The Deadline Effect: Identifying and Correcting Temporal Bias in Human Evaluation

Shiyan Liu
Imperial College London
shiyan.liu26@ic.ac.uk

Abstract

Submission timing offers an inexpensive signal of variation in human evaluation. We analyse 130,023 reviews of 32,955 papers from ICLR 2024, 2025, and 2026. Later reviews have higher within-paper rating contrasts in all 15 reported specifications, although not all estimates are significant. Official-deadline and confidence-adjusted analyses support the association without identifying a causal effect. Early reviews also have lower squared error against recorded committee decisions, with annual gaps of approximately 0.005 to 0.006. Optimal Timing-Calibrated Aggregation (OTCA) fits a linear timing weight to a decision-alignment objective. Combined with confidence, it improves in-sample borderline decision agreement over uniform aggregation by 4.11 to 11.25 percentage points; separate cross-year tests support transfer. Benefits depend on the metric and target, including the threshold used for metareview-derived labels. A complementary Adversarial Temporal Debiasing (ATD) probe reduces symmetric timing-discriminator AUC by 0.243 while retaining acceptance prediction in a review-level split. Text features show modest engagement differences, but blinded, length-controlled assessments do not establish a substantive quality advantage. These results motivate further evaluation of temporal metadata while distinguishing prediction of existing decisions from improvement in evaluation quality. Code and numerical data: https://github.com/shiyan-liu/deadline-effect.

1 Introduction

Human judgments shape scientific peer review, crowdsourced annotation, and reinforcement learning from human feedback (RLHF) [18, 19, 15]. Alongside ratings and confidence, submission timestamps offer readily available information whose meaning and value for aggregation remain unclear.

Timing may reflect time pressure, workload, expertise, assignment timing, or reviewing style. The latest review of a paper need not be near the official deadline. These ambiguities require separating measurement validity, predictive utility, and interpretation.

We study three ICLR cycles through three connected questions. First, does relative submission timing consistently predict within-paper scoring, and does the pattern survive alternative timing measures and reviewer covariates (Section 5)? Second, can this information improve aggregation, transfer across cycles, and retain value under different evaluation targets (Section 6)? Third, what do review text and direct quality assessments reveal about the interpretation of the association (Section 7)?

To examine predictive utility, we introduce Optimal Timing-Calibrated Aggregation (OTCA), which fits a linear timing weight to a paper-level decision-alignment objective. Joint multiplies this weight by confidence, testing whether timing adds information beyond the reviewer's stated certainty. Separately, Adversarial Temporal Debiasing (ATD) probes whether timing predictability can be attenuated while preserving acceptance prediction. Our empirical contribution links these predictive

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

tests to measurement checks and blinded text assessments, establishing where temporal metadata is informative and where a quality interpretation remains unsupported.

2 Related Work

Human evaluation and peer review. Prior work documents prestige effects in single-blind reviewing [21], variation in perceived review quality [6], and broader reliability challenges [18, 17]. Annotator-error models [2] and crowdsourcing methods [8, 16, 22] motivate nonuniform aggregation. Leniency also arises in educational assessment and online ratings [3, 12]. Our focus is the information available from submission timing within a review cycle.

Review text and AI assistance. PeerRead and subsequent NLP research provide resources and methods for analysing scientific reviews [9, 10, 23]. AI assistance may contribute to timing patterns [11]: Liang et al. [13] report corpus-level evidence linking AI-modified text to proximity to review deadlines. LLM judges also exhibit position and verbosity biases [25], which motivate the order-swapped and length-matched checks reported below.

Aggregation and representation learning. Probabilistic aggregation can model reviewer reliability [7]. Timing weighting offers a small, interpretable alternative when stable reviewer identifiers or historical reliability estimates are unavailable. Domain-adversarial learning [4] and adversarial fairness methods [24, 14] motivate our separate ATD experiment. Applications to RLHF remain prospective despite its dependence on annotation quality [20, 5, 1], but we do not evaluate preference-annotation data.

3 Data and Measurements

Sample. We analyse OpenReview records for ICLR 2024, 2025, and 2026. We retain papers with at least two valid rating timestamps and an observed timestamp spread of at least one hour. The sample contains 32,955 papers and 130,023 reviews (Table 1). Acceptance is coded from the recorded decision string. In 2024 and 2025, records without an explicit acceptance label, including empty decision strings, are coded as non-acceptance; the 2026 sample contains explicit accept or reject decisions. This convention limits interpretation of the target as adjudicated rejection. Timestamps record review creation, while ratings and text can include later revisions, so the analysis relates initial submission time to the review record available at collection.

Table 1: Primary ICLR samples. The observed timestamp window spans the earliest to latest retained review of each paper.

	Papers	Reviews	Rev./paper	Accept%	Scale	Med. window (h)
2024	7,261	28,027	3.86	31.1%	$\{1, 3, 5, 6, 8, 10\}$	174.6
2025	11,520	46,748	4.06	32.2%	$\{1, 3, 5, 6, 8, 10\}$	172.9
2026	14,174	55,248	3.90	37.8%	$\{0, 2, 4, 6, 8, 10\}$	157.8
Total	32,955	130,023	*n/a*	*n/a*	*n/a*	*n/a*

To assess generality, we also use a selected NeurIPS 2025 sample of 5,541 papers and 22,373 reviews, with 95.4% recorded acceptance. For target sensitivity, we use 2,378 area-chair metareviews from ICLR 2024 and 2025; 1,466 belong to the fixed borderline subsets.

Data attribution and terms. We credit the ICLR and NeurIPS organisers, OpenReview, and the contributors of the review records. OpenReview's Terms of Use license review and metareview comments under CC BY 4.0 and dedicate associated metadata under CC0; readability remains governed by access controls.[1] These terms concern the records analysed here, rather than the separate licences attached to submitted article PDFs. Appendix E documents the asset terms and research-artifact availability.

[1] `https://openreview.net/legal/terms`, updated 5 November 2026; see "Comment and Configuration Record License" and "Metadata Dedication."

Relative submission position. For paper i, let t_{ij} be the creation timestamp of review j. Define

$$\mathrm{rp}_{ij} = \frac{\max_k t_{ik} - t_{ij}}{\max_k t_{ik} - \min_k t_{ik}} \in [0, 1]. \tag{1}$$

The latest review has $\mathrm{rp} = 0$ and the earliest has $\mathrm{rp} = 1$. Within each year, pooled review quantiles define Early ($\mathrm{rp} \geq q_{66}$), Mid ($q_{33} \leq \mathrm{rp} < q_{66}$), and Late ($\mathrm{rp} < q_{33}$). Endpoint ties mean that bucket sizes need not be equal. The observed spread differs from the allotted review period.

For measurement validation, we also partition the gap to the official reviewer deadline into year-specific tertiles (date anchors in Appendix A.3). Reviewer adjustment uses self-reported confidence. Stable cross-paper reviewer identifiers, workload, seniority, assignment time, and emergency status are unavailable.

Ratings and decision agreement. We normalise ratings as $\tilde{r} = (r - r_{\min})/(r_{\max} - r_{\min})$, using the year's scale bounds. The Brier-style score is $N^{-1} \sum (\tilde{r} - y)^2$, where y is recorded acceptance status; lower values indicate closer numerical agreement. Normalised ratings are score-derived probability proxies, not calibrated probability forecasts. This metric shares the dependence on committee decisions of our accuracy metric and does not resolve circularity.

4 Methods

4.1 Within-Paper Rating Contrasts

We compare each review to the other reviews of its paper using a leave-one-out contrast,

$$d_{ij} = r_{ij} - \bar{r}_{i,-j} = \frac{n_i}{n_i - 1}(r_{ij} - \bar{r}_i), \qquad d_{ij} = \mu + \beta_E E_{ij} + \beta_L L_{ij} + \varepsilon_{ij}. \tag{2}$$

Here E and L indicate Early and Late, with Mid as reference. Paper-specific subtraction removes additive rating offsets. Regressing these contrasts on bucket indicators differs from conventional fixed effects, which demeans both outcome and regressors; the coefficients are descriptive within-paper contrasts.

Reviewer differences remain possible confounders. Confidence-adjusted models add centred confidence and timing interactions. Inference uses paper-clustered standard errors and paper bootstrap ($B = 500$). Alternative hour buckets classify gaps to the latest review as Late for $g_{ij} < h/2$, Early for $g_{ij} > 2h$, and Mid otherwise. These specifications and the soundness regression use mean-centred outcomes; the pooled-bucket model uses Eq. 2.

4.2 Optimal Timing-Calibrated Aggregation

For ratings r_{ij} and positive weights, define

$$w_{ij}(a) = 1 + a\,\mathrm{rp}_{ij}, \qquad \hat{s}_i(a) = \frac{\sum_j w_{ij}(a) r_{ij}}{\sum_j w_{ij}(a)}, \qquad a > -1. \tag{3}$$

A positive a gives greater weight to earlier reviews; $a = 0$ recovers uniform averaging. The feasible bound follows because each paper contains a review with $\mathrm{rp} = 1$. The reported OTCA fit minimises

$$a^\star = \arg\min_{a > -1} \frac{1}{N} \sum_i (\hat{s}_i(a) - y_i)^2. \tag{4}$$

The objective uses *raw ratings* against binary decisions, a scale-dependent criterion distinct from normalised Brier evaluation. Normalising the fitting target changes the optimum (Appendix B.1).

We fit a by feasible grid search and interval refinement, with linear-time objective evaluation. "Optimal" denotes the numerical fit within the searched family, without a global uniqueness guarantee. The 2026 fit is $a^\star \approx 1.204$; search details appear in Appendix B.1.

The five rules in Table 2 aggregate the same ratings and differ only in their weights.

Figure 1 contrasts the continuous and bucket-based weighting rules. The heuristic uses observed bucket-mean contrasts $\hat{\delta}_E$ and $\hat{\delta}_L$, with $\hat{\delta}_M = 0$. Its approximate 2026 weights are

Table 2: Aggregation rules. All methods consume the same ratings. Additional inputs specify the information needed beyond ratings; fit labels are required only when estimating the timing coefficient.

Rule	Weight before per-paper normalisation	Additional inputs
Uniform	1	None
Confidence	c_{ij}	Confidence
Heuristic timing	$1 - \hat{\delta}_{b(ij)}$	Timestamps
OTCA	$1 + a^\star \mathrm{rp}_{ij}$	Timestamps, fit labels
JOINT	$(1 + a^\star \mathrm{rp}_{ij}) c_{ij}$	Timestamps, confidence, fit labels

$(1.096, 1.000, 0.904)$ for Early, Mid, and Late. Joint tests whether timing contributes predictive information beyond confidence; a gain does not establish their statistical independence.

Fitting requires historical decisions or a labelled calibration sample. Once a slope is fixed, inference requires ratings and relative timing, plus confidence for Joint.

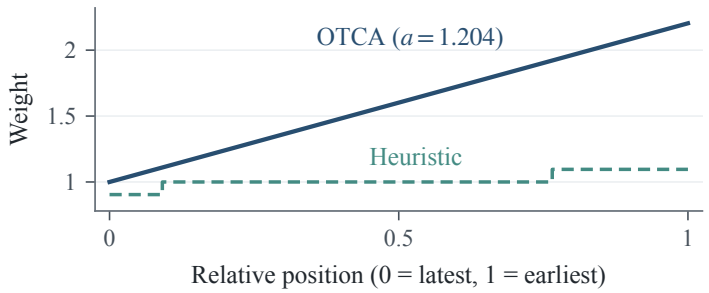


Figure 1: Timing weights in 2026. OTCA varies continuously with relative position; the heuristic is constant within each pooled timing bucket. Weights are shown before paper-level normalisation.

4.3 Adversarial Temporal Debiasing

For a complementary representation-level analysis, an encoder maps four inputs (normalised rating, confidence, soundness, and the gap to the paper's latest review) to a 32-dimensional representation. A predictor estimates recorded acceptance, while a discriminator predicts a timing label. Let $\mathcal{L}_y$ and $\mathcal{L}_\ell$ be binary cross-entropy losses. Their encoder objective is

$$\min_{\theta,\phi} \; \mathcal{L}_y(g_\phi(f_\theta(x)), y) - \lambda \mathcal{L}_\ell(d_\psi(f_\theta(x)), \ell), \qquad \min_{\psi} \; \mathcal{L}_\ell. \tag{5}$$

A gradient reversal layer implements the adversarial sign. The encoder has a ReLU hidden layer and a linear output, avoiding inactive output units.

ATD uses a random 80/20 *review-level* split, 50 epochs, and $\lambda = 1$. Its timing label uses a within-paper lower quantile, distinct from the pooled regression buckets. Paper overlap between partitions and full-snapshot preprocessing limit out-of-paper generalisation; Appendix C gives further details.

5 Timing Associations and Validation

We first establish the within-paper pattern, validate timing measurement and reviewer adjustment, and examine variation across paper groups and venues.

5.1 Within-Paper Timing Associations

Appendix Table 8 reports five specifications in each year; the primary coefficients also appear in Table 3. All 15 late coefficients are positive, but several are not statistically significant, particularly in 2024. The 24-, 48-, and 72-hour specifications use gaps to each paper's *latest observed review*; they are not official-deadline cutoffs. Soundness replaces overall rating as the dependent variable. The

normalised pooled-bucket late coefficient increases from 0.0014 to 0.0096 across years. The 2026 soundness contrast is positive but nonsignificant under clustered inference ($p = 0.067$). With only three cycles, the change in effect size cannot be attributed to time pressure.

Figure 2 shows the pooled-bucket means and paper-bootstrap intervals for all three years.

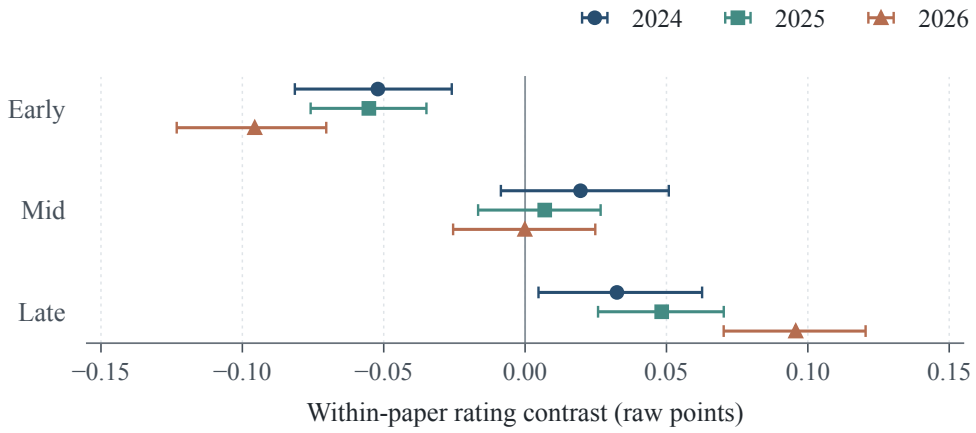


Figure 2: Submission timing and within-paper ratings across three ICLR cycles. Points show mean leave-one-out rating contrasts within each timing bucket; horizontal intervals are 95% paper-bootstrap intervals ($B = 500$). The zero line represents agreement with the other reviews of the same paper. These are bucket means, rather than regression coefficients relative to Mid.

The scoring pattern also raises a question about agreement with the recorded decisions. Applying the metric in Section 3 to individual reviews, Early reviews have lower loss than Late reviews in every year, with gaps of 0.0054, 0.0053, and 0.0059 (Figure 3). This connects relative scoring differences to a decision-based outcome, while leaving reviewer competence unmeasured.

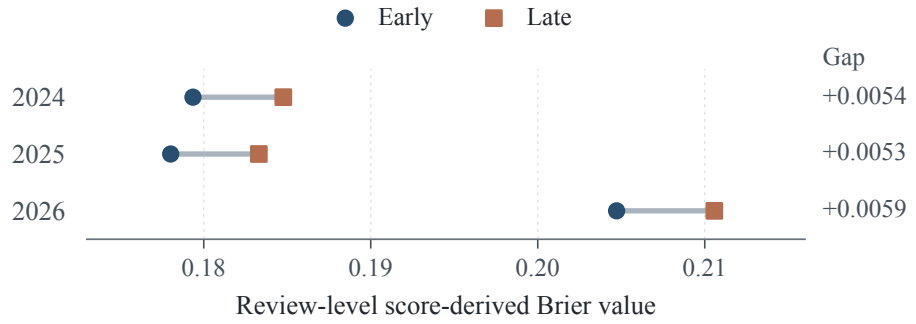


Figure 3: Review-level decision agreement by submission timing. Each pair joins Early and Late score-derived Brier values in one ICLR cycle; lower values indicate closer agreement with recorded decisions. The annotated gap is Late minus Early. This comparison concerns individual reviews, whereas the aggregation evaluation in Section 6.1 concerns paper-level scores.

5.2 Deadline Measurement and Reviewer Covariates

Official review deadlines. Relative position measures ordering within an observed paper-specific window. To test whether the association also appears when timing is anchored to an external deadline, we repeat the analysis using the official-date measure defined in Section 3. Relative position correlates with this gap at Pearson $r = 0.59$ to 0.66 and Spearman $r = 0.70$ to 0.74. Official-deadline tertiles show early strictness in all three years. Late leniency is significant in 2025 and 2026, while the 2024 estimate is negative and nonsignificant. Pooling all 130,023 reviews with year indicators and

paper-clustered standard errors gives $\beta_E = -0.0978$ ($p < 0.001$) and $\beta_L = +0.0333$ ($p = 0.0051$). Thus the pooled association is not specific to within-paper timing, but the late effect does not replicate significantly in every year under both measures.

Confidence and residual confounding. An association under both timing measures could still reflect differences between reviewers. We therefore add confidence as an observed reviewer covariate and test its interactions with timing. Adjustment attenuates the late coefficients in every year, with the strongest remaining contrast in 2026 (Table 3). No confidence by timing interaction is significant ($p = 0.41$ to 0.97). These checks show partial overlap with confidence, not removal of all reviewer-level confounding. The unavailable reviewer attributes described in Section 3 remain potential sources of confounding.

Table 3: Measurement validity and confidence adjustment. Official-deadline coefficients use Mid as reference; relative-timing late coefficients are shown before and after adding confidence. All coefficients are raw rating points. The last row pools years with year indicators.

	Official-deadline timing		Relative-timing β_L	
Year	β_E	β_L	Base	+ confidence
2024	-0.092	-0.019	$+0.013$	$+0.006$
2025	-0.062	$+0.040$	$+0.041$	$+0.033$
2026	-0.131	$+0.054$	$+0.096$	$+0.083$
Pooled	-0.098	$+0.033$	Not pooled	

5.3 Heterogeneity and Cross-Venue Evidence

We next ask where the association is most pronounced and whether any part of it extends beyond ICLR. In 2026, the largest late contrast occurs in the score group with the most evenly split recorded decisions ($\beta_L = 0.125$, clustered $p < 0.001$; 51.4% accepted). The high-score group has $\beta_L = 0.038$ ($p = 0.292$; 88.6% accepted). Figure 4 shows the full score-group pattern with clustered uncertainty; Appendix A.5 gives the exact cutoffs and group sizes. These coefficients use mean-centred outcomes, and tied scores produce unequal groups. These within-group estimates do not establish differences between groups or exclude reviewer self-selection.

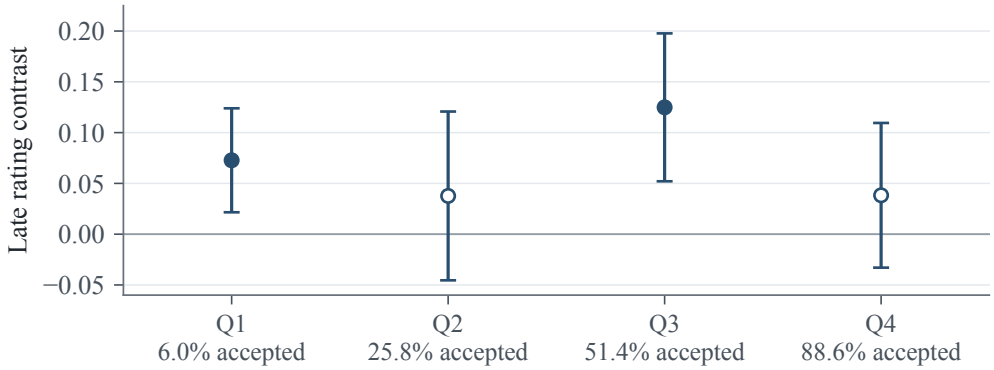


Figure 4: Late versus Mid mean-centred rating contrasts by 2026 paper-score group. Intervals are 95% normal intervals using paper-clustered standard errors. Filled markers indicate $p < 0.05$. Acceptance rates beneath the group labels describe the same papers.

The selected NeurIPS sample provides a limited test of external scope. Relative timing yields early strictness on the normalised scale ($\beta_E = -0.0071$, $p = 0.012$), but no late effect ($\beta_L = -0.0004$, $p = 0.878$). Relative position correlates with the official-deadline gap at Pearson $r = 0.60$ and Spearman $r = 0.72$. With that deadline, the late coefficient is $+0.0078$ ($p = 0.0023$) and the early

coefficient is nonsignificant. Later reviews in this sample also contain 17.7 fewer words and report 0.078 lower confidence on average, consistent with the engagement analysis in Section 7. The mixed scoring results and acceptance-heavy sampling limit the generality claim: parts of the timing pattern extend to a second venue, but venue-wide late leniency is not established.

6 Predictive Utility of Temporal Information

Having established a timing association, we test its predictive utility through aggregation, cross-year transfer, target sensitivity, and representation learning.

6.1 Aggregation Performance

We fix a borderline subset in each year: papers whose uniform mean lies within one raw rating point of the median *individual review rating*. The year-specific median (5, 5, and 4) also classifies aggregate scores. This is an analytical benchmark rather than the committee's decision rule. All methods use the same papers ($n = 4{,}433$, $6{,}941$, and $9{,}511$), with continuity-corrected McNemar tests. Scores equal to the threshold are classified as acceptance; differences within 10^{-12} raw rating points are treated as numerical ties. This rule is shared by all aggregation and target-sensitivity comparisons.

Joint improves over uniform by 4.38, 4.11, and 11.25 percentage points in 2024, 2025, and 2026, respectively (Figure 5; exact values in Appendix Table 13). It has the highest agreement in 2024 and 2026; confidence alone is marginally higher in 2025 (0.6361 versus 0.6355). Joint exceeds confidence significantly only in 2026 ($\chi^2 = 95.95$, $p < 0.001$), with nonsignificant differences in 2024 and 2025 ($p = 0.73$ and 0.88). Timing and confidence therefore offer complementary information in the strongest-signal cycle, but their product does not dominate in every year.

Uniform agreement is 0.5953, 0.5944, and 0.4511 across years. OTCA exceeds the heuristic in 2024 and 2026 ($\chi^2 = 12.01$ and 46.26, both $p < 0.001$). Its agreement is lower in 2025, although the difference is nonsignificant ($p = 0.086$). The full-sample comparison below evaluates whether these threshold-based gains extend to normalised squared error.

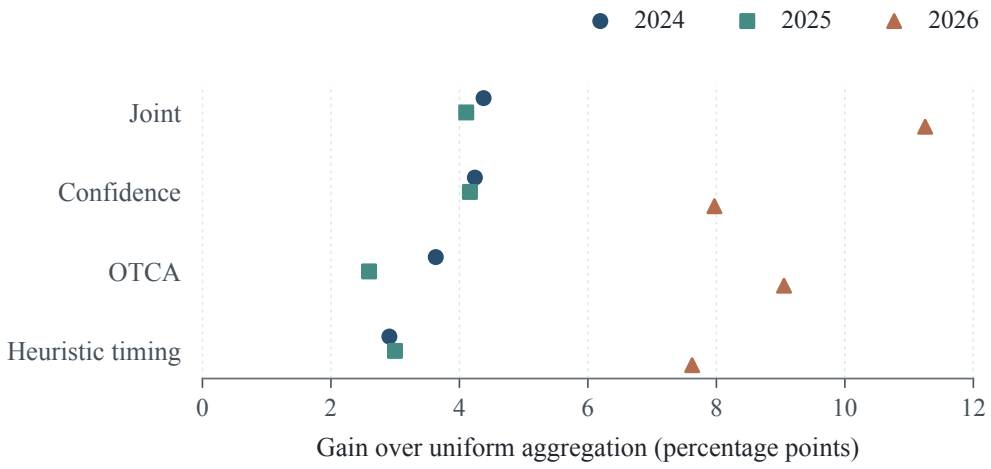


Figure 5: In-sample gains in borderline decision agreement over uniform aggregation. Each point uses the same year's fixed borderline set. Marker shape and colour identify the year; the horizontal axis is in percentage points. Joint performs best in 2024 and 2026, while confidence is slightly higher in 2025.

Full-sample squared error. We also evaluate all retained papers with the score-derived Brier metric (Appendix Table 14). OTCA's loss changes are small and nonsignificant in every year, whereas confidence and Joint improve on uniform in all three years. Confidence has slightly lower loss than Joint in 2026 (0.18347 versus 0.18352), despite Joint's higher borderline agreement. The metrics measure different properties: a threshold crossing can change agreement without materially changing

the full-sample squared error. The objective distinction established in Section 4.2 also means that OTCA need not minimise this evaluation loss. Thus the evidence supports benchmark-specific predictive gains, rather than a general improvement across all notions of evaluation reliability.

6.2 Cross-Year Generalisation

To test transfer beyond the fitting cycle, we fit the OTCA slope on 2024 and evaluate on 2025, and reverse the direction (Table 4). Held-out gains over uniform are 2.98 points on 2024 and 3.37 on 2025 (both McNemar $p < 0.001$). The transferred slope loses 0.65 points on 2024 but gains 0.78 on 2025 relative to within-year fitting, consistent with the distinction between fitted loss and thresholded accuracy. These comparisons support transfer between two cycles; they do not provide a three-year held-out evaluation of Joint.

Table 4: Cross-year OTCA evaluation. The set of borderline papers and classification rule remain fixed within each evaluation year. Every gain over uniform is significant at McNemar $p < 0.001$.

Evaluation	Slope fitted on	Uniform	OTCA	Gain (pp)
2024	2024 (in-sample)	0.5953	0.6316	+3.63
2024	2025 (held-out year)	0.5953	0.6251	+2.98
2025	2025 (in-sample)	0.5944	0.6204	+2.59
2025	2024 (held-out year)	0.5944	0.6282	+3.37

6.3 Sensitivity to the Evaluation Target

Transfer addresses estimation on unseen cycles, but leaves the meaning of the target unchanged. Final decisions are downstream of the reviews, so higher agreement with them does not independently establish better decisions. We examine an alternative signal: LLM-coded enthusiasm in 2,378 area-chair metareviews, with instructions to disregard explicit verdict words. Among 1,466 borderline papers, Joint has agreement 0.6044 at the a priori enthusiasm cutoff of 3, versus 0.6064 for uniform ($p = 0.86$). At cutoff 4, selected by decision-label calibration or aggregate acceptance-rate matching, Joint reaches 0.5430 versus 0.4905 for uniform ($p = 2.2 \times 10^{-11}$). Both rules select the same threshold (Table 5). This sensitivity limits the check, whose target remains part of the decision process (Appendix B.4).

Table 5: Pooled agreement on 1,466 borderline papers under alternative enthusiasm cutoffs. Both calibration rules select cutoff 4 and hence yield the same comparison.

Target and selection rule	Uniform	OTCA	Joint	Joint versus uniform p
≥ 3: a priori midpoint	0.6064	0.6064	0.6044	0.86
≥ 4: calibrated or rate-matched	0.4905	0.5368	0.5430	2.2×10^{-11}

6.4 Representation-Level Timing Suppression

The preceding tests evaluate timing as an aggregation input. We now examine the representation-level question introduced in Section 4.3, using acceptance prediction and residual timing predictability as separate outcomes. For ATD, symmetric discriminator area under the receiver operating characteristic curve (AUC), defined as $0.5 + |\mathrm{AUC} - 0.5|$, decreases from 0.999 to 0.756. This is an absolute reduction of 0.243 AUC units. Acceptance-prediction AUC is 0.716 for the normalised-rating baseline and 0.720 for ATD. Appendix Table 15 reports the timing and acceptance scores. These are review-level results, with no repeated-seed uncertainty estimate, and are not directly comparable to paper-level aggregation accuracy. Residual timing AUC of 0.756 shows that timing remains predictable and does not establish invariance. The result suggests a trade-off worth further testing, but cannot rule out confounding or demonstrate a causal separation of timing and paper quality.

7 Engagement and Alternative Explanations

The predictive results leave open why Early and Late reviews differ. Review text offers evidence about engagement and substance, allowing us to examine whether these differences support a quality interpretation or an alternative explanation.

7.1 Observable Engagement

A reduced-engagement interpretation predicts observable changes in writing behaviour. All five markers in Table 6 are lower for Late reviews in every year. Word-count differences range from 10.4 to 32.8 words, with small standardised effects ($d = -0.132$ to -0.044). These consistent differences motivate a quality assessment, but cannot establish reduced effort: concise reviews can be substantive and longer reviews unhelpful.

Table 6: Engagement markers on the same filtered review samples as the timing analysis. Cells report Late minus Early differences, with Cohen's d in parentheses. Effect sizes are small throughout.

Feature	2024	2025	2026
Total words	-10.4 (-0.044)	-32.8 (-0.132)	-25.7 (-0.116)
Weakness words	-9.6 (-0.057)	-25.6 (-0.149)	-18.6 (-0.130)
Weakness/strength ratio	-0.65 (-0.056)	-0.78 (-0.060)	-0.58 (-0.052)
Numerical density	-0.002 (-0.079)	-0.003 (-0.155)	-0.001 (-0.070)
Question count	-0.11 (-0.037)	-0.18 (-0.058)	-0.22 (-0.071)

7.2 Blinded Assessments of Substantive Quality

Writing behaviour alone cannot determine whether a review is informative or well reasoned. We therefore compare Early and Late reviews using GPT-4.1 and DeepSeek-V4-Pro, with both judges blind to ratings, timing, and confidence. Each pair is presented in both orders and retained as decisive only when the preference is consistent. To distinguish substance from verbosity, we compare the observed win rate with a length-only prediction and repeat the comparison on pairs differing by less than 10% in length.

Early-review win rates among decisive pairs range from 51.7% to 56.1%, close to the length-only predictions (Table 7). With length matching, the 2024 rates are near chance and the early advantage reverses in 2025 and 2026 for both judges. The matched subsets contain only 41 to 95 pairs, so these results establish neither an early nor a late substantive-quality advantage. Together with the engagement measures, they support a narrower interpretation: submission timing predicts writing behaviour more clearly than it predicts quality independently of verbosity.

Table 7: Blinded, order-swapped pairwise review assessments. GPT denotes GPT-4.1 and DS denotes DeepSeek-V4-Pro. Early-win percentages are conditional on decisive pairs. Length-matched pairs differ by less than 10% in length.

Year	Judge	Decisive n	Early win (%)	Length-only (%)	Matched win (%)
2024	GPT	789	54.6	54.9	52.4 ($n = 82$)
2024	DS	557	54.2	53.9	52.2 ($n = 46$)
2025	GPT	814	54.4	56.9	39.3 ($n = 84$)
2025	DS	512	56.1	58.2	41.5 ($n = 41$)
2026	GPT	834	51.7	53.6	43.2 ($n = 95$)
2026	DS	445	52.1	56.0	46.5 ($n = 43$)

7.3 Content-Based Aggregation

The engagement features also raise a practical question: could content provide a more direct aggregation signal than timing? A composite of the five standardised features is negatively associated with ratings, so weighting it may partly re-encode a reviewer's scoring stance. Its raw weighting improves borderline agreement, but removing the component explained by rating changes the gains

to -4.06, -3.70, and $+7.05$ points across years (Appendix D.3). This sign instability provides no consistent basis for replacing timing with a simple content-derived reliability weight. The comparison is exploratory because its fitting and clipping rules differ from constrained OTCA.

7.4 Differential AI Assistance

A distinct explanation is that late reviewers use AI assistance more often, potentially affecting both scores and text [13]. We examine lexical markers as a limited check. The Late minus Early standardised marker-density differences are -0.06, $+0.01$, and -0.01 across the three years; differences in the share containing any marker are -1.73, $+0.03$, and -0.71 percentage points. The effects are small and vary in sign across years. They provide no clear lexical explanation of the timing pattern, but their weak authorship validity means that differential AI use remains unresolved.

8 Discussion and Limitations

Interpretation and deployment. Relative timing is a useful descriptive signal in these snapshots. It can reflect workload, expertise, assignment timing, reviewing style, or time pressure. Our observational data cannot distinguish these explanations. Committee decisions and metareview prose are endogenous to the review process, so improved agreement should not be described as independently verified decision improvement. The raw-scale OTCA objective is also sensitive to rating conventions, and the main gains depend on a particular borderline definition and threshold. Historical or prospective calibration, paper-grouped validation, and externally assessed review quality are needed before operational adoption.

Fairness and incentives. Timing weights could disadvantage reviewers facing caregiving duties, health constraints, heavy workloads, or inconvenient time zones. Publicly known weights could encourage early placeholder submissions without increasing substantive effort. Timestamp creation may also precede important revisions. A platform should audit subgroup effects, monitor changes in behaviour, allow human override, and evaluate structured review support before assigning operational timing weights. The method is computationally simple, but validation and governance have real costs.

Scope and reproducibility. The primary evidence comes from ICLR, with selected NeurIPS data providing only limited external validation. No RLHF or crowdsourcing experiment is reported. The ATD split permits paper overlap and does not measure the combined effect of ATD and OTCA. Asset terms and access restrictions are documented in Appendix E. The public project repository is linked in the abstract; artifact availability is documented in Appendix E. Original execution-resource reporting remains incomplete.

9 Conclusion

Across 130,023 ICLR reviews, relative submission timing is associated with within-paper scoring differences. Official-deadline and confidence checks support this association while leaving reviewer-level confounding unresolved. Timing-aware aggregation improves fitted borderline decision agreement and transfers between two cycles, although gains depend on the evaluation metric and target. Review text supplies evidence of modest engagement differences, but blinded assessments do not establish a substantive quality advantage independent of verbosity. These findings motivate prospective, fair, and independently validated uses of temporal metadata, with a clear separation between predicting existing decisions and improving evaluation quality.

Acknowledgments and Disclosure of Funding

The author thanks the reviewers and area chair for their constructive feedback and the conference organisers for maintaining the evaluation records used in this study. The author received no third-party funding or in-kind support for this work during the 36 months preceding submission and declares no competing financial interests relevant to this work during that period. Language-model assistance was used for editing and for the supplementary text-assessment analyses described in the appendices.

References

[1] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[2] Alexander Philip Dawid and Allan M Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20 to 28, 1979.

[3] Thomas Eckes. *Introduction to many-facet Rasch measurement*. Peter Lang Lausanne, 2023.

[4] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1 to 35, 2016.

[5] Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, page 10835 to 10866. PMLR, 2023.

[6] Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. Peer reviews of peer reviews: A randomized controlled trial and other experiments. *PLOS ONE*, 20(4):e0320444, 2025.

[7] Ilya M Goldin and Kevin D Ashley. Peering inside peer review with Bayesian models. In *International Conference on Artificial Intelligence in Education*, page 90 to 97. Springer, 2011.

[8] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, page 64 to 67, 2010.

[9] 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 the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long papers)*, page 1647 to 1661, 2018.

[10] Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

[11] 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.

[12] Hady W Lauw, Ee-Peng Lim, and Ke Wang. Quality and leniency in online collaborative rating systems. *ACM Transactions on the Web (TWEB)*, 6(1):1 to 27, 2012.

[13] 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. *arXiv preprint arXiv:2403.07183*, 2024. URL `https://arxiv.org/abs/2403.07183`.

[14] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, page 3384 to 3393. PMLR, 2018.

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730 to 27744, 2022.

[16] Vikas C Raykar, Shipeng Yu, Linda H Zhao, Gerardo Hermosillo Valadez, Charles Florin, Luca Bogoni, and Linda Moy. Learning from crowds. *Journal of Machine Learning Research*, 11(4), 2010.

[17] Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in NLP? In *Findings of the association for computational linguistics: EMNLP 2020*, page 1256 to 1262, 2020.

[18] Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76 to 87, 2022.

[19] Rion Snow, Brendan O'Connor, Dan Jurafsky, and Andrew Y Ng. Cheap and fast: But is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, page 254 to 263, 2008.

[20] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008 to 3021, 2020.

[21] Andrew Tomkins, Min Zhang, and William D Heavlin. Reviewer bias in single-versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48):12708 to 12713, 2017.

[22] Jacob Whitehill, Ting-fan Wu, Jacob Bergsma, Javier Movellan, and Paul Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. *Advances in Neural Information Processing Systems*, 22, 2009.

[23] Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171 to 212, 2022.

[24] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, page 335 to 340, 2018.

[25] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL `https://arxiv.org/abs/2306.05685`.

A Inference and Timing Validation

This appendix follows the validation sequence in Section 5: statistical conventions, alternative specifications, external deadline anchors, confidence controls, and score-group heterogeneity.

A.1 Statistical Protocol

Paper-bootstrap intervals in Figure 2 use 500 resamples, retaining all reviews of each sampled paper. The bucket means differ from the Late versus Mid coefficients in Table 8. Regression inference clusters by paper; alternative outcome centring follows Section 4.1. Classification comparisons use continuity-corrected McNemar tests, and paper-level loss comparisons use paired t tests. All p-values are nominal, without multiple-comparison correction. The common numerical-tie rule is specified in Section 6.1.

A.2 Primary and Alternative Timing Specifications

Table 8 complements the primary coefficients in Table 3 by varying the timing partition and rating outcome.

Table 8: Late versus Mid rating contrasts. Each cell gives the coefficient and paper-clustered standard error. Global tertiles use leave-one-out contrasts; hour-cutoff and soundness specifications use mean-centred outcomes. Positive estimates occur in all 15 cells. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

Specification	2024	2025	2026
Global tertiles	$+0.013$ (0.025)	$+0.041^{*}$ (0.019)	$+0.096^{***}$ (0.023)
Gap parameter $h = 24$ h	$+0.014$ (0.023)	$+0.039^{*}$ (0.017)	$+0.052^{**}$ (0.019)
Gap parameter $h = 48$ h	$+0.023$ (0.019)	$+0.032^{*}$ (0.014)	$+0.050^{**}$ (0.016)
Gap parameter $h = 72$ h	$+0.031$ (0.017)	$+0.020$ (0.013)	$+0.062^{***}$ (0.015)
Soundness outcome	$+0.010$ (0.009)	$+0.019^{**}$ (0.007)	$+0.011$ (0.006)

A.3 Official Review Deadlines

The official-date sources are the ICLR 2024 reviewer guide,[2] the ICLR 2025 dates page,[3] and the ICLR 2026 reviewer guide.[4] We use the following UTC anchors: 1 November 2023 at 00:00 UTC (end of the date-only 31 October deadline), 4 November 2024 at 12:00 UTC (end of 3 November Anywhere on Earth), and 2 November 2025 at 00:00 UTC (end of the date-only 1 November deadline). The 2024 and 2026 pages do not specify a time of day in their date-only task listings; their anchors are conventions rather than verified exact cutoffs.

Official-deadline gaps are split using year-specific quantiles. A common shift in the anchor time within a year does not change the rank or tertile assignment. However, reviewer-specific extensions, emergency assignments, and differences between initial creation and final revision can affect construct validity. Reviews with negative gaps are retained; the timestamp spread does not imply that every review met its official deadline.

Table 9: Official-deadline analyses. Coefficients are on each year's raw rating scale, with Mid as reference. The 2024 late estimate reverses sign and is nonsignificant.

Year	Pearson r	Spearman r	β_E	β_L
2024	0.59	0.70	-0.092	-0.019
2025	0.66	0.74	-0.062	$+0.040$
2026	0.64	0.73	-0.131	$+0.054$

[2] `https://iclr.cc/Conferences/2024/ReviewerGuide`
[3] `https://iclr.cc/Conferences/2025/Dates`
[4] `https://iclr.cc/Conferences/2026/ReviewerGuide`

The pooled analysis uses 32,955 paper clusters and 130,023 reviews. With year indicators, the early coefficient is -0.097843 (SE 0.012084; $p < 0.001$), and the late coefficient is $+0.033349$ (SE 0.011917; $p = 0.00514$).

For the NeurIPS 2025 scope check, the official reviewing period ends on 2 July 2025 Anywhere on Earth, corresponding to 3 July 2025 at 12:00 UTC.[5] Relative position and the deadline gap have Pearson correlation 0.6022 and Spearman correlation 0.7210. On the normalised scale, official-deadline coefficients are $\beta_E = 0.00075$ ($p = 0.775$) and $\beta_L = 0.00785$ ($p = 0.0023$), consistent with the measure-dependent comparison in Section 5.3.

A.4 Confidence Adjustment

Table 10 extends the main-text late-coefficient comparison with early coefficients and confidence-by-timing interaction tests.

Table 10: Relative-timing contrast coefficients before and after confidence adjustment. Interaction p-values refer to confidence by timing. These controls do not substitute for reviewer fixed effects.

	β_E		β_L		Interaction p	
Year	Base	+ confidence	Base	+ confidence	Early	Late
2024	-0.072	-0.058	$+0.013$	$+0.006$	0.44	0.65
2025	-0.062	-0.050	$+0.041$	$+0.033$	0.97	0.93
2026	-0.096	-0.078	$+0.096$	$+0.083$	0.52	0.41

A.5 Heterogeneity by Paper Score

Table 11 supplies the exact partition and sample composition underlying Figure 4.

Table 11: Score-group analysis for 2026, using mean-centred ratings and paper-clustered inference. Exact cutoffs are 4, 4.5, and $16/3$; tied paper means produce unequal groups.

Group	Papers	Accept (%)	Mean rating	β_L	p
Q1 (≤ 4)	5,532	6.0	3.42	$+0.073$	0.005
Q2 ($4 < \bar{r} \leq 4.5$)	2,283	25.8	4.49	$+0.038$	0.374
Q3 ($4.5 < \bar{r} \leq 16/3$)	3,211	51.4	4.98	$+0.125$	< 0.001
Q4 ($> 16/3$)	3,148	88.6	5.95	$+0.038$	0.292

B Aggregation and Evaluation

We first document fitting and scale sensitivity, then give the exact aggregation results, and finally examine dependence on the evaluation metric and target.

B.1 Objective, Numerical Search, and Scale Sensitivity

For paper i, write $R_i = \sum_j r_{ij}$, $P_i = \sum_j \mathrm{rp}_{ij}$, and $Q_i = \sum_j \mathrm{rp}_{ij} r_{ij}$. Then $\hat{s}_i(a) = (R_i + aQ_i)/(n_i + aP_i)$. These sufficient statistics make each objective evaluation linear in the number of papers after a single pass over reviews. The same threshold tie rule as in the main benchmark is applied to every reported classification result. The numerical search evaluates a nominal grid from -2 to 2, but rejects nonpositive weights. Since every retained paper contains an earliest review with $\mathrm{rp} = 1$, feasible slopes satisfy $a > -1$. It then uses 20 ternary interval refinements near the best grid point. Ternary search presumes a suitably shaped local interval; no theorem of global unimodality or uniqueness is established.

Table 12 compares slopes fitted with raw and normalised ratings. The different optima quantify the scale sensitivity of the fitting objective; normalised fitting is a sensitivity analysis rather than the main aggregation rule.

[5] `https://neurips.cc/Conferences/2025/AC-Guidelines`

Table 12: Fitted slopes under two objectives. The numerical comparison searches the feasible interval $(-0.999, 2.1)$; all fitted minima shown are interior.

Fit target	2024	2025	2026
Raw rating versus decision	0.658	1.102	1.204
Normalised rating versus decision	0.759	0.662	0.549

B.2 Borderline Decision Agreement

The fixed subsets and classification rule follow Section 6.1. The confidence rule floors confidence at 10^{-3} and uses a default of 3 when absent.

Table 13: In-sample borderline decision agreement. Gains are percentage points over uniform on a fixed set of papers per year. All nonuniform methods improve over uniform with McNemar $p < 0.001$; bold identifies the highest agreement within each year.

	2024 ($n = 4{,}433$)		2025 ($n = 6{,}941$)		2026 ($n = 9{,}511$)	
Method	Agreement	Gain	Agreement	Gain	Agreement	Gain
Uniform	0.5953	0.00	0.5944	0.00	0.4511	0.00
Confidence	0.6377	+4.24	**0.6361**	+4.16	0.5308	+7.97
Heuristic timing	0.6244	+2.91	0.6244	+3.00	0.5273	+7.62
OTCA	0.6316	+3.63	0.6204	+2.59	0.5416	+9.05
JOINT	**0.6391**	+4.38	0.6355	+4.11	**0.5636**	+11.25

B.3 Full-Sample Brier Scores

We apply the metric defined in Section 3 to each paper aggregate, using the complete retained sample and paired tests on per-paper loss differences. Table 14 gives the full values underlying the metric comparison in Section 6.1.

Table 14: Full-sample score-derived Brier values (lower is better). Superscripts give paired-test significance versus uniform: $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. Unmarked values are nonsignificant.

Method	2024	2025	2026
Uniform	0.16456	0.16351	0.18540
Confidence	0.16220^{***}	0.16107^{***}	0.18347^{***}
Heuristic timing	0.16448^{**}	0.16343^{**}	0.18525^{**}
OTCA	0.16426	0.16331	0.18531
Joint	0.16195^{***}	0.16092^{***}	0.18352^{***}

B.4 Metareview Enthusiasm and Target Sensitivity

The supplementary analysis evaluates within-year aggregation on 1,197 metareviews from 2024 and 1,181 from 2025 (2,378 total); 758 and 708 papers, respectively, belong to the fixed borderline sets. An LLM rates enthusiasm from metareview prose with instructions to disregard explicit verdict words. Enthusiasm is therefore a derived signal from a decision-stage document, not an independent expert review-quality annotation.

The coding prompt asks the model to assess the strength and positivity of the area chair's reasoning on a five-point scale, with 1 clearly negative, 3 mixed or lukewarm, and 5 clearly strongly positive. Its explicit instruction is: "Ignore whatever explicit verdict word (accept/reject) appears." The implementation passes at most the first 3,000 characters of metareview text, uses temperature 0 and a 20-token response limit, and parses a single `ENTHUSIASM: 1` to `ENTHUSIASM: 5` response. Verdict words are not mechanically removed, so the instruction does not guarantee that the model ignores them. This is an additional reason to treat enthusiasm as a target-sensitivity analysis rather than independent adjudication.

Youden's criterion uses per-paper decision labels to select a threshold. Base-rate matching uses only the aggregate 40.1% acceptance rate; cutoff 4 yields 34.8% positive enthusiasm labels, closer to this rate than cutoff 3 (57.4%). Agreement at cutoff 4 is therefore conditioned on calibration choices. At the a priori midpoint, Joint provides no gain. The continuous-target analysis also shows a lower pooled Pearson correlation for OTCA (0.425) than uniform (0.444) on the borderline subset. Together, the threshold and continuous-target results limit claims that reweighting generally improves agreement with metareview reasoning.

C Representation-Level Timing Analysis

The representation probe uses a distinct prediction task and split from paper-level aggregation. We separate its implementation from the interpretation of its evaluation.

C.1 Architecture and Training

The encoder has dimensions $4 \to 64 \to 32$, with a ReLU after the first affine layer and a linear output. Predictor and discriminator each have dimensions $32 \to 32 \to 1$, with ReLU and sigmoid activations. Training uses Adam with learning rate 10^{-3}, weight decay 10^{-4}, batch size 512, and 50 epochs. We use random seed 42 and $\lambda = 1$, and update predictor and discriminator on detached embeddings before updating the encoder through the two losses. Confidence is divided by 5; rating and soundness use observed rating bounds, with rating as fallback for missing soundness. The gap feature is divided by its full-snapshot maximum.

C.2 Evaluation and Baseline Comparability

The review-level split differs from a paper-grouped split: multiple reviews of the same paper can contribute the same decision label to training and test sets. As a result, the reported test AUC should be read as an exploratory review-level measure. The timing baseline ranks the gap feature, while the acceptance baseline ranks the normalised rating. They are not matched supervised classifiers trained on the full raw feature vector. The reported AUCs therefore do not constitute an ablation isolating only the gradient reversal objective. We compute ROC AUC with average ranks for tied scores; timing scores are reported as symmetric AUC, and predictor AUCs below 0.5 are mirrored. No fresh adversary is fitted after training, no repeated-seed interval is available, and no numerical sweep over λ substantiates claims about smaller values.

A stronger evaluation would group by paper, fit an independent post-training adversary, compare matched predictive baselines, and report uncertainty across seeds.

Table 15: Exploratory ATD results for 2026. Baselines differ by task as described above. Symmetric AUC measures timing predictability after allowing label reversal.

	Symmetric timing AUC	Acceptance AUC
Task-specific raw-feature baseline	0.999	0.716
ATD representation	0.756	0.720

D Review Text and Alternative Explanations

Following Section 7, we distinguish engagement measurements, blinded quality assessments, content-based aggregation, and lexical indicators of AI assistance.

D.1 Engagement Markers

The five text features measure total word count, weakness-section word count, the weakness-to-strength word-count ratio, numerical-token density, and the number of question marks. Counts and ratios are imperfect proxies for writing behaviour and do not assess whether a criticism is valid, a question is useful, or a review is fair. Welch tests compare pooled Early and Late groups,

without reviewer controls. Their nominal significance should be interpreted in light of within-paper dependence and multiple comparisons.

In the selected NeurIPS sample, later reviews also have lower word counts ($\Delta = -17.7$, $d = -0.082$), numerical density ($\Delta = -0.0021$, $d = -0.104$), question count ($\Delta = -0.184$, $d = -0.063$), and confidence ($\Delta = -0.078$, $d = -0.094$). This selected sample does not support population-wide estimates for NeurIPS.

D.2 Blinded Pairwise Quality Assessments

Table 7 reports the six judge-by-year comparisons under the blinding, order-consistency, and length-matching protocol described in Section 7.2.

The order-consistency filter makes decisive-pair rates conditional on judge stability, while length matching restricts them to a smaller subset. Neither subset is a random sample of all review pairs. The six comparisons in Table 7 therefore assess the robustness of a quality interpretation to presentation order and length, rather than provide an independent expert-annotated quality benchmark.

D.3 Content-Based Aggregation

The supplementary content-weight experiment sums the five year-standardised engagement features and fits a separate weighting coefficient. Papers with nonpositive weights are omitted during this fit, while evaluation floors all weights at 0.01; these choices make the content comparison exploratory rather than a matched constrained-OTCA benchmark. The text composite is associated with the rating being aggregated. Residualisation removes a fitted linear component of that association but does not make the remaining text signal causally exogenous. Table 16 reports the resulting aggregation gains and the regression slope of the standardised composite on rating. The slope describes association, not causal dependence.

Table 16: Content-aware weighting. Gains are in-sample percentage-point changes in borderline agreement versus uniform.

Year	Rating-to-composite slope	Raw content gain	Residualised gain
2024	-0.34	$+4.49$	-4.06
2025	-0.34	$+3.04$	-3.70
2026	-0.39	$+11.23$	$+7.05$

D.4 Lexical Indicators of AI Assistance

Marker density measures hits per 1,000 words, while presence measures the share with at least one hit. Both contrasts use Late minus Early and the same filtered snapshots as the primary analysis. Neither is equivalent to the corpus-level estimation method of Liang et al. [13].

E Asset Terms and Research-Artifact Availability

Review records and metadata. The source records are contributed by reviewers and area chairs to ICLR 2024, 2025, and 2026 and NeurIPS 2025, and hosted by OpenReview. Under the OpenReview Terms of Use cited in Section 3, reviews, metareviews, and discussion are Comments licensed under CC BY 4.0, while associated metadata has a CC0 dedication. We preserve source attribution and comply with record-level readability restrictions. These provisions do not grant a uniform licence to the submitted article PDFs, which are not redistributed as a research asset here. Analyses and tables are derived summaries, not reproductions of the underlying papers.

Hosted language models. We credit OpenAI for GPT-4.1 and DeepSeek for DeepSeek-V4-Pro in the blinded assessments. Access to hosted models is governed by the applicable provider and, where used, intermediary service terms, rather than a licence to redistribute model weights. Relevant terms include the OpenAI Services Agreement,[6] the DeepSeek Open Platform Terms of Service,[7] and OpenRouter's terms for API routing.[8] The metareview implementation uses an OpenAI-compatible API through OpenRouter. No third-party model weights are released with this paper.

Public project repository. Code and numerical data are publicly available at `https://github.com/shiyan-liu/deadline-effect`. The package includes fixed rating and timestamp snapshots with source record identifiers, numerical text features, the 2,378 enthusiasm labels, analysis and figure-generation scripts, and the six judge-by-year summaries. The README gives installation and execution instructions; `./reproduce.sh` recomputes the statistical results and checks 705 numerical values against the reference outputs. The data card documents the sample, fields, provenance, and limitations. Original code is licensed under MIT; the data licence document distinguishes CC0 metadata from CC BY 4.0 derived features and outputs. Review prose, identities, article PDFs, and credentials are excluded. Optional scripts disclose the LLM prompts and require separately obtained source text and user-provided API access; the six historical comparisons are released at summary granularity.

Execution resources. A complete statistical replay on an Apple M5 with 16 GB memory, Python 3.14, NumPy 2.4.1, and SciPy 1.17.0 takes approximately 8 seconds and less than 0.5 GB resident memory. A separate 50-epoch ATD replay using PyTorch 2.10.0 on one CPU thread takes approximately 9 seconds and 0.5 GB resident memory. The repository provides the ATD outputs underlying Table 15 together with the execution logs. Historical hardware records and total LLM execution costs are not supplied. These resource measurements do not resolve the split and repeated-seed limitations in Appendix C.

[6] `https://openai.com/policies/services-agreement/`

[7] `https://cdn.deepseek.com/policies/en-US/deepseek-open-platform-terms-of-service.html`

[8] `https://openrouter.ai/terms`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction distinguish observational associations, in-sample decision agreement, cross-year transfer, and the exploratory ATD result. They explicitly limit claims about causality and substantive quality.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 and the appendices discuss reviewer confounding, endogenous targets, raw-scale fitting, in-sample evaluation, ATD paper overlap, and selected cross-venue data.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper presents no formal theorem. OTCA is a numerical fit, and the within-paper contrast design is not claimed to identify a causal effect.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sections 3 and 4 and Appendices A, B.1, C, and D.2 specify the samples, estimands, objectives, statistical tests, training protocol, and text-assessment controls. Appendix B.4 gives the enthusiasm-coding instructions. The six judge-by-year comparisons use the stated blinding, order-consistency, and length-matching controls. Public artifact availability is reported separately below.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The public repository linked in the abstract supplies fixed numerical inputs, stored enthusiasm labels, statistical and figure-generation code, optional ATD and LLM scripts, reference outputs, and executable reproduction instructions. Appendix E describes the package and identifies the summary granularity of the six historical judge comparisons.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: §4.3 and Appendix C report all ATD hyperparameters (Adam, lr 10^{-3}, weight decay 10^{-4}, batch 512, 50 epochs, 80/20 split, $\lambda = 1.0$). The OTCA grid-search and refinement procedure are fully described in §4.2 and Appendix B.1.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Table 8 reports paper-clustered standard errors. Figure 2 reports paper-bootstrap intervals. Paired McNemar and squared-error tests are described in Appendix A.1.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Appendix E gives measured hardware, memory, and execution time for the statistical and ATD replays. Original hardware and complete historical LLM execution-resource records remain unavailable.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The study analyses existing review records without recruiting participants. Section 8 discusses privacy, access restrictions, and potential disparate impact; deployment is not recommended without further validation.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: §8 discusses potential positive impacts (better calibrated aggregation, subject to prospective validation) and negative/fairness risks (potential disadvantage to reviewers with caregiving or time-zone constraints), with a call for disparate-impact audits before deployment.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: No new deployed system is released. Access restrictions and privacy must be respected when sharing review records; operational timing weights require fairness and incentive evaluation.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section 3 and Appendix E credit the review-record contributors, organisers, platform, and model providers, and identify the CC BY 4.0 comment licence, CC0 metadata dedication, and hosted-service terms. Article-PDF licences are distinguished from the records used in this study; no model weights are redistributed.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public repository includes a README, numerical data card, provenance manifest with source identifiers and input hashes, reproduction instructions, asset licences, and measured replay resources; see Appendix E.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper analyses existing, publicly released review data; no new human subjects were recruited and no crowdsourcing was conducted.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or any primary data collection from human subjects; IRB approval is therefore not applicable.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs were used for supplementary pairwise review assessments and metareview-enthusiasm coding, as well as editing. Methods and limitations are described in Appendices D.2 and B.4.