTRODUCTION

Large language models (LLMs) are becoming part of scientific peer review, with corpus-level evidence of LLM-associated changes in review language (Liang et al., 2024). Their potential role extends from drafting evaluations to helping authors formulate responses. A favorable evaluation alone does not establish that a review invites a response or that discussion leads to reconsideration.

Observational work associates detector-defined AI assistance with higher review scores and acceptance rates (Russo Latona et al., 2024), while peer-review benchmarks examine the detection of generated and edited reviews (Yu et al., 2026; Chen et al., 2026). A randomized intervention at ICLR 2025 found that providing reviewers with LLM feedback prompted review revisions and longer author–reviewer discussions (Thakkar et al., 2026). Less is known about how AI-like language in these routine exchanges relates jointly to evaluation, author participation, response organization, and reviewer updating.

We study this question using 150,635 ICLR Official Reviews from 2024–2026, linked to available author responses and reviewer follow-up. We examine four outcomes: reviewer rating, rebuttal submission, visible response organization measured by table count, and subsequent overall-rating increase. EditLens and CoCoDet provide complementary continuous measurements of AI-likeness for review text and author-response extensions. We assess their measurement basis using human and directly generated peer reviews, together with 2022–2023 historical cohorts that help distinguish existing textual associations from those specific to the period of widespread LLM use. The scores serve as population-level textual proxies, and the analyses describe associations rather than causal effects of individual AI use.

Our findings reveal *evaluation–response asymmetry*. First, both main measurements associate higher review-side AI-likeness with higher ratings and lower measured rates of positive reviewer updating in the primary models. Second, author participation separates by measurement: EditLens is associated with less rebuttal submission in 2025–2026, whereas CoCoDet is associated with more only in 2026, despite their shared rating direction. Third, higher AI-likeness generally accompanies more tabular response organization in pooled analyses, with larger review-side point estimates in 2026.

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.

These findings motivate evaluating AI-mediated peer review through participation and updating as well as the favorability of its judgments.

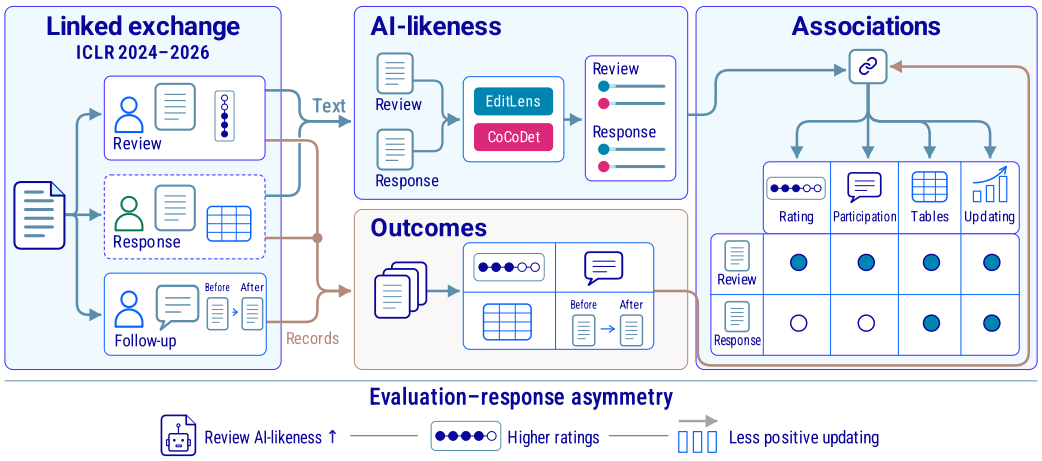


Figure 1: Study design and central association pattern. EditLens and CoCoDet score review and response text; linked ICLR 2024–2026 records supply four outcomes. The lower strip summarizes the shared primary-model association: higher review-side AI-likeness accompanies higher ratings and less positive updating.

2 Related Work

LLMs in review writing and evaluation. Corpus-level analysis estimates LLM-associated modification of conference reviews (Liang et al., 2024), while detector-based studies track the growth of AI-like academic review text (Shen & Wang, 2026). Russo Latona et al. (2024) use GPTZero-defined assistance, within-paper score comparisons, and matched acceptance analyses to report associations with higher scores and acceptance rates. Our study uses continuous textual measurements across ICLR 2024–2026 and links evaluation to author responses and reviewer follow-up.

Reviewer feedback and subsequent interaction. A randomized ICLR 2025 study found that LLM feedback intended to improve clarity and actionability prompted review revisions and longer author–reviewer discussions (Thakkar et al., 2026). Other studies relate score updates to pre-discussion ratings, co-reviewer signals, and author–reviewer interaction (Kargaran et al., 2025); review-trajectory analysis finds that initial-review structure already predicts much of the subsequent score movement (Louis et al., 2026). The intervention provides evidence about an explicitly introduced feedback tool; our observational analysis examines how AI-like language relates to these outcomes in routine exchanges.

Detecting AI-generated and AI-edited text. Zero-shot likelihood methods assess text through language-model probabilities: DetectGPT uses probability curvature under text perturbations, Fast-DetectGPT estimates conditional-probability curvature through sampling, and Binoculars contrasts perplexity with cross-perplexity from a pair of models (Mitchell et al., 2023; Bao et al., 2024; Hans et al., 2024). Supervised approaches instead learn targets relevant to generation and editing. CoCoDet, introduced with CoCoNUTS, distinguishes human, mixed, and AI-generated peer-review content (Chen et al., 2026); EditLens predicts the extent of AI editing using similarity-derived supervision (Thai et al., 2025). A further approach models human–AI collaboration through structured representations: DETree uses hierarchical representation learning to capture relationships among text sources (He et al., 2025). Peer-review benchmarks evaluate detection under direct generation and LLM-assisted editing (Yu et al., 2026; Chen et al., 2026), making these distinctions relevant to reviews that may combine human judgments with machine-assisted expression.

Connecting measurement to the review process. EditLens and CoCoDet target editing and content composition, respectively, with CoCoDet additionally trained on peer-review content. Comparing their associations with evaluation and interaction distinguishes shared patterns from those dependent on the textual proxy. Methods reports measurement validation and selection, including historical controls and auxiliary DETree evidence.

3 METHODS

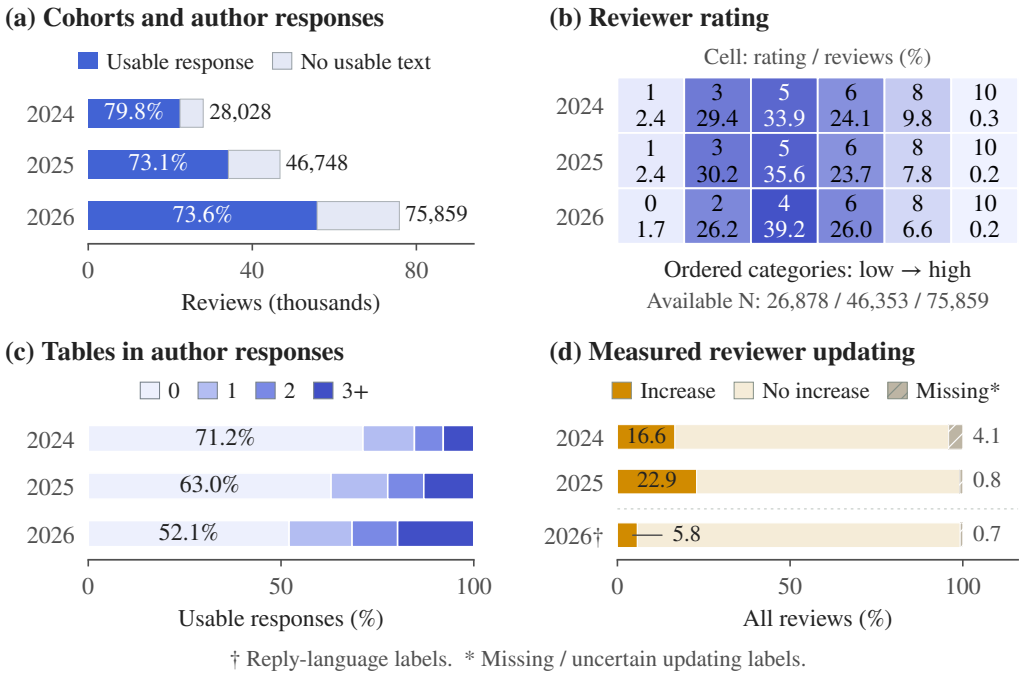


Figure 2: Cohort and outcome distributions. (a) Reviews with and without usable author responses. (b) Actual ratings and their percentages among available ratings, with a common 0–40% color scale and year-specific rating sources. (c) Table counts among usable responses; 3+ pools counts of three or more. (d) Updating-label composition among all reviews, including missing or uncertain labels. The 2026† outcome uses reply language; 2024–2025 use rating snapshots. The cohorts cover 38,256 year-specific paper records; these descriptive denominators differ from the regression samples.

3.1 INTERACTION DATA AND OUTCOMES

The dataset comprises 150,635 public ICLR Official Reviews from OpenReview for 2024–2026, covering 38,256 year-specific paper records and 112,342 nonempty usable author responses.[1] Each observation links a review to its paper, available author response, and reviewer follow-up, with inference clustered by paper; Figure 2 and Appendix Table 1 summarize the annual counts. These are full-cohort counts; outcome-specific analysis samples appear in Appendix J.

Reviewer rating uses pre-discussion snapshots for 2024–2025 and ratings recorded at collection for 2026, which lacks corresponding snapshots. *Rebuttal submitted* indicates nonempty author text linked to a review after table removal, and *rebuttal table count* counts contiguous Markdown table blocks as a measure of visible response organization. *Follow-up overall-rating increase* records increases between rating snapshots in 2024–2025 and explicit increase language in each reviewer's last available reply, classified by a local Qwen3.5-9B model, in 2026†; the dagger distinguishes the semantic label from observed score changes. We use *positive reviewer updating* to refer to this outcome; it measures upward rating revision or its explicit expression, rather than all forms of reconsideration.

[1] OpenReview API documentation: https://docs.openreview.net/reference/api-v2.

3.2 CANDIDATE AI-LIKENESS MEASUREMENTS

EditLens measures editing through the expected value of four ordered edit-magnitude buckets (Thai et al., 2025). CoCoDet distinguishes human, mixed, and AI-generated peer-review content (Chen et al., 2026); we use its AI-versus-rest log-odds $s = \ell_{\text{AI}} - \text{logsumexp}(\ell_{\text{Human}}, \ell_{\text{Mix}})$ to spread out probabilities concentrated near zero while retaining the pure-AI versus human-or-mixed contrast.

Higher scores indicate greater AI-likeness, not calibrated AI-use probability. Review inputs exclude metadata; response inputs exclude Markdown tables, retained separately as an outcome. Checkpoints, alternative candidates, and implementation checks appear in Appendices C and D.

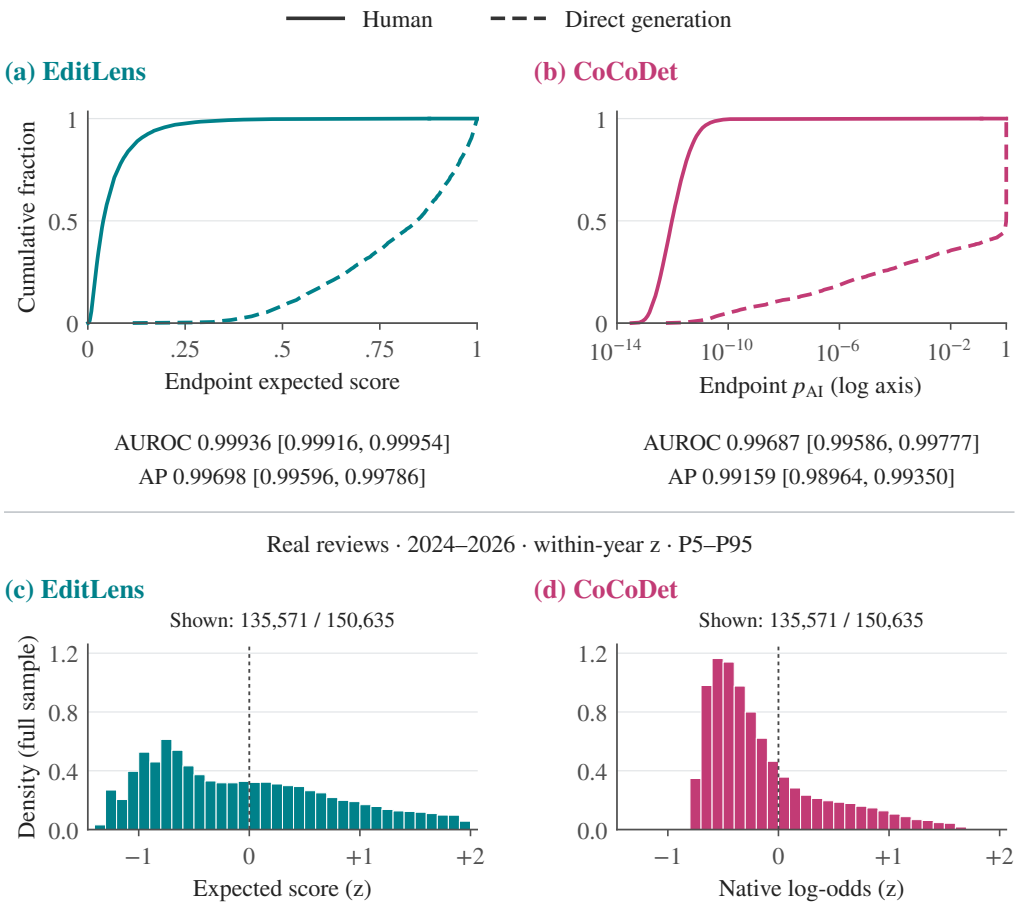


Figure 3: Endpoint discrimination and observed score distributions. (a,b) ECDFs compare 14,339 human and 3,126 directly generated reviews from 3,793 papers. Solid/dashed curves denote human/direct generation. CoCoDet's endpoint panel uses the original softmax probability on a base-10 log axis. AUROC and AP intervals use 1,000 paper-cluster bootstrap draws. (c,d) Histograms summarize all 150,635 real reviews in 2024–2026: EditLens expected scores and CoCoDet native AI-versus-rest log-odds, standardized within year and method using sample SD, then pooled. Each method displays its P5–P95 range; density uses full-sample N without renormalizing the retained subset. Display trimming does not change regression samples; bin construction is documented in Appendix C.

3.3 MEASUREMENT VALIDATION AND SELECTION

We compare human ICLR 2023 reviews with reviews directly generated from the corresponding papers. Both main measurements achieve AUROC above .996 and average precision above .991 (Figure 3), but this endpoint does not validate detection of naturally edited reviews.

Historical controls assess temporal specificity: 2022 is the stronger pre-LLM control, while 2023 is a low-exposure cohort overlapping the human endpoint. Fast-DetectGPT and Binoculars already show negative rating and participation associations in these cohorts, without stronger negative associations later; we therefore exclude them from the primary results. EditLens and CoCoDet have no clear historical rating or participation associations and positive post-minus-historical rating contrasts. Their support is outcome-specific: EditLens's table-count association predates the later cohorts, and CoCoDet's participation era contrast includes zero. Appendix C reports aligned specifications, scale sensitivities, and complete contrasts. Historical rating comparisons use recorded ratings, not the primary 2024–2025 pre-discussion outcome; historical updating lacks comparable snapshots.

We retain EditLens and CoCoDet as complementary main proxies and DETree as auxiliary evidence under an approximate implementation (Appendix D). Review-domain validation does not establish validity for author responses; rebuttal-side analyses remain domain extensions.

3.4 Downstream estimands and inference

Models are fitted by year, measurement, and text side, adjusting for the scored text's length. Scores and lengths are standardized within each final sample; intervals use paper-clustered covariance, with Benjamini–Hochberg correction within analysis families (Benjamini & Hochberg, 1995). Appendix F specifies token windows, exclusions, and multiplicity; Appendix I reports primary estimates. Binary-outcome effects are percentage-point probability differences from score $z = 0$ to $z = 1$, averaged over observed covariates.

Evaluation. OLS relates reviewer rating to review-side AI-likeness, reporting original rating-point change per score SD.

Author participation. Binomial logit models relate usable rebuttal submission to review-side AI-likeness.

Response organization. Among usable responses, negative-binomial models relate table count to AI-likeness on each text side, reporting percent change in expected count per score SD. Supplementary within-paper comparisons use a different outcome scale (Appendix H).

Reviewer updating. Among usable responses with available, non-uncertain labels, binomial logit models relate overall-rating increase to AI-likeness. Review-side models control categorical recorded rating; response-side models control categorical pre-discussion rating in 2024–2025 and recorded rating in 2026†. Supplementary initial-rating and ceiling checks cover 2024–2025 (Appendix G).

4 Results

4.1 Research questions

Using the measurements assessed in Section 3.3, we ask four questions about evaluation and the surrounding exchange:

- **RQ1 (Evaluation):** How does review-side AI-likeness relate to reviewer rating?
- **RQ2 (Author participation):** How does review-side AI-likeness relate to whether authors submit a response?
- **RQ3 (Response organization):** How does AI-likeness in reviews and author responses relate to response table count?
- **RQ4 (Reviewer updating):** How does AI-likeness in reviews and author responses relate to subsequent overall-rating increase?

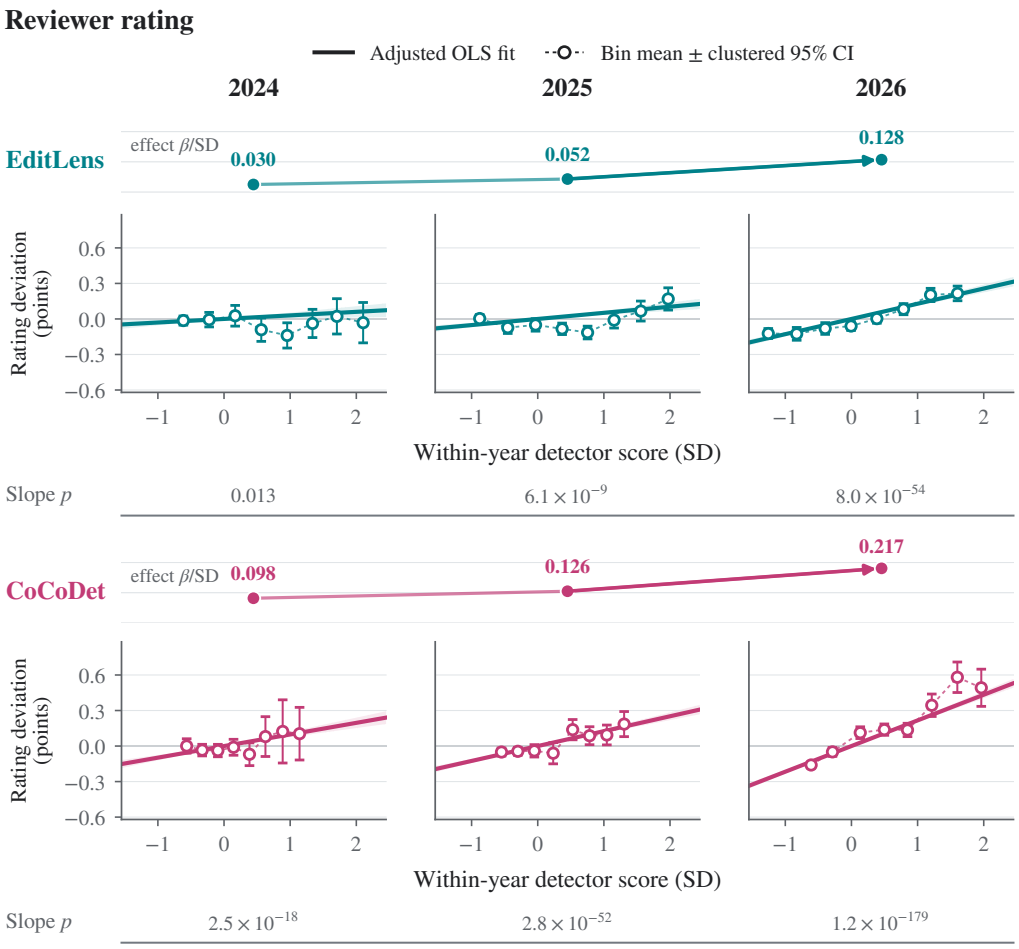


Figure 4: Reviewer-rating associations across years. Rows denote detectors; columns denote 2024–2026. Points and whiskers summarize eight equal-width bins within each method-year's P5–P95 score range: token-adjusted mean rating deviations and descriptive paper-clustered 95% intervals, conditional on the estimated length adjustment. Lines show OLS fits with slope-confidence ribbons. Fitting and standardization use the full model sample before display trimming. The upper trajectories compare annual rating-point slopes; connecting lines do not test temporal change. The *Slope p* rows report unadjusted two-sided paper-clustered Wald tests of the length-adjusted score coefficient, not goodness-of-fit tests. Ratings are pre-discussion in 2024–2025 and recorded in 2026.

4.2 Evaluation

Higher review-side scores are associated with higher ratings for both methods in all three cohorts (Figure 4). Per score SD, estimates are $+.030$, $+.052$, and $+.128$ rating points for EditLens and $+.098$, $+.126$, and $+.217$ for CoCoDet in 2024, 2025, and 2026, respectively, with point estimates increasing across successive cohorts for both methods. CoCoDet remains positive under raw-probability and rank-normal sensitivity specifications in all three cohorts. EditLens's rank-normal and clipped-logit sensitivities retain positive associations in 2025 and the extended 2026 diagnostics, whereas its 2024 support is limited to the raw score. The shared positive association describes the full-sample primary fits; early-cohort estimates, especially EditLens, are sensitive to score-range restrictions and nonlinear specifications (Appendix K).

4.3 AUTHOR PARTICIPATION

Higher review-side AI-likeness is associated with less author participation under EditLens in 2025–2026 and more under CoCoDet in 2026; both 2024 intervals and CoCoDet's 2025 interval cross zero (Figure 5). For the $z = 0$ to $z = 1$ score contrast, EditLens is associated with 0.54–0.78 percentage points less rebuttal submission across the three cohorts. CoCoDet estimates are -0.16, $+0.23$, and $+0.62$ points in 2024, 2025, and 2026, respectively. Shared favorable evaluation therefore coexists with measurement-dependent author participation, with the clearest opposing associations in 2026.

Adjusted probability difference (pp)

Year	EditLens	CoCoDet
2024	−0.54 [−1.15, +0.08]	−0.16 [−0.73, +0.40]
2025	−0.78 [−1.29, −0.26]	+0.23 [−0.26, +0.73]
2026	−0.66 [−1.07, −0.25]	+0.62 [+0.23, +1.00]

Paper-clustered 95% CIs; $z = 0 \rightarrow 1$.

Figure 5: Author participation.

4.4 RESPONSE ORGANIZATION

Higher AI-likeness generally accompanies more response tables in pooled analyses, with patterns differing by measurement and text side (Figure 6). Among submitted responses, review-side EditLens scores are associated with 9.8%, 10.7%, and 17.0% more expected tables per score SD across 2024–2026, yielding successively larger point estimates. Review-side CoCoDet estimates are +1.8%, +4.0%, and +12.5% in 2024, 2025, and 2026, respectively, with intervals excluding zero in 2025–2026. Rebuttal-side estimates range from +7.1% to +14.4% for EditLens and +7.5% to +13.6% for CoCoDet, with all intervals excluding zero. Supplementary same-sample rebuttal-side comparisons distinguish this pooled pattern from variation among responses to the same paper: EditLens has negative within-paper coefficients, while CoCoDet's intervals include zero (Appendix H).

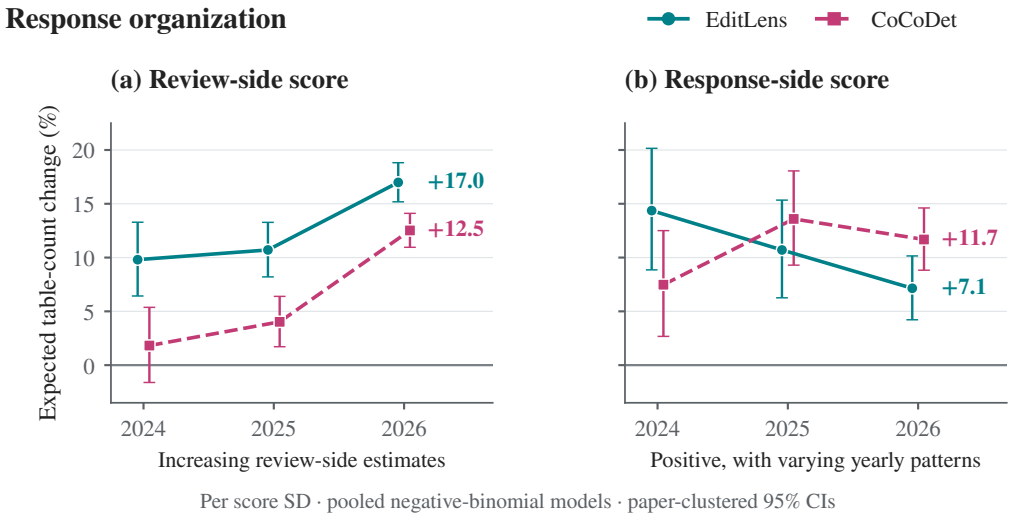


Figure 6: Response organization. Points show percent changes in expected response table count per score SD from pooled negative-binomial models, using (a) review-side and (b) response-side scores and adjusting for the scored text's length. Whiskers are pointwise paper-clustered 95% intervals. Lines connect independent annual estimates for visual guidance, not a fitted temporal trend. Panels share the same scale. CoCoDet uses native AI-versus-rest log-odds; sample filtering is specified in Appendix F. Samples and standardizations are model-specific; this is not a test of differences between sides or methods.

4.5 REVIEWER UPDATING

Primary models link higher review-side AI-likeness to less positive updating under both measurements; rebuttal-side associations differ (Figure 7). For the $z = 0$ to $z = 1$ contrast, EditLens corresponds to 0.89–1.92 percentage points less positive updating; CoCoDet corresponds to 0.98–3.15 points less. Both methods' intervals exclude zero in every cohort. The association persists with initial-rating adjustment and maximal-rating exclusion in 2024–2025, but CoCoDet's rank-based form is less consistent (Appendix G); tail-exclusion and nonlinear checks appear in Appendix K.

Rebuttal-side EditLens is negative throughout (-2.59 to -0.98 points), whereas CoCoDet spans -1.11 to $+0.16$ and excludes zero only in 2025 (Figure 7b). Shared negative associations concern review language; DETree supports negative response-side updating only in 2025 (Appendix D).

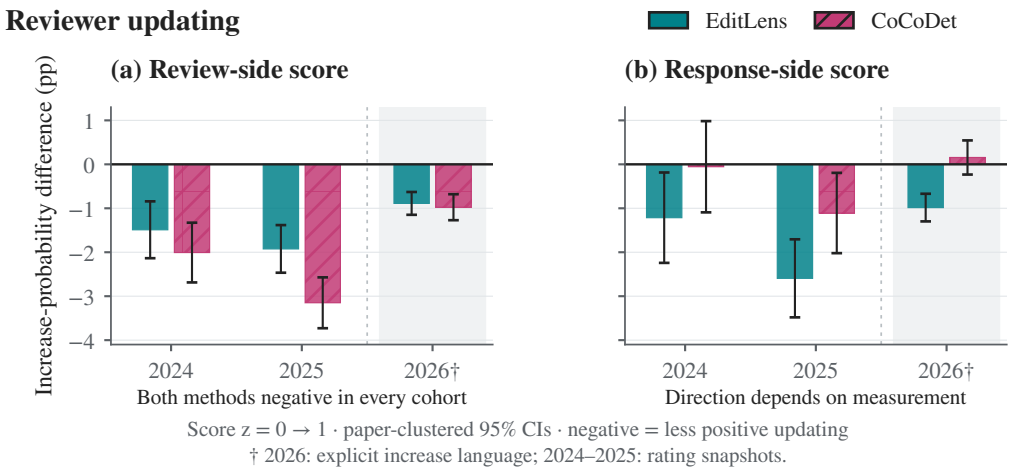


Figure 7: Reviewer updating. Bars show adjusted differences in positive-updating probability from score $z = 0$ to $z = 1$, averaged over observed covariates, in percentage points. Whiskers are pointwise paper-clustered 95% intervals. All models adjust for the scored text's length, with rating controls that differ by text side and cohort.

5 DISCUSSION

The results define *evaluation–response asymmetry*: the shared association with favorable evaluation does not extend uniformly to the exchange. Higher review-side AI-likeness accompanies higher ratings and less positive updating under both measurements, whereas participation has no shared direction and pooled table-count associations depend on measurement and text side.

The shared rating association could reflect preferential use of assistance to polish favorable judgments, changes in judgments during assistance, or textual styles associated with favorable evaluation. Selection predicts favorable judgments before assistance; a shift toward leniency predicts rating changes during assisted review composition. Review drafts and tool logs could test these contrasting predictions while accounting for stylistic differences.

Increasing rating and review-side table-count estimates across 2024–2026 suggest a growing role for AI-associated language in evaluation and response organization. The rating pattern also holds with recorded ratings throughout (Appendix B), but these descriptive patterns cannot distinguish changes in AI assistance from changing review conventions or establish a causal trend. Rating and review-side updating associations are more consistent across score-range and functional-form checks in 2026; early-cohort results include reversed EditLens rating contrasts (Appendix K).

Response availability is largely shared among reviews of the same paper, with small within-paper estimates (Appendix H), so paper-level response propensity matters when interpreting these associations. Rebuttal-side EditLens reverses from positive pooled to negative within-paper organization associations while CoCoDet's intervals include zero, and review-side within-paper associations are

clearly positive in 2026 (Appendix H). These patterns motivate examining shared response material versus reviewer-specific evidence. Tables answering requests for comparisons or experiments could be reused across replies without distinguishing responses to individual reviewers. EditLens's historical review-side table association points to established review and response conventions as a possible source of the later pattern.

The 2024–2025 initial-rating and ceiling checks weaken a ceiling-only explanation of less positive updating, although CoCoDet's 2024 rank-based result remains sensitive (Appendix G). Polished reviews might express already-settled judgments, or responses might leave rating-determining concerns unresolved. Initial confidence, concerns addressed, and reviewers' reasons for retaining scores could distinguish these accounts. Less upward revision does not imply poorer exchange quality.

For peer-review systems, these findings motivate evaluating assistance through review specificity, response relevance, and reasons for retaining or revising judgments; the ICLR 2025 randomized intervention shows that designed assistance can change interaction (Thakkar et al., 2026). Our observational findings identify complementary questions about routine reviewing: whether favorable language leaves authors clear issues to address, and whether visible response organization is followed by consideration of those issues. Supporting these functions would connect fluent scholarly communication to the exchange of evidence on which peer review depends.

6 Limitations

Measurement validity and selection. Uncalibrated scores are neither AI-use probabilities nor AI-written fractions; CoCoDet's log-odds need not track actual use linearly. Direct-generation endpoints do not validate natural editing, mixed authorship, or humanization; CoCoDet's matched training domain limits generalization of its near-ceiling endpoint discrimination. EditLens implementation and CoCoDet scale choices followed exploratory analyses (Appendix C). Rebuttal-domain historical validation is unavailable; DETree provides approximate, auxiliary evidence. Distinct training targets and a shared evaluation corpus mean agreement is not independent replication.

Historical controls. Historical cohorts lack verified non-use labels; 2023 overlaps the human endpoint and may include reviews finalized after ChatGPT's release. Historical nulls do not establish equivalence, and associations may reflect stable textual features. Rating contrasts concern recorded ratings rather than the primary pre-discussion outcome in 2024–2025.

Observational confounding. Paper quality, topic, reviewer expertise, confidence, and language background, assignment, and timing may confound these associations, which establish neither causality nor individual AI use. Separate outcome models cannot trace individual exchanges. Pre-discussion rating snapshots do not ensure review text predates responses or rating revisions.

Outcome and input coverage. Rating availability varies by cohort. Updating compares actual score increases in 2024–2025 with explicit increase language in 2026[†]; rating controls differ by text side, and ceiling checks cover only 2024–2025. Token-window selection varies across years and score distributions; EditLens truncates roughly 40% of rebuttals in the primary window, and DETree reads 512-token heads. Table counts indicate visible organization, not quality or effort.

Multiplicity and generalizability. Family-specific FDR corrections do not provide a joint error guarantee across all analyses. Findings from public ICLR reviews may not transfer to other venues, review forms, languages, or confidential reviewing systems.

7 Conclusion

Across three ICLR cohorts, AI-like review language is associated with higher ratings and less positive reviewer updating under both main measurements. Participation has no shared direction, and pooled table-count associations vary by measurement and text side. This *evaluation–response asymmetry* places favorable evaluation within a differentiated pattern of response and reconsideration: understanding AI-mediated peer review requires tracing how judgments are expressed and answered.

AI USE STATEMENT

For endpoint validation, DeepSeek-V4-Flash-0731 generated synthetic reviews; Qwen3.5-9B classified reviewer follow-up language. An OpenAI coding agent aided retrieval and discovery of related work, citation metadata verification, and subsequent manuscript polishing, LaTeX editing, and consistency checks. The authors developed the research ideas, wrote the initial manuscript, and remain responsible for every claim, analysis choice, citation, and submitted text.

ETHICS STATEMENT

Although the source records are public, attributing detector scores to individuals could create privacy and reputational harms. We therefore restrict interpretation to aggregate patterns; the measurements should not be used to identify or rank reviewers or authors, or to allege individual AI use or misconduct.

REPRODUCIBILITY STATEMENT

The Methods and appendices document cohort construction, measurement definitions, model specifications, sample flows, and sensitivity analyses. The supplement provides frozen aggregate outputs, de-identified record-level numerical data, and code for score computation and statistical analysis. Text and public identifiers are excluded; random identifiers retain analysis groupings. These data are not guaranteed to be unlinkable. Further materials will be released after anonymity is lifted.

REFERENCES

Mervat Abassy, Kareem Elozeiri, Alexander Aziz, Minh Ngoc Ta, Raj Vardhan Tomar, Bimarsha Adhikari, Saad El Dine Ahmed, Yuxia Wang, Osama Mohammed Afzal, Zhuohan Xie, Jonibek Mansurov, Ekaterina Artemova, Vladislav Mikhailov, Rui Xing, Jiahui Geng, Hasan Iqbal, Zain Muhammad Mujahid, Tarek Mahmoud, Akim Tsvigun, Alham Fikri Aji, Artem Shelmanov, Nizar Habash, Iryna Gurevych, and Preslav Nakov. LLM-DetectAIve: a tool for fine-grained machine-generated text detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 336–343. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-demo.35. URL https://aclanthology.org/2024.emnlp-demo.35/.

Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient zero-shot detection of machine-generated text via conditional probability curvature. In *International Conference on Learning Representations*, 2024.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Yihan Chen, Jiawei Chen, Guozhao Mo, Xuanang Chen, Ben He, Xianpei Han, and Le Sun. CoCoNUTS: Concentrating on content while neglecting uninformative textual styles for AI-generated peer review detection. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26921–26950. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1240.

Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with binoculars: Zero-shot detection of machine-generated text. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17519–17537. PMLR, 2024.

Yongxin He, Shan Zhang, Yixuan Cao, Lei Ma, and Ping Luo. DETree: Detecting human–AI collaborative texts via tree-structured hierarchical representation learning. In *Advances in Neural Information Processing Systems*, 2025.

Amir Hossein Kargaran, Nafiseh Nikeghbal, Jing Yang, and Nedjma Ousidhoum. Insights from the ICLR peer review and rebuttal process. *arXiv preprint arXiv:2511.15462*, 2025.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. 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.

Mathieu Louis, Tibo Vanleke, Vincent Ginis, and Andres Algaba. Rebuttals move peer-review scores, but initial-review structure bounds the movement. *arXiv preprint arXiv:2606.22166*, 2026.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24950–24962. PMLR, 2023.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*, 2024.

Siyuan Shen and Kai Wang. Detecting AI-generated content in academic peer reviews. *arXiv preprint arXiv:2602.00319*, 2026.

Katherine Thai, Bradley Emi, Elyas Masrour, and Mohit Iyyer. EditLens: Quantifying the extent of AI editing in text. *arXiv preprint arXiv:2510.03154*, 2025.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8(3):326–336, 2026. doi: 10.1038/s42256-026-01188-x. URL https://doi.org/10.1038/s42256-026-01188-x.

Sungduk Yu, Man Luo, Avinash Madasu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an LLM? benchmarking AI text detection in peer review. In *International Conference on Learning Representations*, 2026.

Unless stated otherwise, appendix intervals are pointwise, paper-clustered 95% confidence intervals. Interval exclusion of zero does not imply significance after multiple-testing correction; reported q-values use the analysis families specified with each analysis.

A Interaction data and sample accounting

Review inputs are canonical review-only texts from the collected versions. Author-response text linked to a review is called its rebuttal throughout; inputs exclude Markdown table blocks, with table count retained separately. Zero response tokens mark no usable text, which need not mean no author activity anywhere in the paper's discussion. Counts refer to review-linked response records, not unique authors or deduplicated messages.

The rating supports are $\{1, 3, 5, 6, 8, 10\}$ in 2024–2025 and $\{0, 2, 4, 6, 8, 10\}$ in 2026. Main coefficients use raw rating points within year. Figure 4 shows token-adjusted rating deviations in original rating points. The six ordered categories map to $\{0, .2, .4, .6, .8, 1\}$ only for the historical rating contrasts in Appendix C.

For 2024–2025 follow-up outcomes, incomplete snapshot pairs are excluded. For 2026†, the local classifier processes nonempty final replies; empty replies receive a rule-based negative label and uncertain outputs are excluded from the binary models.

Table 1: Observational sample. Responses are nonempty usable author responses before outcome-specific token windows and filters.

Year	Reviews	Papers	Responses	Response rate
2024	28,028	7,262	22,368	79.8%
2025	46,748	11,520	34,170	73.1%
2026	75,859	19,474	55,804	73.6%
Total	150,635	38,256	112,342	74.6%

B Rating-source sensitivity

For 2024–2025 we compare the pre-discussion snapshot with the value recorded at collection on exactly the same reviews. The main rating coefficients are larger under the pre-discussion source for both primary methods. Direct models of initial minus recorded rating reject equality of the score coefficients in all six method–year cells (including auxiliary DETree) after BH correction. The pre-discussion source anchors the main 2024–2025 rating outcome before the exchange; this comparison shows source sensitivity without identifying why ratings changed.

Table 2: Same-sample rating-source comparison. Entries are rating-point changes per score SD; parentheses contain nominal p-values.

Method	Year	Pre-disc.	Recorded	Difference p
CoCoDet	2024	+.0980 (2.54e-18)	+.0769 (1.97e-11)	1.03e-5
CoCoDet	2025	+.1257 (2.79e-52)	+.0944 (2.20e-27)	2.32e-12
EditLens	2024	+.0300 (.0128)	+.0041 (.742)	2.41e-6
EditLens	2025	+.0517 (6.11e-9)	+.0129 (.176)	2.79e-17
DETree	2024	+.0059 (.620)	-.0109 (.380)	.00331
DETree	2025	+.0102 (.246)	-.0213 (.0263)	3.00e-11

Using recorded ratings in all three cohorts provides an additional view of the annual pattern (Table 3). These models use the same three-detector sample within each year, a 256–768-token review window, and OLS adjusted for review length, with scores standardized within year and intervals clustered by paper. Their samples include reviews without initial-rating snapshots and therefore differ from the paired comparison above. EditLens and CoCoDet retain successively larger point estimates; DETree falls slightly in 2025 before becoming positive in 2026. The former pattern thus also appears under a common recorded-rating definition, although these annual estimates do not constitute a fitted time-trend test.

Table 3: Recorded-rating associations in all three cohorts. Entries are rating points per score SD [95% CI], adjusted for review length. The common samples contain 17,728, 30,940, and 52,383 reviews in 2024, 2025, and 2026, respectively.

Measurement	2024	2025	2026
EditLens	$+.007[-.017, +.031]$	$+.013[-.006, +.032]$	$+.128[+.112, +.145]$
CoCoDet	$+.076[+.054, +.098]$	$+.094[+.077, +.111]$	$+.217[+.202, +.231]$
DETree	$-.008[-.032, +.016]$	$-.022[-.040, -.003]$	$+.054[+.038, +.070]$

C Measurement implementation and historical controls

Fast-DetectGPT supplies a conditional-probability discrepancy score (Bao et al., 2024); Binoculars supplies a perplexity/cross-perplexity ratio (Hans et al., 2024). Fast-DetectGPT uses the paired Llama-3.1-8B base and Instruct models with a 1,024-token head; Binoculars uses the Falcon-7B base and Instruct pair with a 512-token head. Their inference specifications are held fixed across the historical and later cohorts, and Binoculars scores are oriented so that higher values indicate more AI-like text.

The public EditLens Llama checkpoint produces four edit-magnitude buckets; the primary score is their expected value. Inference uses a 1,024-token head. Our runner agrees with the current public inference script on 24 check texts to a maximum absolute score difference of 1.2×10^{-7}. Against the upstream repository's 2,400 precomputed validation scores, however, the mean absolute difference is 3.0×10^{-4}, exceeding the prespecified 10^{-4} tolerance; the original inference environment is not fully documented. All study scores were recomputed with the fixed public checkpoint and our recorded environment, rather than imported from that validation score file. This remains a revised/post-hoc implementation, not an exact reproduction of the historical scores. CoCoDet uses the released checkpoint with a ModernBERT backbone and a maximum input of 2,048 model tokens. Its complete logits and three class probabilities are retained; downstream models use AI-versus-rest log-odds computed with a numerically stable log-sum-exp. This scale was selected after inspecting initial analyses; raw p_{AI} and rank-normal scores are retained as sensitivity specifications. The transformation preserves the underlying ordering in exact arithmetic and does not calibrate AI-use intensity.

The endpoint's directly generated reviews were produced from corresponding paper content without supplying ratings, reviewer identities, confidence, or review metadata. Metrics and intervals use the paper as the bootstrap unit. Fast-DetectGPT and Binoculars also retain endpoint ranking signal; this alone does not establish the temporal specificity of downstream associations. The historical comparison uses 10,202 reviews from 2022 and 14,339 from 2023 (24,541 total) and models 2024–2026 as the post-era group. Figure S2 reports formal post-minus-historical contrasts on the natural effect scale.

Figure 3 uses frozen endpoint display points. Its real-review histograms use complete 0.1-z bins extending to the nearest grid edge, without including observations outside the displayed P5–P95 range. Display trimming changes neither standardization nor regression samples.

EditLens's raw post-minus-historical rating contrast is +.0083 aligned units per SD, and all three transformations are positive; its participation contrasts range from -1.00 to $-.81$ percentage points and exclude zero. Its table-count associations exist in both eras. For CoCoDet's native log-odds score, the all-recorded rating contrast is +.0162 [.0127, .0198] aligned units per SD. The post-era interaction is also positive for table count but not clearly different from zero for participation. Historical CoCoDet rating estimates are $-.024$ $[-.066, +.018]$ in 2022 and $+.008$ $[-.027, +.043]$ in 2023; participation intervals also cross zero. The 2023 table-count estimate is negative but its interval crosses zero under the retained trimming rule. The original probability and rank-based specifications provide scale sensitivities, rather than calibrated estimates of AI use.

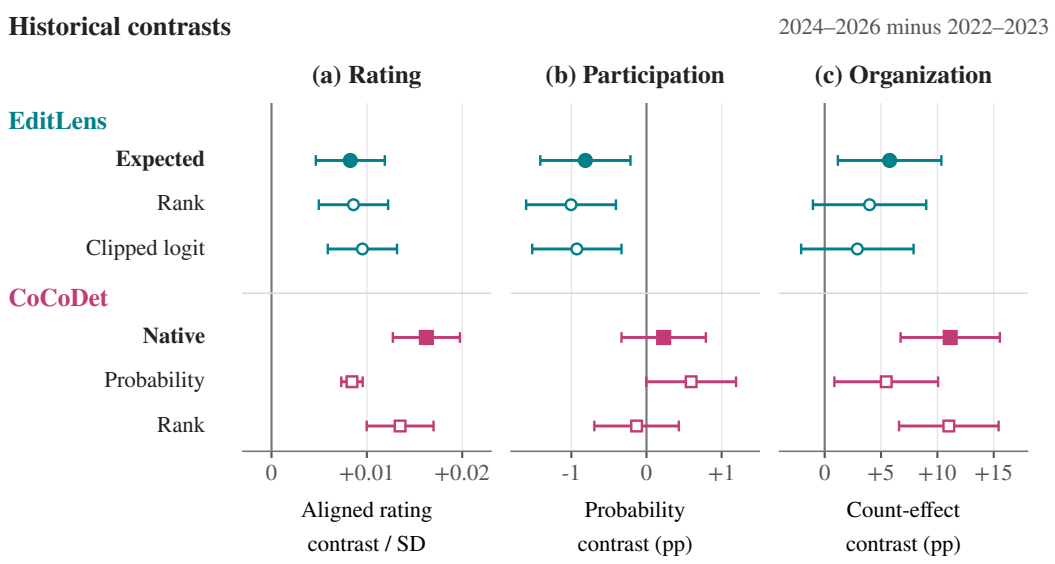


Figure S2: Historical-control contrasts. Points compare review-side associations in 2024–2026 with 2022–2023. Filled symbols mark the primary EditLens expected score and CoCoDet native AI-versus-rest log-odds; open symbols show sensitivity scales. Rank denotes rank-normal scores; clipped logit is an EditLens sensitivity, and probability denotes the original CoCoDet AI probability. Shared rows align the three outcomes. Whiskers are paper-clustered 95% contrast intervals, using the delta method for nonlinear effects. Rating uses recorded ratings mapped to 0–1 in every year. Participation contrasts are differences in average probability effects ($z = 0 \to 1$, pp); organization contrasts are percentage-point differences between expected-count percentage effects per score SD. Zero denotes no era contrast, not absence of association within either era.

C.1 Evidence from measurements not retained as primary proxies

Historical associations help assess whether a score–outcome relation is specific to the period of widespread LLM use. Table 4 summarizes 2022–2023 rating and participation estimates for Fast-DetectGPT, Binoculars, and LLM-DetectAIve (LLMDA), a supervised discriminator (Abassy et al., 2024); Table 5 compares these cohorts with 2024–2026 using score-by-era interactions and year effects. Within each analysis, implementation and statistical specifications are held fixed across eras. Rating uses recorded values in every year; era contrasts use the six-category $[0, 1]$ mapping. Historical updating is unavailable because comparable audited rating snapshots do not exist. The 2022 cohort is the stronger temporal control; 2023 is a lower-exposure cohort rather than verified absence of AI use.

Fast-DetectGPT and Binoculars already have negative rating and participation associations in both historical years. Conversely, LLMDA has positive historical rating associations despite sharing the main results' direction; its endpoint AUROC is .657 [.645,.669] and AP is .354 [.340,.370]. Its primary aligned-rating era contrast includes zero, while the rank-normal contrast is negative, $-.00399$ $[-.00758, -.00040]$. Thus the direction of an association alone does not justify retaining a measurement or attributing its relation to AI use.

Table 4: Historical rating and participation associations for measurements not retained as primary proxies. Entries are estimates [95% CI]; rating uses recorded-rating points per score SD, and participation uses probability differences from $z = 0$ to $z = 1$ (pp). Rating samples contain 6,611/9,395 reviews and participation samples 6,543/9,301 in 2022/2023.

Method	Year	Rating	Participation
Fast-DetectGPT	2022	$-.107[-.148, -.065]$	$-1.14[-2.05, -.24]$
	2023	$-.059[-.095, -.024]$	$-1.30[-2.08, -.53]$
Binoculars	2022	$-.156[-.198, -.114]$	$-1.47[-2.43, -.51]$
	2023	$-.084[-.120, -.049]$	$-1.32[-2.08, -.57]$
LLMDA	2022	$+.065[+.027, +.103]$	$+.63[-.31, +1.56]$
	2023	$+.079[+.040, +.118]$	$-.13[-.79, +.54]$

The direct era comparisons show less negative rating slopes for both likelihood methods and no shared strengthening of their participation associations. Fast-DetectGPT's response-side organization relation does become more negative, an exception not corroborated by Binoculars. LLMDA has no response-side table model in this comparison.

Table 5: Formal era contrasts: 2024–2026 minus 2022–2023. Rating uses aligned units; participation uses pp. Organization pre/post entries are percent count effects, and their difference is in percentage points. The final column tests the link-scale interaction, not the displayed nonlinear contrast.

Method	Relation	Pre	Post	Difference [95% CI]	Link q
Fast-DetectGPT	Rating	$-.0100$	$-.0043$	+.0056 [+.0021, +.0092]	0.007
Binoculars	Rating	$-.0144$	$-.0086$	+.0058 [+.0022, +.0094]	0.007
Fast-DetectGPT	Participation	-1.23	$-.86$	+.37 [$-.29$, +1.02]	0.028
Binoculars	Participation	-1.41	-1.66	$-.25$ [$-.92$, +.41]	0.213
Fast-DetectGPT	Org., review	-3.25	-4.94	-1.69 [-5.94, +2.55]	0.575
Binoculars	Org., review	-3.25	-4.03	$-.78$ [-5.00, +3.44]	0.716
Fast-DetectGPT	Org., response	$-.78$	-10.80	-10.02 [-16.87, -3.16]	0.007
Binoculars	Org., response	-1.42	$+.47$	+1.89 [-5.18, +8.96]	0.689
LLMDA	Rating	$+.0088$	$+.0115$	+.0027 [$-.0009$, +.0063]	0.430
LLMDA	Participation	$+.11$	$+.44$	+.33 [$-.25$, +.91]	0.625
LLMDA	Org., review	-3.16	$-.31$	+2.85 [-2.57, +8.27]	0.463

Original BH families are retained: eight interactions for the two likelihood methods; the supervised analysis uses its own declared detector family.

For nonlinear models, the link-scale interaction and natural-scale marginal contrast test different quantities: Fast-DetectGPT's participation interaction has $q = .028$, while the probability contrast includes zero. Historical controls constrain individual relations rather than assigning universal validity to a detector. EditLens, for example, already has positive historical table-count associations. Primary-measurement selection remains post hoc with respect to parts of the observed evidence.

D DETREE AUXILIARY EVIDENCE

DETree represents generation processes in a tree-structured embedding and uses prototype retrieval for few-shot adaptation (He et al., 2025). We use a fixed, auditable approximate reproduction with the released model and prototype bank. The score is an uncalibrated retrieval score. Six prespecified layer/k variants reproduce the main qualitative organization results. This auxiliary analysis uses the same observational corpus, not an independent validation sample.

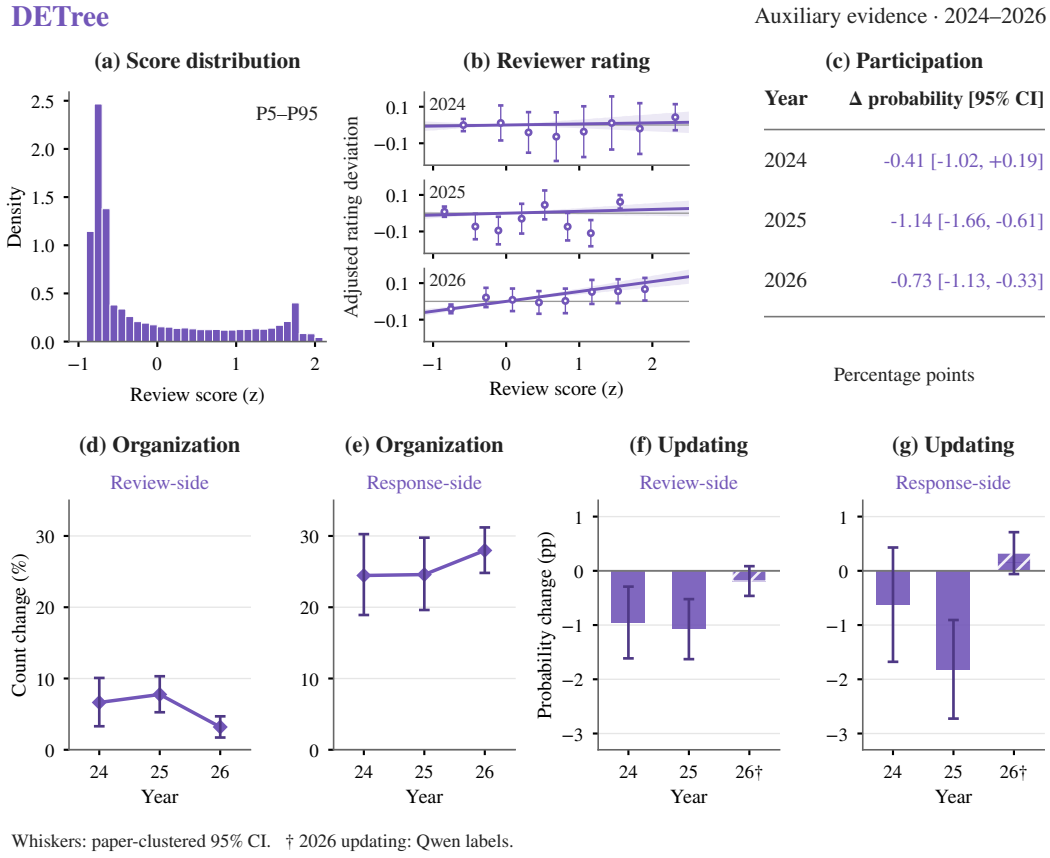


Figure S3: DETree auxiliary evidence. (a) Within-year standardized review scores, pooled over 2024–2026; P5–P95 display, full-sample density normalization. (b) Full-sample rating fits and 95% slope bands on identical axes; points show eight equal-width P5–P95 bins of token-adjusted rating deviations, with descriptive 95% intervals conditional on the token adjustment. Ratings are initial in 2024–2025 and recorded in 2026. (c,f,g) Probability changes from $z = 0$ to $z = 1$ (pp); (d,e) expected-count changes per score SD (%). All intervals are paper-clustered; probability intervals use the delta method. Updating controls for recorded rating on the review side, and initial (2024–2025) or recorded (2026) rating on the response side. Hatched 2026 updating bars use Qwen labels.

Under the primary year-specific rating definition, DETree is near zero in 2024–2025 and positive in 2026. Its review-side participation estimates are negative in 2025–2026, and its review-side table-count estimates are positive in all three years after historical nulls. Rebuttal-side table-count estimates are strongly positive at the pooled level, but within-paper diagnostics are near zero; rebuttal-side updating is clearly negative only in 2025. Review-side updating is negative in 2024–2025, while its 2026 interval includes zero. The method therefore supplies corroboration for the organization associations alongside differences by year and text side in evaluation and updating.

E QWEN LABEL AUDIT

The 2026† outcome uses Qwen3.5-9B to classify final reviewer replies as `RAISE` when they explicitly report or commit to an increase in overall rating, and `NO_CHANGE` otherwise; uncertain replies are excluded. Here `NO_CHANGE` means no explicit overall-rating increase, not verified retention of the same score; decreases and other acknowledgments also fall outside `RAISE`. Empty replies receive the rule-based negative label described in Appendix A. An author manually checked a stratified random sample of 100 nonempty replies from each predicted class (seed 20260903), with predicted labels visible, and found no errors among the 200 inspected replies.

Qwen3.5-9B

```
SYSTEM PROMPT

You will receive exactly one complete academic reviewer reply and must
produce exactly one classification label. The entire message is one
indivisible reply and one classification target, even when it contains
numbered lists, bullets, headings, multiple questions, or point-by-point
responses. Never split it into separate items.

Classify whether the reviewer explicitly raised the review's overall
score or overall rating after an author rebuttal.

Labels:
- RAISE: The reviewer clearly states that they already
raised/increased/upgraded the overall score/rating, or makes an
unambiguous decision, commitment, or plan to do so (for example, "I will
raise" or "I plan to increase"). A numerical change counts only when the
new value is higher than the old value.
- NO_CHANGE: There is no explicit upward overall-score/rating change.
This includes an unchanged score, an explicit score decrease, a decision
to keep the score after considering an increase, no overall-score/rating
discussion, or changes only to other metrics.
- UNCERTAIN: The reply concerns a possible upward overall-score/rating
change, but the direction or commitment is unresolved, ambiguous,
conditional, or non-committal.

Rules:
1. Judge only the overall recommendation score or overall rating. Changes
to confidence, soundness, correctness, contribution, novelty,
presentation, or any other subscore do not count.
2. A clear overall-score decrease is NO_CHANGE. "May," "might,"
"consider," or similarly unresolved upward language is UNCERTAIN. A score
change with no direction is also UNCERTAIN.
3. Use the reviewer's final resolved position. Remaining criticism does
not cancel an explicit increase, but praise, satisfaction, confidence, or
support for acceptance alone does not establish one.

Return exactly one bare label: RAISE, NO_CHANGE, or UNCERTAIN. Return no
JSON, punctuation, explanation, markdown, or additional text.



USER PROMPT

This request contains exactly one complete reviewer reply and requires
exactly one label. Classify the single reviewer_reply string below as a
whole:

{"reviewer_reply": "<complete reviewer reply>"}
```

F Detailed statistical specification

Review-side models use texts with 256–768 Llama-3.1-8B tokens; rebuttal-side models use 768–1,280 tokens. Counts include the beginning-of-sequence token and precede truncation. Standardization follows outcome-specific exclusions. Rank-normal scores use average ranks for ties followed by an inverse-normal transform; clipped-logit scores transform bounded scores away from the endpoints before standardization.

Submission and table models retain score percentiles 0.5–99.5; for CoCoDet, the original p_{AI}-based retained samples are held fixed across score scales, including ties at the trimming boundaries. Table models exclude counts above 30. Response-side models require usable responses, and follow-up excludes missing or uncertain outcomes. Diagnostics examine EditLens truncation and CoCoDet untrimmed and rank-based table-count specifications.

BH correction spans year cells within each rating method and specification, the 15 recorded-rating, participation, and table-count cells across 2022–2026 within each review-side detector analysis, and year cells within each review-side updating specification. The rebuttal analysis corrects its 18 primary models as one family. CoCoDet historical interaction tests are corrected across nine relation–scale combinations.

G Initial-rating controls for reviewer updating

Supplementary review-side analyses examine whether limited upward room at higher initial ratings accounts for the negative updating association in 2024–2025. The sample requires usable author responses, complete rating pairs and scores for both main measurements, 256–768 review tokens, and at least two eligible reviews per paper. Scores and review length are standardized after selection; pooled linear probability models (LPMs) and logistic models adjust for categorical initial rating and review length, with paper-clustered intervals. Effects below compare each score's empirical P10 and P90. These exploratory checks use their own matched samples and do not replace the recorded-rating-adjusted primary estimates in Figure 7.

For both primary scores, negative associations persist with initial-rating adjustment (Table 6). For EditLens, they also persist under rank-normal scores and after excluding reviews initially rated 10: the latter raw-score LPM contrasts are -3.45 and -4.99 percentage points in 2024 and 2025. Stratified analyses locate the clearest negative associations in the middle initial-rating range (4–6); estimates vary across rating strata. Thus, initial-rating adjustment and removal of maximal ratings retain the negative pooled association.

Table 6: Updating associations after initial-rating adjustment. Entries are P10–P90 probability differences in percentage points [95% CI] from pooled models with categorical initial rating and review length. Both methods use 12,658 reviews from 4,540 papers in 2024 and 21,233 reviews from 7,240 papers in 2025.

Method	Year	Linear probability model	Logistic model
EditLens	2024	$-3.42[-4.90, -1.94]$	$-3.63[-5.25, -2.02]$
CoCoDet	2024	$-.97[-1.55, -.39]$	$-1.24[-2.07, -.41]$
EditLens	2025	$-4.98[-6.50, -3.47]$	$-5.16[-6.72, -3.59]$
CoCoDet	2025	$-2.63[-3.50, -1.75]$	$-3.01[-4.08, -1.94]$

For CoCoDet's native log-odds score, excluding initially maximal ratings retains negative LPM contrasts of -0.97 and -2.63 points. Scale sensitivity remains: original-probability P10–P90 contrasts are near zero, while rank-normal LPM contrasts are $+0.13$ $[-1.64, +1.90]$ in 2024 and -3.25 $[-4.79, -1.70]$ in 2025. These checks do not extend to 2026†, which lacks initial-rating snapshots.

G.1 Raise, retain, and lower: matched snapshot decomposition

To identify what accompanies less positive updating, we decompose actual 2024–2025 rating changes into raise, retain, and lower on identical records. The samples require usable responses, complete initial and recorded ratings, both primary scores, 256–768 review tokens, and at least two

eligible reviews per paper. All three linear probability models include review length and categorical initial rating; their score coefficients and P10-to-P90 contrasts sum to zero. Pooled and within-paper fits share the sample and score SD. Within-paper covariance corrections account for the absorbed paper intercepts. CoCoDet uses native AI-versus-rest log-odds; EditLens uses its expected score.

Table 7: Initial-rating-adjusted snapshot changes: empirical P10-to-P90 probability differences (pp) from three linear probability models on the same records. Effects sum to zero across raise/retain/lower. Samples: 12,658 reviews/4,540 papers in 2024; 21,233/7,240 in 2025.

Year	Method	Model	Raise	Retain	Lower
2024	EditLens	Pooled	−3.42 [−4.90, −1.94]	+3.28 [+1.74, +4.82]	+.14 [−.36, +.64]
2024	EditLens	Within	−1.61 [−3.71, +.50]	+1.45 [−.78, +3.67]	+.16 [−.55, +.87]
2024	CoCoDet	Pooled	−.97 [−1.55, −.39]	+.93 [+.29, +1.57]	+.04 [−.26, +.34]
2024	CoCoDet	Within	−.85 [−1.77, +.07]	+.71 [−.30, +1.73]	+.13 [−.26, +.53]
2025	EditLens	Pooled	−4.98 [−6.50, −3.47]	+4.79 [+3.24, +6.33]	+.20 [−.22, +.62]
2025	EditLens	Within	−3.26 [−5.35, −1.16]	+3.47 [+1.30, +5.63]	−.21 [−.87, +.45]
2025	CoCoDet	Pooled	−2.63 [−3.50, −1.75]	+2.84 [+1.95, +3.73]	−.21 [−.44, +.02]
2025	CoCoDet	Within	−1.34 [−2.58, −.10]	+1.74 [+.47, +3.01]	−.40 [−.77, −.04]

In pooled comparisons, both primary scores associate with fewer increases and more unchanged ratings in both years. The decrease contrasts have intervals including zero. Within-paper comparisons show the same fewer-raise, more-retain pattern in point estimates for both measurements; the corresponding intervals exclude zero in 2025 and include zero in 2024. The 24 coefficient tests form one exploratory BH family; the displayed intervals are pointwise. There is no analogous 2026 three-way decomposition: its reply-language labels do not identify actual retain/lower outcomes.

H WITHIN-PAPER COMPARISONS

These comparisons distinguish pooled associations from differences among reviews or responses to the same paper. Within each year and outcome, we intersect the two primary methods' eligible records, retain papers with at least two records, and then standardize each score and text length. Both methods and both estimators therefore use identical records within each comparison; different outcomes may use different samples. Papers with constant outcomes are retained.

Pooled OLS and paper-demeaned OLS share the outcome transform, score SD, and controls. Binary outcomes use linear probability models, and table counts use $\log(1 + \text{count})$, distinct from the primary logit/NB effect scales. Review-side updating controls categorical recorded rating; response-side updating controls initial rating in 2024–2025 and recorded rating in 2026†. Paper-clustered CR1 covariance for within-paper fits accounts for the absorbed paper intercepts. The 72 coefficient tests form one exploratory BH family.

Table 8: Review-side rating: matched-sample pooled and within-paper associations. Coefficients are rating points per score SD [95% CI]. Both measurements and both models use identical records within year.

Method	Year	N	Papers	Pooled	Within paper
EditLens	2024	15,935	5,704	+.032 [+.008, +.057]	+.049 [+.013, +.085]
CoCoDet	2024	15,935	5,704	+.101 [+.078, +.123]	+.101 [+.066, +.135]
EditLens	2025	29,384	9,946	+.053 [+.035, +.070]	+.086 [+.062, +.111]
CoCoDet	2025	29,384	9,946	+.127 [+.110, +.143]	+.138 [+.114, +.161]
EditLens	2026	50,403	17,241	+.130 [+.114, +.147]	+.179 [+.156, +.202]
CoCoDet	2026	50,403	17,241	+.217 [+.202, +.232]	+.229 [+.207, +.250]

Table 9: Review-side participation: matched-sample pooled and within-paper associations. Coefficients are percentage points (linear probability model) per score SD [95% CI]. Both measurements and both models use identical records within year.

Method	Year	N	Papers	Pooled	Within paper
EditLens	2024	16,187	5,844	−.58 [−1.22, +.07]	−.03 [−.20, +.15]
CoCoDet	2024	16,187	5,844	−.22 [−.82, +.38]	+.00 [−.14, +.13]
EditLens	2025	28,934	9,886	−.77 [−1.30, −.24]	−.09 [−.21, +.02]
CoCoDet	2025	28,934	9,886	+.34 [−.16, +.83]	−.12 [−.23, −.01]
EditLens	2026	49,459	17,052	−.68 [−1.10, −.27]	−.04 [−.12, +.03]
CoCoDet	2026	49,459	17,052	+.58 [+.19, +.97]	−.05 [−.12, +.03]

Papers with differing response availability among eligible reviews: 48/5,844 (2024), 115/9,886 (2025), and 136/17,052 (2026).

Table 10: Review-side organization: matched-sample pooled and within-paper associations. Coefficients are $\log(1 + \text{count})$ units per score SD [95% CI]. Both measurements and both models use identical records within year.

Method	Year	N	Papers	Pooled	Within paper
EditLens	2024	12,868	4,662	+.030 [+.021, +.040]	−.003 [−.013, +.007]
CoCoDet	2024	12,868	4,662	+.003 [−.006, +.013]	−.007 [−.016, +.003]
EditLens	2025	20,909	7,191	+.039 [+.031, +.048]	+.003 [−.005, +.012]
CoCoDet	2025	20,909	7,191	+.017 [+.008, +.026]	+.002 [−.007, +.010]
EditLens	2026	36,258	12,528	+.077 [+.070, +.084]	+.022 [+.015, +.029]
CoCoDet	2026	36,258	12,528	+.060 [+.052, +.067]	+.025 [+.018, +.031]

Table 11: Review-side updating: matched-sample pooled and within-paper associations. Coefficients are percentage points (linear probability model) per score SD [95% CI]. Both measurements and both models use identical records within year. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Method	Year	N	Papers	Pooled	Within paper
EditLens	2024	12,658	4,540	−1.42 [−2.09, −.75]	−.92 [−1.95, +.11]
CoCoDet	2024	12,658	4,540	−1.90 [−2.47, −1.33]	−1.85 [−2.82, −.88]
EditLens	2025	21,233	7,240	−1.97 [−2.54, −1.39]	−2.14 [−3.01, −1.26]
CoCoDet	2025	21,233	7,240	−3.26 [−3.81, −2.71]	−3.00 [−3.84, −2.16]
EditLens	2026†	36,492	12,574	−.91 [−1.20, −.63]	−.97 [−1.40, −.54]
CoCoDet	2026†	36,492	12,574	−.89 [−1.15, −.62]	−.88 [−1.27, −.49]

Table 12: Response-side organization: matched-sample pooled and within-paper associations. Coefficients are $\log(1 + \text{count})$ units per score SD [95% CI]. Both measurements and both models use identical records within year.

Method	Year	N	Papers	Pooled	Within paper
EditLens	2024	4,391	1,895	+.039 [+.022, +.056]	−.041 [−.073, −.009]
CoCoDet	2024	4,391	1,895	+.019 [+.002, +.036]	−.017 [−.043, +.008]
EditLens	2025	6,004	2,589	+.032 [+.017, +.048]	−.049 [−.074, −.025]
CoCoDet	2025	6,004	2,589	+.046 [+.030, +.061]	−.008 [−.031, +.015]
EditLens	2026	11,722	4,820	+.032 [+.020, +.045]	−.058 [−.077, −.039]
CoCoDet	2026	11,722	4,820	+.048 [+.036, +.061]	−.006 [−.022, +.010]

Table 13: Response-side updating: matched-sample pooled and within-paper associations. Coefficients are percentage points (linear probability model) per score SD [95% CI]. Both measurements and both models use identical records within year. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Method	Year	N	Papers	Pooled	Within paper
EditLens	2024	4,332	1,869	−1.74 [−3.00, −.48]	−2.05 [−5.20, +1.09]
CoCoDet	2024	4,332	1,869	−.39 [−1.65, +.87]	+1.09 [−1.45, +3.63]
EditLens	2025	6,127	2,638	−3.55 [−4.68, −2.43]	−3.14 [−6.00, −.27]
CoCoDet	2025	6,127	2,638	−1.19 [−2.35, −.03]	+1.08 [−1.30, +3.45]
EditLens	2026†	11,939	4,896	−1.25 [−1.67, −.82]	−.98 [−2.13, +.16]
CoCoDet	2026†	11,939	4,896	+.14 [−.32, +.59]	+1.05 [+.05, +2.06]

For the positive CoCoDet within-paper estimate in 2026†, $q = .060$ after BH adjustment over the 72 exploratory coefficient tests.

Rating associations remain positive within papers for both measurements. Participation estimates shrink substantially, with most within-paper intervals including zero. Fewer than 1.2% of papers in each year have differing participation outcomes among eligible reviews (Table 9). This limited within-paper outcome variation constrains conclusions about selective responses to individual reviewers; the small estimates do not rule out paper-level associations. Review-side organization is clearly positive within papers only in 2026. On the response side, EditLens reverses from positive pooled to negative within-paper estimates, while all three CoCoDet within-paper intervals include zero. Auxiliary DETree likewise has no clearly nonzero within-paper response-organization association in its separate method-available samples (Appendix D). If authors share tables across responses, this contributes to overall activity but little within-paper variation. The pooled comparisons describe overall variation in table use, while the within-paper comparisons describe its distribution among responses to the same paper.

Review-side updating remains negative in point estimates, with EditLens 2024 crossing zero. Response-side EditLens is negative but clearly separated from zero only in 2025. CoCoDet changes from negative pooled estimates in 2024–2025 to positive within-paper point estimates; its 2026† interval excludes zero, but the corresponding BH-adjusted $q = .060$ in the 72-test family does not meet the .05 threshold. The 2026 estimate concerns explicit increase language; the earlier estimates concern actual rating changes.

I PRIMARY EFFECT ESTIMATES

Table 14: Primary effect estimates. Probability differences compare score $z = 0$ with $z = 1$, averaged over observed covariates; rating and count effects are per score SD. Intervals use paper-clustered covariance. The p-values are nominal; BH correction follows the families in Appendix F. Sample sizes appear in Table 15.

		EditLens				CoCoDet			
Side	Year	Est.	CI low	CI high	p	Est.	CI low	CI high	p
Reviewer rating rating points per score SD									
Rev.	2024	+0.030	+0.006	+0.054	.013	+0.098	+0.076	+0.120	2.54e-18
Rev.	2025	+0.052	+0.034	+0.069	6.11e-9	+0.126	+0.110	+0.142	2.79e-52
Rev.	2026	+0.128	+0.112	+0.145	7.98e-54	+0.217	+0.202	+0.231	1.18e-179
Rebuttal submitted percentage-point difference									
Rev.	2024	-0.537	-1.149	+0.075	.082	-0.165	-0.728	+0.399	.566
Rev.	2025	-0.776	-1.292	-0.261	.003	+0.232	-0.261	+0.725	.358
Rev.	2026	-0.658	-1.066	-0.251	.001	+0.616	+0.230	+1.002	.002
Follow-up rating increase percentage-point difference									
Rev.	2024	-1.489	-2.136	-0.842	9.75e-6	-2.007	-2.687	-1.328	1.79e-8
Rev.	2025	-1.924	-2.466	-1.382	5.74e-12	-3.150	-3.729	-2.571	1.86e-25
Rev.	2026†	-0.889	-1.148	-0.629	2.16e-10	-0.975	-1.271	-0.680	1.39e-9
Resp.	2024	-1.214	-2.242	-0.186	.023	-0.055	-1.093	+0.983	.918
Resp.	2025	-2.594	-3.481	-1.706	2.65e-8	-1.107	-2.021	-0.194	.019
Resp.	2026†	-0.984	-1.298	-0.670	2.10e-8	+0.156	-0.232	+0.544	.426
Rebuttal table count expected-count change (%)									
Rev.	2024	+9.803	+6.428	+13.286	4.35e-9	+1.821	-1.612	+5.373	.302
Rev.	2025	+10.709	+8.204	+13.272	2.97e-18	+4.026	+1.714	+6.392	5.79e-4
Rev.	2026	+16.986	+15.182	+18.818	3.71e-87	+12.522	+10.956	+14.109	3.51e-61
Resp.	2024	+14.366	+8.857	+20.154	9.88e-8	+7.476	+2.673	+12.504	.002
Resp.	2025	+10.707	+6.258	+15.341	1.17e-6	+13.591	+9.295	+18.056	9.28e-11
Resp.	2026	+7.143	+4.217	+10.152	1.04e-6	+11.681	+8.825	+14.612	6.36e-17

J ANALYSIS-SAMPLE COUNTS

Table 15: Full cohorts and final analysis-cell counts. "Both" denotes equal review counts for EditLens and CoCoDet; paper counts can differ after method-specific filtering.

Outcome	Side	Method	2024	2025	2026
Full cohort	—	—	28,028	46,748	75,859
Reviewer rating	Review	Both	17,022	30,693	52,383
Rebuttal submitted	Review	EditLens	17,550	30,630	51,859
Rebuttal submitted	Review	CoCoDet	17,550	30,658	52,121
Rating increase	Review	Both	13,558	22,296	38,127
Rating increase	Response	Both	6,361	9,206	16,718
Table count	Review	EditLens	13,993	22,267	38,122
Table count	Review	CoCoDet	13,993	22,279	38,315
Table count	Response	Both	6,578	9,198	16,674

K SCORE-RANGE AND MODEL-SPECIFICATION SENSITIVITY

These exploratory analyses assess how the primary associations depend on score range, transformation, functional form, and text-input coverage. They were designed after the main analyses to characterize the scope of the primary findings. Confidence intervals are pointwise; BH-adjusted tests use the families specified below and in Appendix F.

K.1 Empirical P10-to-P90 contrasts and functional form

Tables 16 and 17 express the full-sample models over each score's observed P10-to-P90 range. Probability contrasts average over observed covariates, with delta-method intervals using the full covariance. These predictions retain the full estimation sample and therefore do not independently test tail influence. The two detectors' percentile spans need not represent equal changes in actual AI use.

The linear-score columns report the primary-model contrasts. On identical records, the quadratic-score models add a squared standardized-score term while retaining the model family, dispersion rule, controls and clustering. The 36 quadratic-term tests share one BH family; these tests assess curvature, not whether the displayed percentile contrast is nonzero.

For EditLens rating, the quadratic-model P10-to-P90 contrast is negative in 2024 ($-.096$ $[-.168, -.023]$) and includes zero in 2025 ($+.041$ $[-.007, +.088]$), despite positive full-sample linear slopes. CoCoDet rating contrasts remain positive in all three years. Review-side updating contrasts include zero for both methods in 2024, but remain negative in 2025–2026. Both rating contrasts remain positive in 2026. Thus early-cohort results are more dependent on functional form; neither linear nor quadratic contrasts establish monotonicity throughout the score range.

Table 16: P10-to-P90 contrasts and functional-form diagnostics (review side): empirical P10-to-P90 contrasts from linear-score and quadratic-score models on identical primary samples. Rating: points; participation/updating: pp; organization: expected-count percent change. The linear column gives primary-model predictions, not central-80% refits; sample sizes appear in Table 15. The final column is the quadratic-term Wald test, BH-adjusted over all 36 shape tests. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Method	Year	Linear	Quadratic	Shape q
Rating	EditLens	2024	+.066 [+.014, +.118]	−.096 [−.168, −.023]	< .001
Rating	CoCoDet	2024	+.106 [+.082, +.129]	+.115 [+.066, +.163]	0.709
Participation	EditLens	2024	−1.19 [−2.55, +.16]	−1.16 [−3.01, +.69]	0.974
Participation	CoCoDet	2024	−.24 [−1.06, +.58]	−1.49 [−2.85, −.14]	0.044
Organization	EditLens	2024	+23.11 [+14.86, +31.96]	+45.25 [+31.50, +60.43]	< .001
Organization	CoCoDet	2024	+2.65 [−2.33, +7.87]	−.06 [−7.54, +8.03]	0.444
Updating	EditLens	2024	−3.28 [−4.72, −1.85]	−1.02 [−3.03, +.99]	0.002
Updating	CoCoDet	2024	−2.22 [−3.00, −1.45]	−.80 [−2.21, +.61]	0.044
Rating	EditLens	2025	+.132 [+.087, +.176]	+.041 [−.007, +.088]	< .001
Rating	CoCoDet	2025	+.201 [+.175, +.227]	+.177 [+.130, +.223]	0.257
Participation	EditLens	2025	−1.96 [−3.26, −.67]	−3.26 [−4.68, −1.85]	< .001
Participation	CoCoDet	2025	+.40 [−.45, +1.25]	−1.91 [−3.36, −.46]	< .001
Organization	EditLens	2025	+29.52 [+22.20, +37.29]	+41.59 [+32.83, +50.92]	< .001
Organization	CoCoDet	2025	+6.86 [+2.90, +10.98]	+16.33 [+9.10, +24.04]	0.002
Updating	EditLens	2025	−4.94 [−6.34, −3.55]	−2.44 [−3.94, −.93]	< .001
Updating	CoCoDet	2025	−5.00 [−5.93, −4.07]	−5.44 [−7.00, −3.89]	0.553
Rating	EditLens	2026	+.361 [+.315, +.407]	+.345 [+.298, +.391]	< .001
Rating	CoCoDet	2026	+.386 [+.360, +.413]	+.463 [+.419, +.507]	< .001

Table 16 continued

Outcome	Method	Year	Linear	Quadratic	Shape q
Participation	EditLens	2026	−1.84 [−2.97, −.71]	−2.05 [−3.19, −.91]	0.001
Participation	CoCoDet	2026	+1.10 [+.41, +1.80]	+.34 [−.79, +1.47]	0.122
Organization	EditLens	2026	+55.24 [+48.62, +62.15]	+62.03 [+54.76, +69.64]	< .001
Organization	CoCoDet	2026	+23.43 [+20.39, +26.56]	+42.31 [+36.35, +48.53]	< .001
Updating	EditLens	2026†	−2.61 [−3.41, −1.81]	−2.64 [−3.45, −1.84]	0.553
Updating	CoCoDet	2026†	−1.81 [−2.39, −1.24]	−2.59 [−3.37, −1.80]	0.013

Table 17: P10-to-P90 contrasts and functional-form diagnostics (response side): empirical P10-to-P90 contrasts from linear-score and quadratic-score models on identical primary samples. Organization: expected-count percent change; updating: pp. The linear column gives primary-model predictions, not central-80% refits; sample sizes appear in Table 15. The final column is the quadratic-term Wald test, BH-adjusted over all 36 shape tests. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Method	Year	Linear	Quadratic	Shape q
Organization	EditLens	2024	+38.87 [+23.07, +56.70]	+64.78 [+41.76, +91.53]	$< .001$
Organization	CoCoDet	2024	+19.89 [+6.86, +34.51]	+20.09 [+2.36, +40.87]	0.974
Updating	EditLens	2024	−2.91 [−5.41, −.41]	−2.14 [−5.10, +.82]	0.340
Updating	CoCoDet	2024	−.14 [−2.71, +2.44]	−.02 [−3.30, +3.26]	0.964
Organization	EditLens	2025	+29.14 [+16.49, +43.16]	+57.45 [+38.95, +78.41]	$< .001$
Organization	CoCoDet	2025	+38.15 [+25.28, +52.34]	+42.62 [+28.05, +58.84]	0.044
Updating	EditLens	2025	−6.56 [−8.85, −4.26]	−5.26 [−7.87, −2.66]	0.046
Updating	CoCoDet	2025	−2.80 [−5.12, −.48]	−3.32 [−5.73, −.91]	0.146
Organization	EditLens	2026	+20.03 [+11.55, +29.15]	+35.07 [+24.18, +46.92]	$< .001$
Organization	CoCoDet	2026	+33.30 [+24.61, +42.60]	+37.35 [+27.81, +47.60]	$< .001$
Updating	EditLens	2026†	−2.83 [−3.84, −1.83]	−2.74 [−3.79, −1.70]	0.375
Updating	CoCoDet	2026†	+.40 [−.59, +1.39]	+.24 [−.76, +1.25]	0.153

K.2 Tail-exclusion refits

We remove scores above P95 or outside the inclusive P10–P90 range of each primary outcome sample, retain boundary ties, and refit the same model family and controls. Score and length SDs are recomputed after exclusion; methods can retain different records. These restrictions change both the population and the score span, so differences in standardized coefficients do not quantify the fraction of an effect attributable to tails.

Table 18: Tail exclusions (review side). Rating: points/SD; organization: count change (%/SD); binary outcomes: pp from $z = 0$ to $z = 1$. Each retained sample is restandardized; therefore these are not identical exposure contrasts. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Method	Year	Primary	Drop top 5%	Central 80%
Rating	EditLens	2024	+.030 [+.006, +.054]	−.027 [−.052, −.003]	−.022 [−.048, +.004]
Rating	CoCoDet	2024	+.098 [+.076, +.120]	+.005 [−.019, +.030]	−.018 [−.045, +.009]
Participation	EditLens	2024	−.54 [−1.15, +.08]	−.32 [−.94, +.31]	+.05 [−.62, +.73]
Participation	CoCoDet	2024	−.16 [−.73, +.40]	−.77 [−1.40, −.15]	−.66 [−1.34, +.03]
Organization	EditLens	2024	+9.80 [+6.43, +13.29]	+10.55 [+7.01, +14.22]	+9.47 [+5.66, +13.43]
Organization	CoCoDet	2024	+1.82 [−1.61, +5.37]	−2.84 [−6.12, +.57]	−3.74 [−7.29, −.05]
Updating	EditLens	2024	−1.49 [−2.14, −.84]	−.29 [−.98, +.40]	−.37 [−1.13, +.38]
Updating	CoCoDet	2024	−2.01 [−2.69, −1.33]	+.20 [−.47, +.88]	+.38 [−.38, +1.14]
Rating	EditLens	2025	+.052 [+.034, +.069]	−.003 [−.020, +.015]	−.021 [−.040, −.002]
Rating	CoCoDet	2025	+.126 [+.110, +.142]	+.055 [+.038, +.072]	+.037 [+.019, +.056]
Participation	EditLens	2025	−.78 [−1.29, −.26]	−1.31 [−1.85, −.78]	−1.56 [−2.15, −.97]
Participation	CoCoDet	2025	+.23 [−.26, +.73]	−.72 [−1.24, −.21]	−.71 [−1.27, −.15]
Organization	EditLens	2025	+10.71 [+8.20, +13.27]	+11.88 [+9.34, +14.47]	+9.52 [+6.84, +12.27]
Organization	CoCoDet	2025	+4.03 [+1.71, +6.39]	+5.35 [+2.99, +7.76]	+5.12 [+2.56, +7.75]
Updating	EditLens	2025	−1.92 [−2.47, −1.38]	−.57 [−1.14, +.00]	−.08 [−.71, +.54]
Updating	CoCoDet	2025	−3.15 [−3.73, −2.57]	−1.56 [−2.12, −1.00]	−1.25 [−1.86, −.65]
Rating	EditLens	2026	+.128 [+.112, +.145]	+.102 [+.085, +.119]	+.093 [+.075, +.110]
Rating	CoCoDet	2026	+.217 [+.202, +.231]	+.164 [+.148, +.179]	+.124 [+.107, +.141]

Table 18 continued

Outcome	Method	Year	Primary	Drop top 5%	Central 80%
Participation	EditLens	2026	−.66 [−1.07, −.25]	−.83 [−1.26, −.41]	−.43 [−.87, +.02]
Participation	CoCoDet	2026	+.62 [+.23, +1.00]	+.16 [−.24, +.56]	−.18 [−.62, +.25]
Organization	EditLens	2026	+16.99 [+15.18, +18.82]	+18.11 [+16.25, +19.99]	+14.06 [+12.15, +16.00]
Organization	CoCoDet	2026	+12.52 [+10.96, +14.11]	+12.70 [+11.05, +14.38]	+10.66 [+8.89, +12.47]
Updating	EditLens	2026†	−.89 [−1.15, −.63]	−.86 [−1.13, −.59]	−.91 [−1.20, −.62]
Updating	CoCoDet	2026†	−.98 [−1.27, −.68]	−.91 [−1.19, −.63]	−.72 [−1.03, −.41]

Table 19: Tail exclusions (response side). Organization: count change (%/SD); updating: pp from $z = 0$ to $z = 1$. Each retained sample is restandardized; therefore these are not identical exposure contrasts. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Method	Year	Primary	Drop top 5%	Central 80%
Organization	EditLens	2024	+14.37 [+8.86, +20.15]	+18.30 [+12.34, +24.59]	+14.05 [+7.90, +20.55]
Organization	CoCoDet	2024	+7.48 [+2.67, +12.50]	+3.71 [−1.30, +8.98]	+3.21 [−2.03, +8.72]
Updating	EditLens	2024	−1.21 [−2.24, −.19]	−1.00 [−2.06, +.05]	−.48 [−1.63, +.66]
Updating	CoCoDet	2024	−.05 [−1.09, +.98]	−.03 [−1.10, +1.05]	−.29 [−1.41, +.83]
Organization	EditLens	2025	+10.71 [+6.26, +15.34]	+15.51 [+10.77, +20.46]	+11.51 [+6.78, +16.45]
Organization	CoCoDet	2025	+13.59 [+9.29, +18.06]	+12.72 [+8.21, +17.42]	+10.88 [+6.18, +15.79]
Updating	EditLens	2025	−2.59 [−3.48, −1.71]	−1.73 [−2.65, −.81]	−1.29 [−2.29, −.28]
Updating	CoCoDet	2025	−1.11 [−2.02, −.19]	−1.62 [−2.54, −.70]	−1.91 [−2.90, −.92]
Organization	EditLens	2026	+7.14 [+4.22, +10.15]	+11.94 [+8.80, +15.18]	+4.37 [+1.43, +7.39]
Organization	CoCoDet	2026	+11.68 [+8.82, +14.61]	+11.90 [+8.98, +14.90]	+9.41 [+6.43, +12.48]
Updating	EditLens	2026†	−.98 [−1.30, −.67]	−.73 [−1.08, −.38]	−.68 [−1.05, −.30]
Updating	CoCoDet	2026†	+.16 [−.23, +.54]	+.13 [−.27, +.52]	+.13 [−.29, +.55]

Early-cohort rating and updating associations depend on the retained score range. After deleting the highest 5%, the EditLens 2024 rating estimate is negative and the CoCoDet interval includes zero; in the central 80%, both 2024 intervals include zero, while EditLens is negative and CoCoDet positive in 2025. Review-side updating intervals cross zero for both methods in 2024 and for EditLens in 2025. Both measurements retain positive rating and negative review-side updating associations under these restrictions in 2026. Response-side updating also weakens in 2024. High-score observations remain part of the target population; these checks limit extrapolation of the full-sample association across the score range rather than establishing that high scores are invalid observations.

K.3 Score scales and trimming rules

Table 20 compares CoCoDet's primary native log-odds with its original probability score on identical records, using the same controls and each scale's own SD. Positive rating and negative review-side updating associations persist across both scales. Participation and response-side findings depend more on the scale: the 2025 positive participation and negative response-updating intervals exclude zero under only the probability and native scales, respectively. These are alternative parameterizations, not equal changes in actual AI use.

Table 20: CoCoDet scale sensitivity on identical primary samples. Rating: points/SD; organization: expected-count change (%/SD); participation/updating: probability differences from $z = 0$ to $z = 1$ (pp). Native denotes AI-versus-rest log-odds; probability is the original p_{AI}. Each scale uses its own sample SD with unchanged records and controls. Intervals are paper-clustered 95% CIs. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Side	Outcome	Year	Native (primary)	Probability
Review	Rating	2024	$+.098[+.076, +.120]$	$+.066[+.044, +.087]$
Review	Rating	2025	$+.126[+.110, +.142]$	$+.100[+.084, +.116]$
Review	Rating	2026	$+.217[+.202, +.231]$	$+.146[+.131, +.160]$
Review	Participation	2024	$-.16[-.73, +.40]$	$+.35[-.29, +.99]$
Review	Participation	2025	$+.23[-.26, +.73]$	$+.80[+.26, +1.33]$
Review	Participation	2026	$+.62[+.23, +1.00]$	$+.64[+.25, +1.03]$
Review	Organization	2024	$+1.82[-1.61, +5.37]$	$+.70[-2.85, +4.39]$
Review	Organization	2025	$+4.03[+1.71, +6.39]$	$+.58[-1.66, +2.88]$
Review	Organization	2026	$+12.52[+10.96, +14.11]$	$+6.16[+4.72, +7.62]$
Review	Updating	2024	$-2.01[-2.69, -1.33]$	$-2.18[-3.17, -1.18]$
Review	Updating	2025	$-3.15[-3.73, -2.57]$	$-2.15[-2.73, -1.57]$
Review	Updating	2026†	$-.98[-1.27, -.68]$	$-.51[-.81, -.21]$
Response	Organization	2024	$+7.48[+2.67, +12.50]$	$+2.45[-1.16, +6.19]$
Response	Organization	2025	$+13.59[+9.29, +18.06]$	$+6.02[+3.59, +8.51]$
Response	Organization	2026	$+11.68[+8.82, +14.61]$	$+2.55[+.36, +4.79]$
Response	Updating	2024	$-.05[-1.09, +.98]$	$-.24[-1.06, +.59]$
Response	Updating	2025	$-1.11[-2.02, -.19]$	$+.08[-.55, +.71]$
Response	Updating	2026†	$+.16[-.23, +.54]$	$-.18[-.54, +.19]$

Rank-normal scores use within-sample average ranks for ties followed by the inverse normal CDF. CoCoDet ranks are computed from native log-odds, which can distinguish values tied in stored probabilities. The no-trim specification removes only outcome-specific score trimming, retaining token windows, count caps and other eligibility rules. Its estimates equal the primary estimates for outcomes without score trimming, marked by a dash in Tables 21 and 22. Primary CoCoDet participation/table samples retain the original probability-defined trimming IDs.

Table 21: Rank and score-trimming sensitivity (review side). Rating: points/SD; organization: count change (%/SD); binary outcomes: pp from $z = 0$ to $z = 1$. Primary estimates appear in Table 14. A dash means score trimming was absent in the primary model, so removing it changes nothing. Each retained sample is restandardized. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Method	Year	Rank-normal	No score trim
Rating	EditLens	2024	−.004 [−.028, +.021]	—
Rating	CoCoDet	2024	+.054 [+.030, +.078]	—
Participation	EditLens	2024	−.75 [−1.39, −.11]	−.49 [−1.09, +.12]
Participation	CoCoDet	2024	−.58 [−1.18, +.02]	−.08 [−.65, +.50]
Organization	EditLens	2024	+16.22 [+12.28, +20.30]	+9.31 [+5.93, +12.80]
Organization	CoCoDet	2024	−.96 [−4.24, +2.43]	+1.48 [−1.75, +4.81]
Updating	EditLens	2024	−.88 [−1.51, −.24]	—
Updating	CoCoDet	2024	−.64 [−1.27, −.01]	—
Rating	EditLens	2025	+.032 [+.014, +.050]	—
Rating	CoCoDet	2025	+.088 [+.071, +.105]	—
Participation	EditLens	2025	−.87 [−1.40, −.34]	−.80 [−1.32, −.29]
Participation	CoCoDet	2025	−.32 [−.83, +.19]	+.39 [−.10, +.88]
Organization	EditLens	2025	+14.29 [+11.60, +17.06]	+10.57 [+8.09, +13.12]
Organization	CoCoDet	2025	+3.55 [+1.09, +6.07]	+3.98 [+1.62, +6.40]
Updating	EditLens	2025	−1.14 [−1.68, −.59]	—
Updating	CoCoDet	2025	−2.32 [−2.85, −1.79]	—
Rating	EditLens	2026	+.125 [+.108, +.142]	—
Rating	CoCoDet	2026	+.197 [+.181, +.212]	—
Participation	EditLens	2026	−.67 [−1.08, −.26]	−.67 [−1.08, −.26]
Participation	CoCoDet	2026	+.39 [+.01, +.77]	+.62 [+.23, +1.00]
Organization	EditLens	2026	+17.95 [+16.10, +19.84]	+17.08 [+15.28, +18.91]
Organization	CoCoDet	2026	+13.59 [+11.86, +15.34]	+12.49 [+10.93, +14.07]
Updating	EditLens	2026†	−.80 [−1.06, −.55]	—
Updating	CoCoDet	2026†	−.88 [−1.13, −.63]	—

Table 22: Rank and score-trimming sensitivity (response side). Organization: count change (%/SD); updating: pp from $z = 0$ to $z = 1$. Primary estimates appear in Table 14. A dash means score trimming was absent in the primary model, so removing it changes nothing. Each retained sample is restandardized. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Method	Year	Rank-normal	No score trim
Organization	EditLens	2024	+17.20 [+11.21, +23.51]	+13.28 [+7.79, +19.04]
Organization	CoCoDet	2024	+6.73 [+1.52, +12.20]	+7.61 [+2.81, +12.64]
Updating	EditLens	2024	−1.08 [−2.09, −.06]	—
Updating	CoCoDet	2024	+.07 [−.97, +1.11]	—
Organization	EditLens	2025	+13.94 [+9.20, +18.89]	+9.49 [+5.04, +14.12]
Organization	CoCoDet	2025	+14.87 [+10.35, +19.58]	+14.28 [+9.97, +18.76]
Updating	EditLens	2025	−2.37 [−3.25, −1.50]	—
Updating	CoCoDet	2025	−.79 [−1.70, +.13]	—
Organization	EditLens	2026	+7.77 [+4.80, +10.83]	+6.62 [+3.66, +9.66]
Organization	CoCoDet	2026	+12.57 [+9.50, +15.72]	+11.66 [+8.81, +14.58]
Updating	EditLens	2026†	−.95 [−1.26, −.64]	—
Updating	CoCoDet	2026†	+.09 [−.31, +.48]	—

K.4 TEXT-INPUT TRUNCATION

The input-coverage check restricts EditLens to responses that fit completely within its detector input limit (Table 23). Updating associations remain negative in all years; organization remains positive in point estimates, with the 2026 interval including zero. The comparison jointly reflects input coverage, subset composition, and the score SD recomputed within that subset.

Table 23: EditLens response-side input-truncation sensitivity. Primary versus texts that fit the detector input limit; score SD is recomputed in the subset. Organization is count change (%/SD); updating is pp from $z = 0$ to $z = 1$. † Updating in 2026 uses reply-language labels; 2024–2025 use rating snapshots.

Outcome	Year	Primary N	Subset N	Primary	Untruncated
Organization	2024	6,578	3,983	+14.37 [+8.86, +20.15]	+15.56 [+8.30, +23.32]
Updating	2024	6,361	3,862	−1.21 [−2.24, −.19]	−2.03 [−3.26, −.79]
Organization	2025	9,198	5,330	+10.71 [+6.26, +15.34]	+6.73 [+1.03, +12.76]
Updating	2025	9,206	5,348	−2.59 [−3.48, −1.71]	−3.32 [−4.44, −2.20]
Organization	2026	16,674	9,939	+7.14 [+4.22, +10.15]	+3.19 [−.49, +7.01]
Updating	2026†	16,718	10,004	−.98 [−1.30, −.67]	−1.19 [−1.53, −.85]