WHERE THE GROWTH OF PEER REVIEW GOES: A DECADE OF ICLR DISCUSSIONS ON OPENREVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Peer review in AI and machine learning conferences now runs at a size that no reviewing system was designed for. ICLR itself grew $40\times$ between 2017 and 2026 with its review process in public on OpenReview. In this paper, we use this public record to understand how each part of peer review scales with the size of a venue, what drives the growth, and what the discussion still tells us about the reviewers' judgment. Each part is summarized by a power law in the number of submissions, with the exponent showing who absorbs the growth. We find that authors absorbed most of it, as their rebuttals grew faster per paper than the reviews that they answer, while public attention did not grow at all. A small group of prolific authors holds a growing share of all submissions, which exhibits the property of cumulative advantage. We also find that the discussion became routine, with reviewer follow-ups that are frequent, short, and polite, and the review wording predicts scores somewhat less well than it did. These findings show where the load of a growing venue lands and which rules affect it.

1 INTRODUCTION

Peer review decides what a scientific field publishes, who receives credit, and who is hired, and in AI and machine learning conferences it now runs at a size that no reviewing system was designed for, as ICLR grew from 490 submissions in 2017 to 19,814 in 2026. Two independent committees at NeurIPS disagreed on about a quarter of decisions (Cortes & Lawrence, 2021; Beygelzimer et al., 2023), so for an accepted paper the second decision was close to chance. If the process rewarded the number of submissions, the field would concentrate on the most prolific authors, and if only the assigned reviewers take part, open review loses the outside scrutiny that it was created to provide (Soergel et al., 2013). Program chairs respond to growth by changing the process almost every year, e.g., with a limit on submissions per author and a duty to answer rebuttals, and whether such a rule helps depends on who absorbs the growth and on what the answers convey.

Physics, biology, and urban science describe growth with power laws, i.e., a property of a system proportional to its size raised to a fixed power, e.g., patents in a city against its population (Bettencourt et al., 2007), and we are not aware of such descriptions for peer review, because few venues have kept a public record long enough to fit them. ICLR has run its review process in public on OpenReview since 2013, and earlier analyses of this record treated the size of the venue as fixed background and left the discussion itself unstudied, which leads to the following questions.

> *How does each part of peer review scale with the size of a venue, who absorbs the extra work, and what drives that growth? Does the discussion still reflect what the reviewers think, and how far is the decision from random choice?*

The questions are hard for four reasons. The venue grew every year, so size and calendar time are correlated, and a change cannot be attributed to either alone. Forms, scales, and policies changed almost every year, so a break in a series can be a change of the form and not of the reviewers. More papers are withdrawn before the rebuttal each year, so a per-paper average mixes changes in behavior and in population. A model of such data has alternative explanations, e.g., fixed differences between authors in place of cumulative advantage, so each model is checked against them.

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.

Our work makes the following contributions along with their main findings.

A scaling description of peer review that shows who absorbs the growth. Every per-year property of the process is summarized by a power law in the number of submissions, next to a fit in the year. We found that, among the papers under review, authors add rebuttal words faster than reviewers add review words. A growing share of a year's submissions comes from a few prolific authors, whose tail grows heavier each year, and public attention shows no relation with venue size.

Measurements of the discussion and a model of when the decision breaks. The review form changed three times, so we train a text-to-score model on the reviews of years that share a form and evaluate it out of sample in every year, next to a tone-score fit on the assessment fields, i.e., the fields that hold the reviewer's judgment. We found that the wording of a review predicts its score with a lower correlation now compared to 2017 to 2021, in a gradual decline, and that reviewer follow-ups became frequent, short, and polite. A one-factor model of the ratings gives the signal share of a rating, the share of coin-flip decisions, and the point at which the decision equals random choice. We found that the signal share decreased by about 40% while a fourth review kept the panel reliability flat, and that the decision accepts about three quarters of the true top papers while over a third of the decisions are coin flips, which a fixed panel would raise to almost half by 2035.

Mechanisms behind the growth results. We infer the process that produces a growth result, which we refer to as the mechanism, e.g., who adds the extra papers or words and why. An author's papers next year rise with this year's papers at a fixed earlier count, showing cumulative advantage, i.e., those who have more gain more, and the public consists of one-time commenters and not a persistent community. Within a year, effort per paper does not depend on the size of an area or topic group, and every group follows the venue trends, while team size, reviewer tone and confidence, and withdrawal differ across different topic groups.

Related work. Surveys of peer review list the strain of reviewer load and the experiments that venues have run on it (Shah, 2022; Kuznetsov et al., 2024), and position papers describe a review crisis (Kim et al., 2025) and argue for preserving OpenReview as a research resource (Sun et al., 2025). Empirical studies of the OpenReview data have measured score variation between the reviewers of a paper, institutional effects, and review sentiment against decisions at ICLR 2017 to 2020 (Tran et al., 2020; Wang et al., 2023; Ghosal et al., 2019), and score movement during the rebuttals of 2024 and 2025 (Kargaran et al., 2025; Louis et al., 2026). All of these works treat the size of a venue as fixed, so none of them measures how any property of peer review grew with that size, and the discussion itself, i.e., who replies to whom, at what length, and in what tone, has not been measured across the lifetime of a venue. Appendix A extends the related work discussion.

2 Data and Measurements

Source. We use the official OpenReview client to fetch every submission and every note in its forum, i.e., its discussion page, for 2013, 2014, and 2017 to 2026, without PDFs. A security incident removed many notes of ICLR 2026, so note-based measurements stop at 2025, while submission-based measurements, e.g., authors, titles, and decisions, include 2026. Each note receives a role (reviewer, author, area chair, program chair, public) from its signature and invitation, i.e., OpenReview's name for a post type, and the authors' notes are author notes, which form the rebuttal, while a single such note is an author reply. Appendix B gives more details.

Overview. Figure 1 maps each quantity to its model and the difficulty that the model meets.

Growth fits. Let N_y be the submissions in year y and q_y a per-year quantity, with the power law

$$q_y \approx c\, N_y^{\alpha}, \tag{1}$$

where c is a constant and α is the exponent, fitted by least squares on the logarithms of the yearly values, with a fit standard error from their spread around the line. Because $\log N_y$ and the year have a correlation of at least 0.98, each fit is compared with a straight line in the calendar year through the R^2, which is at least as high for all but four rows (Appendix D), so the exponents summarize a decade of growth and are not evidence of a size effect. Two exponents are compared by fitting (1) to the ratio of the two series, whose slope is the difference of the exponents with its own fit standard error, because the yearly deviations of the two series move together. The area fit applies the same form across the primary areas of a single year, i.e., the areas that authors choose on submission, and so varies the count of papers at a fixed era, i.e., a year with its policies, tools, and norms.

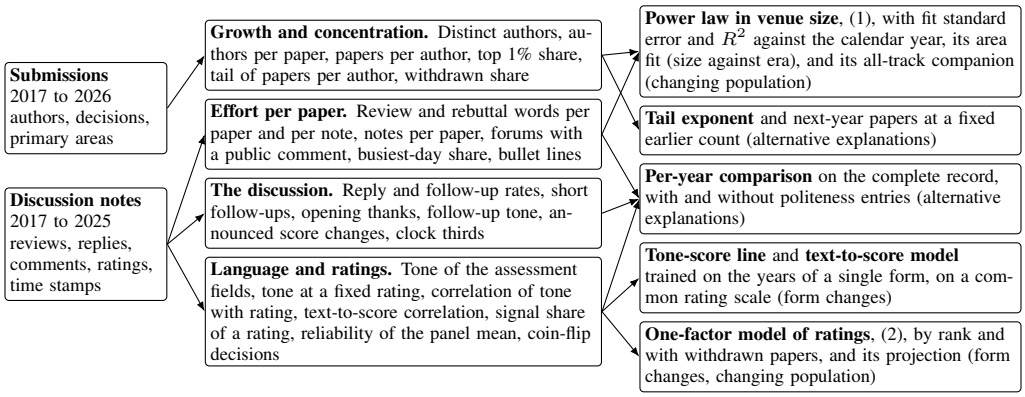


Figure 1: Data, quantities, models, and the difficulty (in parentheses) that each model meets.

Per-year quantities. Every per-year quantity uses all papers of the year, i.e., the complete record and not a sample, so it carries no sampling error, and only a fitted exponent carries a fit standard error, from the spread of the points around the line. In the findings, italics mark the proposed mechanism, i.e., the process that we propose produces the finding, and the text says whether the data support it or leave it a conjecture.

Authors and the tail of papers per author. Within a year we identify an author by profile ID (from 2021) and by email otherwise, and across years a slot without an ID takes the profile ID with its name, else the email. For each year, let k be the papers of an author, and we estimate the tail exponent γ for which the number of authors with $k \geq 2$ papers is proportional to $k^{-\gamma}$ by the discrete maximum likelihood estimate of Clauset et al. (2009), where a smaller $\hat{\gamma}$ means a heavier tail.

The forum as a tree. Every note replies to an earlier note, so the notes of a forum form a tree. A review is replied to if an author note appears below it, and followed up if the same reviewer posts again after the first author reply, which gives the author reply rate, the reviewer follow-up rate, and the share of forums with a public comment, i.e., a note from outside the review team.

Tone. Every note with at least five words receives a tone, i.e., the mean over its sentences of the sentiment score from -1 to $+1$ of the dictionary-based program VADER (Hutto & Gilbert, 2014). Politeness words such as thank are positive entries, so every tone measure is also computed without them, and for reviews we also compute the tone of the assessment fields alone. Let s_i be the tone of note i and r_i its rating on the 1-to-10 scale, or the rating of the review answered if note i is an author reply. Within each year we fit the tone-score line $s_i \approx \beta_0 + \beta_1 r_i$ by least squares and report their correlation, and the fitted tone at ratings 3 and 8 compares wording at a fixed rating.

The text-to-score model. This model predicts the rating from the words of the assessment text by a ridge regression on up to 100,000 words and word pairs, trained on the years that share a form and evaluated on every held-out year, where the correlation between predicted and actual rating measures how much of the score the wording reveals. We also count bullet lines per 1,000 review words, scan reviewer comments for announced raised, promised, kept, or lowered scores, and measure working hours by the share of a role's notes in each third of the day in coordinated universal time (UTC). Appendix C gives the form history, every word list and pattern, and the topic groups.

3 Findings

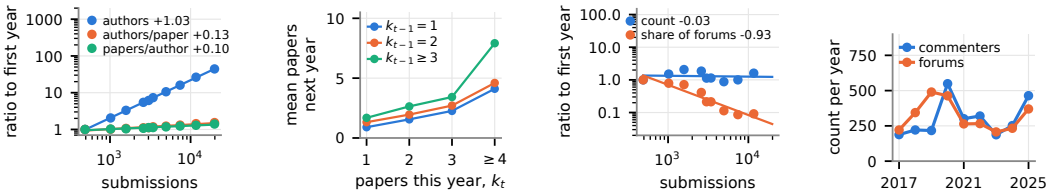
3.1 How the Process Grew and What May Drive It

Distinct authors grow in proportion to submissions, and teams grow by adding authors already on other papers of the year. Each submission still adds about three distinct authors, as in 2017 (Figure 2a), while authors per paper rose by half and papers per author by almost as much (Table 1), so a larger share of author slots belongs to authors with several submissions. *Teams grew by adding authors who already appear on other submissions of the same year, and not new participants.*

Submissions concentrate on a few prolific authors, because this year's submissions raise the author's submissions next year. The share of author slots held by the top 1% nearly doubled

Table 1: Exponents α from (1), fit standard errors (SE), and first and last fitted values, over 2017 to 2026 for submission-based rows (withdrawn share over 2018 to 2025) and 2017 to 2025 for note-based rows. Indented rows multiply to the row above.

Quantity	$\alpha \pm$ SE	First	Last	Quantity	$\alpha \pm$ SE	First	Last
Distinct authors	1.03 ± 0.01	1,458	64,649	Review words per paper	0.24 ± 0.06	912	1,800
Authors per paper	0.13 ± 0.01	3.7	5.7	Reviews per paper	0.11 ± 0.03	3.1	4.1
Papers per author	0.10 ± 0.01	1.24	1.74	Words per review	0.13 ± 0.04	296	444
Top 1% share of author slots	0.16 ± 0.01	4.8%	8.8%	Rebuttal words per paper	0.40 ± 0.06	1,416	4,525
Withdrawn share of submissions	0.50 ± 0.09	8.2%	25.6%	Author notes per paper	0.28 ± 0.08	5.6	12.3
Public-comment forums, share	-0.93 ± 0.11	45%	4.1%	Words per author note	0.12 ± 0.03	251	367
Public-comment forums, count	-0.03 ± 0.12	221	355	Follow-up words per paper	0.52 ± 0.14	86	402
Busiest-day share of reviews	-0.14 ± 0.05	30%	22%	Follow-up notes per paper	0.56 ± 0.16	0.8	3.5
Bullet lines per 1,000 words	0.34 ± 0.04	6.5	18.4	Words per follow-up	-0.04 ± 0.05	113	113



(a) Authors, ratio to first year

(b) Next-year papers at fixed past

(c) Public comments, ratio to first year

(d) Public commenters

Figure 2: Subfigures (a) and (c) are ratios to the first year, subfigure (b) next year's papers against this year's at a fixed earlier count k_{t-1}, pooled over years, and (d) commenters and forums.

(Table 1), and the tail exponent decreased from 3.6 to 2.7. *The mechanism is cumulative advantage, i.e., those who have more gain more, as a senior author with more students joins more papers.*

A rise of next year's mean count with this year's count cannot separate cumulative advantage from fixed differences between authors, but among authors present in both earlier years, next year's mean count still rises with this year's count at every fixed earlier count (Figure 2b), and a least-squares prediction from both counts gives this year's count about three times the weight of the earlier count, whereas a fixed personal rate would make the two counts equally informative. State dependence thus adds to fixed differences, so growth came increasingly from returning authors, on whom a limit on submissions per author acts.

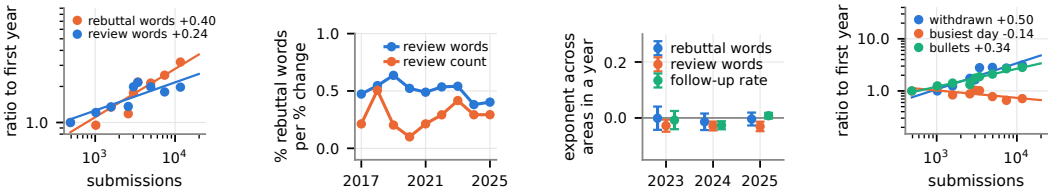
Public attention shows no relation with venue size, because public commenters are one-time visitors and not a persistent community. Forums with a public comment stayed at a few hundred a year with no trend (Table 1), so their share decreased almost as the inverse of the venue (Figure 2c). *The public is a flow of commenters who each come for a single paper and not a fixed community whose attention is divided among more papers.* A few hundred distinct commenters a year (Figure 2d) post fewer than two notes each since 2018, and few of them return the next year. *The flow does not grow with the venue, because a comment is a public good, i.e., it benefits everyone whether or not they contributed, so a visitor has no private reason to supply it (Olson, 1965).*

Reviewers and authors both add words to the discussion, and the difference between them lies in the number of notes. Rebuttal words per paper grow faster than review words per paper, by 2.8 fit standard errors in the exponents (Figure 3a), whereas per note the two sides grew alike, so the difference lies in the number of notes, as reviews per paper rose by a third while author notes per paper more than doubled (Table 1). *We conjecture that the split of a rebuttal into several notes, which OpenReview's limit on comment length forces, explains the count of notes, and that the extra words come from the drafting tools of 2023 and the expectation that a rebuttal answers every point.*

The faster growth on the authors' side belongs to the papers that stayed under review, because withdrawn papers leave the main track, i.e., the papers that reach a decision, after their reviews, so over all submissions with quoted review text removed the difference lies within its fit standard error.

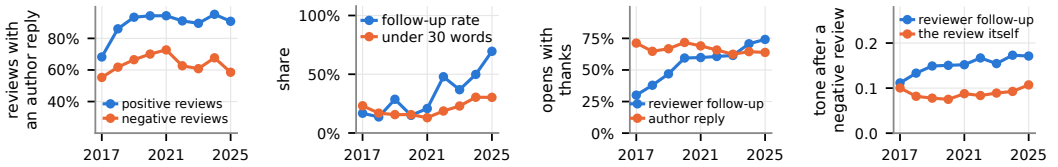
Within a year, effort per paper does not depend on the size of an area, as the exponents across areas are near zero (Figure 3c), and the extra rebuttal words do not depend on the review answered, as a review 10% longer went with a rebuttal 5% longer up to 2023 and 4% longer after (Figure 3b).

The share of reviews posted on the busiest day decreased, the withdrawn share of submissions tripled, and reviews became itemized lists. The busiest-day share of reviews decreased by about a quarter, while the share within a day of it showed no trend (Figure 3d). *A spread over time zones or*



(a) Words per paper, ratio to first year (b) Rebuttal response (c) Scaling across areas (d) Withdrawn, busiest day, bullets

Figure 3: Effort per paper and three further quantities. Subfigures (a) and (d) are each quantity as a ratio to its first year, subfigure (b) the response of rebuttal length to review length and count, and subfigure (c) the exponent across primary areas with its fit SE.



(a) Author reply rate (b) Reviewer follow-ups (c) Opens with thanks (d) Tone after negative review

Figure 4: The discussion. Subfigure (a) is the author reply rate by review sign, subfigure (b) the follow-up rate and the short-follow-up share, subfigure (c) the thanks-opening share of replies by side, and subfigure (d) the tone of a negative review and its follow-up.

earlier completion lowers only the single-day share, whereas a laxer deadline would raise the share posted after it, which decreased. Bullet lines per 1,000 words nearly tripled (Table 1). *Withdrawn papers carried a mean rating more than half a point below that of rejected papers in every year, so we conjecture that withdrawal responds to low scores and is worth more with more alternative venues.* The withdrawn share of papers rated below 5 rose from 1 in 8 in 2018 to 1 in 2 in 2025, so the papers under review lose a growing share of their lowest-rated members.

3.2 The Discussions: Who Talks, How Much, How Kindly, and When

Reviewers now answer rebuttals, but a growing share of their answers are short routine replies. The follow-up rate of replied-to reviews rose $4\times$, and the share of first follow-ups under 30 words decreased to 2021 and then more than doubled (Figure 4b). *We conjecture that the duty to respond and the expectation that a good rebuttal earns a higher score made replies routine.* Authors reply to almost every positive review, i.e., rated 6 or above, and to three in five negative ones (Figure 4a).

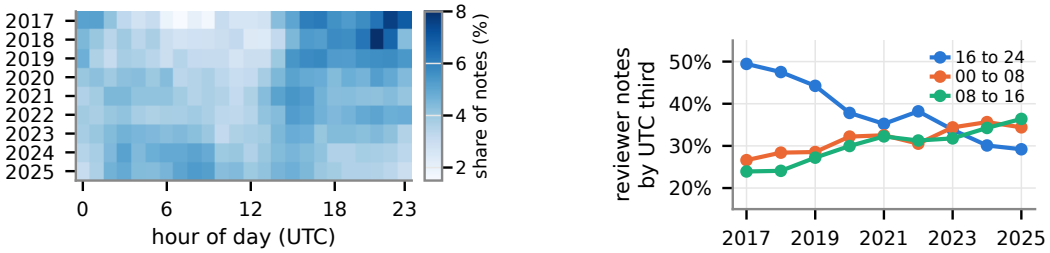
Reviewer follow-ups adopted the opening thanks of author replies, whose share changed little. About two thirds of first author replies opened with thanks in every year, whereas among reviewer follow-ups the share more than doubled to three quarters (Figure 4c). *We conjecture that the follow-up became a routine reply that confirms receipt, and a routine reply opens with a formula.*

The added positive tone of reviewer follow-ups is mostly politeness formulas. After a negative review, the first follow-up scores as more positive than the review in every year, and its tone rose by half (Figure 4d), whereas without the politeness entries it is flat and within 0.035 of the review's tone. *Most of the added positive tone is therefore the opening formula and not a milder judgment.*

The working hours of reviewers and authors shifted from the Americas towards Asia. Reviewer activity peaked in the American afternoon in 2017 and in the East Asian afternoon in 2025 (Figure 5a), and the share of notes in the last UTC third decreased by two fifths for reviewers and by almost a third for authors (Figure 5b). *We conjecture that the cause is the composition of the reviewer pool, drawn from the authors, and that submissions grew largely in Asia.*

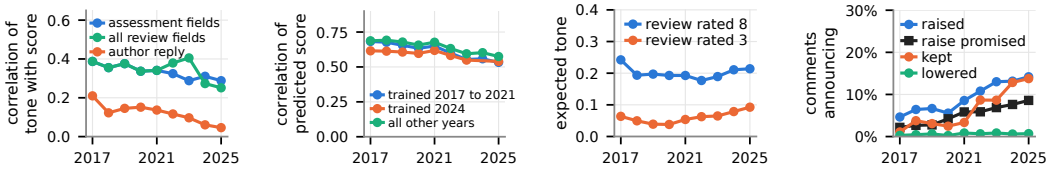
3.3 Alignment: What Reviews Say and What They Score

The wording of a review predicts its score with a lower correlation than it did, in a gradual decline, and the apparent break in 2024 was a change of the review form. On the full text, the correlation of tone with rating was stable to 2023 and decreased by a third in 2024, whereas the



(a) Reviewer activity by UTC hour (b) Thirds of the UTC day

Figure 5: Subfigure (a) is the share of reviewer notes by UTC hour, and (b) the share by UTC third.



(a) Tone-score correlation (b) Text-to-score predictor (c) Tone at a given score (d) Score announcements

Figure 6: Tone and scores. Subfigure (a) is the correlation of tone with rating by field set and for replies, subfigure (b) the held-out correlation of the text-to-score model, subfigure (c) the fitted tone at ratings 3 and 8, and subfigure (d) announced score changes per comment.

assessment fields show none (Figure 6a). *The 2024 form dropped the verdict summary field and added a questions field, so the full text became less verdict-like and the assessment did not.*

The text-to-score model agrees (Figure 6b). Trained on the single-field years 2017 to 2021, it predicts a held-out year of that period better than any later year, the decline is gradual, about a fifth, and 2024 is no lower than 2023 under any training set.

Negative reviews are worded more positively than in 2019 and 2020, as the fitted tone of a review rated 3 more than doubled from its low of those years (Figure 6c), and the tone of an author reply carries a quarter of its 2017 correlation with the rating of the review answered (Figure 6a).

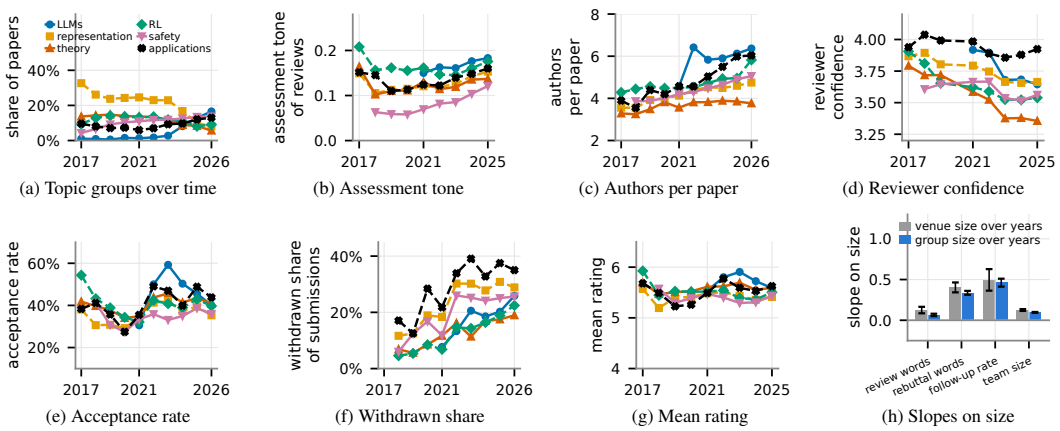
Announced score changes rose in every category except lowerings, and they capture a minority of the actual changes. Announced raises tripled and kept scores rose from almost none to as many as raises, while lowerings were announced in under 1% of comments (Figure 6d). The announced raises of 2024 and 2025 are about two fifths, and the lowerings about a third, of the changes that two studies found from scores before and after the rebuttal (Kargaran et al., 2025; Louis et al., 2026).

3.4 Topic Groups: Shared Trends and Distinct Levels

Every topic group grew, and the mix of the venue shifted from representation learning and theory towards language models. A text classifier on the keywords, title, and abstract, trained on the primary areas of 2025 and 2026, assigns every submission to a single group among 11 groups that coarsen those areas, e.g., LLMs and foundation models, and agrees with the authors' area for three in five held-out papers. The LLM share rose more than $10\times$ to a sixth of the venue and the safety share tripled, whereas representation and transfer learning decreased from a third to under a tenth and theory lost half of its share (Figure 7a).

Effort per paper and reviewer engagement rose in every group with the calendar year, and within a group they did not rise with the size of the group. Rebuttal words and the follow-up rate rose in every group, whose rebuttal exponent exceeds its review exponent by over a fit standard error. The slope of each quantity on group size over the years matches its slope on venue size (Figure 7h), and a fit on the group-year panel with a group term, log group size, and the year separates the two, as groups grew at different rates. With the year held fixed, the slope on group size is within about a tenth of zero for review words, rebuttal words, and follow-ups, whereas the year term carries the trend. *Effort per paper therefore grows with the era and not with size*, and team size is the exception.

Reviewers assess the safety group least positively in every year, and the LLM and RL groups most positively. The assessment tone of the safety group lies below every other group's in every year since 2018, by a fifth to a half of the venue mean, whereas the RL group up to 2021 and the



(a) Topic groups over time (b) Assessment tone (c) Authors per paper (d) Reviewer confidence

(e) Acceptance rate (f) Withdrawn share (g) Mean rating (h) Slopes on size

Figure 7: Six of the 11 groups, with fit SEs in (h) and the legend of (a) throughout. Subfigure (d) skips 2020, whose form asked for experience instead of confidence, and (h) gives the log-log slope of each quantity on venue size (Table 1) and on group size.

LLM group from 2022 hold the most positive tone (Figure 7b). *We conjecture that the gap reflects the topic's vocabulary, e.g., attack, bias, and harm, which a lexicon scores as negative at any rating.*

Teams in the theory group are the smallest in every year and grew least, whereas teams in the LLM and applications groups grew most. Authors per paper in the theory group rose by under a fifth over the decade, against a half in the venue, so an LLM paper now carries more than $1.5\times$ the authors of a theory paper (Figure 7c). *We conjecture that team size follows the divisibility of the work, because experiments divide among contributors whereas a proof does not.*

Reviewer confidence and withdrawal are highest for applied papers and lowest for theory papers. Reviewer confidence has been highest in the applications group in every year but 2022 and lowest in theory since 2021 (Figure 7d), and the withdrawn share has been highest in the applications group since 2020 and lowest in the theory and RL groups since 2023 (Figure 7f). Acceptance tracks mean rating across groups, as since 2022 the LLM and applications groups rank highest on both and safety lowest, so its less positive tone goes with lower scores, and in 2023 the LLM group was accepted almost twice as often as the safety group (Figures 7e and 7g). *We conjecture that confidence follows how directly a claim can be checked, as a benchmark score is checked more directly than a proof, and that withdrawal is worth more to applied work, which has more venues.*

4 When Peer Review Breaks: A Model of the Signal in a Rating

We study how much of a rating is due to the paper and how much to the reviewer, and what that mix means for the accept-or-reject decision. In our model, each rating is the paper's true quality plus the reviewer's noise. Let q_i denote the true quality of paper i, r_{ij} the rating of reviewer j on paper i, and m the number of reviews of a paper (the panel). We assume the following. **A1.** $r_{ij} = q_i + e_{ij}$, where the quality q_i has mean μ and variance σ_q^2 across papers, the noise e_{ij} has mean zero and variance σ_e^2, and quality and noise are independent of each other and across reviewers, because ICLR assigns reviewers independently and hides other reviews until all are in. **A2.** quality and noise are Gaussian, which fits the bell-shaped ratings. **A3.** the share a of papers with the highest panel mean $\bar{r}_i$, i.e., the mean of its m ratings, is accepted, which holds for most decisions.

By A1 the variance of a rating is $\sigma_q^2 + \sigma_e^2$. Let the signal share ρ denote the fraction of it that comes from the paper, which by A1 also equals the correlation between two ratings of the same paper, as they share only the quality. The panel mean is $q_i + \bar{e}_i$, whose mean noise has variance σ_e^2/m because independent errors partly cancel, and let the reliability R_m denote the signal share of the panel mean, i.e., its squared correlation with the true quality, so that

$$\rho = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_e^2}, \qquad R_m = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_e^2/m} = \frac{m\rho}{1 + (m-1)\rho}, \tag{2}$$

where the last form divides numerator and denominator by $(\sigma_q^2 + \sigma_e^2)/m$. For example, a signal share of one half means that two ratings of the same paper correlate at one half, and a panel of four then

reaches a reliability of four fifths. We estimate ρ per year from all main-track ratings by analysis of variance (Appendix F), and any venue with public ratings can compute the same yardstick.

From the model to the decision quantities. Let $Z_i = (\bar{r}_i - \mu)/\sqrt{\mathrm{Var}(\bar{r}_i)}$ and $Z_i^q = (q_i - \mu)/\sigma_q$ denote the standardized panel mean and quality. By A1 their covariance is σ_q^2 and $\mathrm{Var}(\bar{r}_i) = \sigma_q^2/R_m$ by (2), so by A2 they are standard Gaussian with correlation $\sqrt{R_m}$, and by A3 paper i is accepted if and only if $Z_i > z_a$ with $z_a = \Phi^{-1}(1-a)$, where Φ and ϕ are the standard Gaussian distribution and density. Every quantity below follows from these two facts through R_m and a alone. Regressing Z_i^q on Z_i over the accepted tail gives the expected quality of an accepted paper, $\mathbb{E}[q_i \mid Z_i > z_a] - \mu = \sqrt{R_m}\,\sigma_q\,\phi(z_a)/a$. The true-top share S is the joint tail probability over a, and a second panel's standardized mean Z_i' has $\mathrm{Corr}(Z_i, Z_i') = R_m$, as both share q_i only, so the two-panel disagreement rate D, unobservable with one panel per paper, is twice the split probability, i.e.,

$$S = \frac{\Pr(Z_i > z_a,\ Z_i^q > z_a)}{a}, \qquad D = 2\big(a - \Pr(Z_i > z_a,\ Z_i' > z_a)\big), \tag{3}$$

where the joint probabilities are bivariate Gaussian tail probabilities at correlations $\sqrt{R_m}$ and R_m. Given q_i, the panel mean is q_i plus Gaussian noise of standard deviation $\sigma_e/\sqrt{m} = \tau\sigma_q$ with $\tau = \sqrt{(1-R_m)/R_m}$, so it is accepted with probability $p(q_i) = \Phi\big((Z_i^q - z_a/\sqrt{R_m})/\tau\big)$, and we count the decision as a coin flip if $p(q_i)$ lies between one quarter and three quarters, i.e., if Z_i^q lies within $z_{0.75}\tau$ of $z_a/\sqrt{R_m}$ with $z_{0.75} = \Phi^{-1}(0.75)$, so integrating the standard Gaussian quality over that band gives the coin-flip share $\kappa = \Phi(z_a/\sqrt{R_m} + z_{0.75}\tau) - \Phi(z_a/\sqrt{R_m} - z_{0.75}\tau)$. Random choice is $R_m = 0$, where Z_i is independent of Z_i^q and Z_i', so the gain is zero, $S = a$, and $D = 2a(1-a)$, and perfect knowledge is $R_m = 1$, where $S = 1$ and $D = \kappa = 0$. Peer review thus equals random choice exactly at $\rho = 0$, and the approach is gradual, as $R_m \approx m\rho$ for a small share.

The signal share of a rating decreased over the decade, mostly by 2018, and a fourth review kept the panel reliability flat. Most of the decrease happened by 2018 and continued, partly because low-rated withdrawn papers left the main track (Figure 8a). The reliability stayed near its 2018 value because the venue moved to four reviews in 2021, whereas R_3 at the 2025 share lies below every observed R_m. A fourth assumption links ρ to venue size. **A4.** the noise variance is inversely proportional to attention per review, as more claims take more looks, and under A4 the load per reviewer of Table 1 explains at most half of the fitted slope, so attention per review decreased faster.

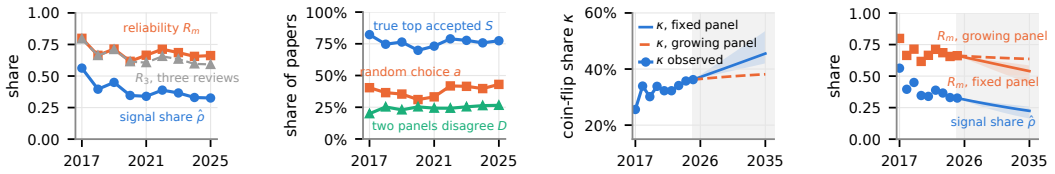
The decision still accepts most of the true top papers, but the share of coin-flip decisions rose. Since 2018 the decision accepted three quarters of the true top papers every year, so S lies far above a (Figure 8b). D rose to the level measured with two real panels at NeurIPS (Cortes & Lawrence, 2021; Beygelzimer et al., 2023), and κ rose by more (Figure 8c), because the panel noise widened the band while the acceptance rate moved towards one half, where quality is densest.

The prediction model. The projection has three inputs per future year y, the venue size N_y, the signal share ρ_y, and the panel size m_y, and holds a at its 2025 value. Submissions double about every two years along their fitted growth, ρ follows its fitted power law in venue size, and the panel size stays at its 2025 value or follows the reviews-per-paper power law of Table 1, so that

$$N_y = N_{2025}\, e^{b\,(y-2025)}, \qquad \rho_y = \hat{\rho}_{2025}\Big(\frac{N_y}{N_{2025}}\Big)^{\beta}, \qquad m_y = m_{2025}\Big(\frac{N_y}{N_{2025}}\Big)^{\gamma}, \tag{4}$$

where b is the fitted growth rate, β is the fitted slope of $\log\hat{\rho}$ on $\log N$, and γ is zero for the fixed panel and the reviews-per-paper exponent of Table 1 for the growing panel. Then R_m, S, D, and κ of year y follow from ρ_y, m_y, and a by (2) and (3), with β fitted over 2018 to 2025 and a band over the slopes within one standard error of that fit and of the fit over 2017 to 2025. The growing panel continues the observed trend, grows by about 8% per doubling, and reaches about 5 reviews in 2030 and 6 in 2035, whereas holding R_m at its 2025 value requires, by (2), $m_y = m_{2025}\,\hat{\rho}_{2025}(1-\rho_y)/(\rho_y(1-\hat{\rho}_{2025}))$, i.e., about 7 by 2035.

If the fitted decline continues, a fixed panel of four makes almost half of the decisions coin flips by 2035, whereas a growing panel holds the share near its 2025 value only if the added reviews carry the same signal. Under (4) with the fixed panel, κ passes two fifths by 2030 and R_m decreases by about a fifth by 2035, whereas the growing panel holds both near their 2025 values (Figures 8c and 8d). Both rules share ρ_y, so the growing panel assumes that reviewer attention and expertise do not decline with the added load. If instead each added review lowers ρ through A4 in proportion to the load, the noise variance grows as m_y/m_{2025} while the panel averages m_y ratings,



(a) Signal share and reliability (b) What the decision recovers (c) Coin-flip decisions (d) Projection

Figure 8: Main-track papers. Subfigure (a) shows $\hat{\rho}$, R_m, and R_3, subfigure (b) the true-top share S, random choice a, and disagreement D, and subfigures (c) and (d) project κ, $\hat{\rho}$, and R_m by (4) with a band over the fitted slopes.

so the noise of the panel mean stays at the fixed panel's and all four quantities coincide with the fixed-panel curves. The two curves bound the growing panel's outcome, and a faster decline of reviewer expertise, which the decreasing confidence of Section 3.4 hints, would lower both.

5 Conclusion

Suggestions for authors. A long rebuttal does not stand out, because rebuttal words per paper grew $3\times$ while words per review grew by half (Table 1), and a review 10% longer now draws a rebuttal only 4% longer (Section 3.1). Author notes per paper doubled, so a reply to every point is the norm, and the tone of a reply carries a quarter of the information about the rating that it did in 2017 (Section 3.3), so the effort belongs in the content. Public comments are rare (Section 3.1) and about two in five of them concern prior work, so they will not surface a flaw, and reviewer confidence is lowest for theory papers (Section 3.4), so a theory paper gains from a checkable claim.

Suggestions for reviewers. 7 in 10 replied-to reviews receive a follow-up, but the share under 30 words more than doubled and their added tone is a formula (Section 3.2), while announcements capture two fifths of the raises and a third of the lowerings (Section 3.3), so a follow-up carries information only if it states the new judgment and its reason. The expected tone at a fixed rating rose (Section 3.3), so polite wording no longer signals a score, and a rating carries a third of its variance as signal (Section 4), so the reasons in the text are worth more than the score.

Suggestions for area chairs. Half of the papers rated below 5 were withdrawn in 2025 (Section 3.1), so decided papers are borderline. A reviewer who changes a score should state the reason, as three quarters of first follow-ups open with thanks (Section 3.2), so a short follow-up marks compliance instead of agreement. The fourth review kept the panel reliability flat while the signal share decreased (Section 4), so a panel of four is worth keeping, and a growing panel helps only if the added reviews carry the same signal. Coin-flip decisions sit in a band around the threshold, so an extra review is worth most for a paper near it, whereas a paper far from it is decided.

Limitations. Appendix E matches the mechanisms to established theories and does not compare them with alternatives. The 9 or 10 yearly points cannot tell a power law from other smooth curves, and year, venue size, the review form, and the topic mix are correlated at about 0.8 or more, so the venue-wide fits identify none of them. The true quality of a paper is not observed, so the signal share measures how far reviewers agree, and a shared bias counts as quality. The ratings are final, so convergence in the discussion overstates the share. The projection assumes that the fitted slope continues, whereas future growth may depart from it, depending on how AI develops. These caveats bound the interpretation and not the measurements, which cover the complete public record of the largest venue in the field and are released with the code. The comparisons within a year and across topic groups need no separation of size from era.

Conclusion. In this paper, we have measured how each part of ICLR's review process grew with the venue and checked a mechanism for each result. Among papers that stayed under review, authors added rebuttal words faster than reviewers added review words, through the number of notes. Submissions concentrated on a few prolific authors, as cumulative advantage predicts, public attention did not grow with the venue, and every topic group followed these trends, with effort rising by the year and not by size, while the groups differ in team size, tone, and withdrawal. Review wording predicts scores less well, the signal share of a rating decreased while a fourth review kept the reliability flat, over a third of decisions are coin flips, and follow-ups became short and polite. A cross-venue comparison within a year is future work.

AI USE STATEMENT

In this work, we used generative AI tools to help develop conceptual frameworks, design or provide feedback on research methodology or experiments, implement methods, clean and reformat the dataset, support qualitative and thematic data analysis, and interpret results. We have not used generative AI tools to generate synthetic data sets, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, or assist with translation. Among the recommended disclosures, we used generative AI tools to create and edit software code, summarize or analyze existing literature, and edit a research paper to improve readability. We have reviewed all AI-assisted work. The authors set the research direction and questions, reviewed and corrected every AI-assisted step, ran all code end to end on the data, checked every number in the paper against the output tables, and verified every reference against its published source. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

All data used in this work are public on OpenReview and were retrieved with the official client under its terms of use. No PDF was downloaded and no attempt was made to identify anonymous reviewers. Author names are used only to count distinct participants and are not reported. The clock shares are aggregates over tens of thousands of notes and do not locate any individual.

REPRODUCIBILITY STATEMENT

The code is available at `https://anonymous.4open.science/r/iclr-growth-DA42`. It contains the downloader, the loader with role inference, every measurement, and the figure and table scripts, with a README that lists the reproduction steps in order. The raw per-year dumps are not included, because the downloader fetches them from OpenReview through the official client in about a quarter of an hour, and the whole pipeline from the download to every figure and table runs in under an hour on a laptop. Appendix B lists the invitations, role rules, and decision labels for each year, and Appendix C gives every lexicon, pattern, and template.

REFERENCES

Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *Science*, 286 (5439):509–512, 1999.

Luís M. A. Bettencourt, José Lobo, Dirk Helbing, Christian Kühnert, and Geoffrey B. West. Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National Academy of Sciences*, 104(17):7301–7306, 2007.

Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? The NeurIPS 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.

Joydeep Biswas, Sheila Schoepp, Gautham Vasan, Anthony Opipari, Arthur Zhang, Zichao Hu, Sebastian Joseph, Matthew Lease, Junyi Jessy Li, Peter Stone, Kiri L. Wagstaff, Matthew E. Taylor, and Odest Chadwicke Jenkins. AI-assisted peer review at scale: The AAAI-26 AI review pilot. *arXiv preprint arXiv:2604.13940*, 2026.

Lutz Bornmann and Rüdiger Mutz. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11):2215–2222, 2015.

Penelope Brown and Stephen C. Levinson. *Politeness: Some Universals in Language Usage*. Cambridge University Press, 1987.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703, 2009.

Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? Insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1274–1290, 2019.

Tirthankar Ghosal, Rajeev Verma, Asif Ekbal, and Pushpak Bhattacharyya. DeepSentiPeer: Harnessing sentiment in review texts to recommend peer review decisions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1120–1130, 2019.

Alvin W. Gouldner. The norm of reciprocity: A preliminary statement. *American Sociological Review*, 25(2):161–178, 1960.

Albert O. Hirschman. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States*. Harvard University Press, 1970.

C. J. Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, pp. 216–225, 2014. doi: 10.1609/icwsm.v8i1.14550.

Sangkeun Jung, Goun Pyeon, Inbum Heo, and Hyungjin Ahn. What drives paper acceptance? A process-centric analysis of modern peer review. *arXiv preprint arXiv:2509.25701*, 2025.

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)*, pp. 1647–1661, 2018. doi: 10.18653/v1/N18-1149.

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.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81634–81651, 2025.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into LLM-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27): eadt3813, 2025.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K. Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, Sheng Lu, Mausam, Margot Mieskes, Aurélie Névéol, Danish Pruthi, Lizhen Qu, Roy Schwartz, Noah A. Smith, Thamar Solorio, Jingyan Wang, Xiaodan Zhu, Anna Rogers, Nihar B. Shah, and Iryna Gurevych. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

Grant Lewison and James Hartley. What's in a title? Numbers of words and the presence of colons. *Scientometrics*, 63(2):341–356, 2005.

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, 2024.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, Diyi Yang, Christopher Potts, Christopher D. Manning, and James Zou. Quantifying large language model usage in scientific papers. *Nature Human Behaviour*, 9(12):2599–2609, 2025. doi: 10.1038/s41562-025-02273-8.

Alfred J. Lotka. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, 16(12):317–323, 1926.

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.

Shama Maganur and Mayank Kejriwal. Structural shifts in institutional participation and collaboration within the AI arXiv preprint research ecosystem. *arXiv preprint arXiv:2602.03969*, 2026.

Robert K. Merton. The Matthew effect in science. *Science*, 159(3810):56–63, 1968.

John W. Meyer and Brian Rowan. Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, 83(2):340–363, 1977.

M. E. J. Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

Mancur Olson. *The Logic of Collective Action: Public Goods and the Theory of Groups*. Harvard University Press, 1965.

Derek de Solla Price. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27(5):292–306, 1976.

Giuseppe Russo, 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. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28, 2025. doi: 10.1145/3757667.

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

Herbert A. Simon. On a class of skew distribution functions. *Biometrika*, 42(3/4):425–440, 1955.

Herbert A. Simon. Designing organizations for an information-rich world. In Martin Greenberger (ed.), *Computers, Communications, and the Public Interest*, pp. 37–72. Johns Hopkins Press, 1971.

David Soergel, Adam Saunders, and Andrew McCallum. Open scholarship and peer review: A time for experimentation. In *ICML Workshop on Peer Reviewing and Publishing Models*, 2013.

Ivan Stelmakh, Charvi Rastogi, Nihar B. Shah, Aarti Singh, and Hal Daumé III. A large scale randomized controlled trial on herding in peer-review discussions. *PLOS ONE*, 18(7):e0287443, 2023.

Hao Sun, Yunyi Shen, and Mihaela van der Schaar. OpenReview should be protected and leveraged as a community asset for research in the era of large language models. *arXiv preprint arXiv:2505.21537*, 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.

David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of OpenReview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

Gang Wang, Qi Peng, Yanfeng Zhang, and Mingyang Zhang. What have we learned from OpenReview? *World Wide Web*, 26(2):683–708, 2023.

Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper Copilot: Tracking the evolution of peer review in AI conferences. In *The Fourteenth International Conference on Learning Representations*, 2026.

APPENDIX CONTENTS

A Extended Related Work **14**

B Data Details **16**

C Measurement Details **18**

D Full Per-Year Tables and Further Trends **27**

E Further Analyses and Interpretations **32**

F The Signal Share of a Rating: Derivations, Estimator, and Projection **39**

A Extended Related Work

Studies of the OpenReview data. Wang et al. (2023) compiled 11,915 submissions and their 41,276 reviews and related scores to citations and preprint versions. Jung et al. (2025) analyzed over 28,000 ICLR submissions from 2017 to 2025 and found that substantive author-reviewer communication during the rebuttal raises the odds of acceptance. This agrees with the borderline result in Appendix D. Yang et al. (2026) built collections of reviews across computer-science venues and documented reviewer workload, templated reviews, and the interventions venues adopted in response. Sun et al. (2025) argue that the OpenReview record should be preserved as a community asset for research on peer review and language models. These works describe levels and policies and do not fit scaling exponents or propose mechanisms.

The review crisis and AI-assisted review. Shah (2022) surveys the challenges of peer review at large sizes, among them reviewer load, and the experiments and computational remedies tried against them. Kuznetsov et al. (2024) survey what natural language processing can contribute to each step of the process. Kim et al. (2025) describe the strain that submission growth puts on review quality and propose author feedback and reviewer rewards. Biswas et al. (2026) report the first venue-wide pilot in which every AAAI-26 submission received an identified AI review, and Thakkar et al. (2026) randomized LLM feedback to reviewers at ICLR 2025. Maganur & Kejriwal (2026) track team sizes and academic-industry co-authorship in AI preprints from 2021 to 2025. These works motivate the question of size, and Yang et al. (2026) document reviewer workload across venues, but none measures how the discussion changed with the size of a single venue.

Datasets of reviews. PeerRead (Kang et al., 2018) gathered reviews from ACL 2017, NeurIPS 2013 to 2017, and ICLR 2017 and framed acceptance and score prediction as natural language processing (NLP) tasks. Gao et al. (2019) released reviews and author responses from ACL 2018 with before- and after-rebuttal scores and found that after-rebuttal scores are largely predicted by the initial ones. Ghosal et al. (2019) used the sentiment of ICLR 2017 and 2018 reviews to predict decisions, and Stelmakh et al. (2023) ran a randomized trial at ICML 2020 on whether the first argument in a reviewer discussion sways the outcome. These works cover a year or two of a venue and do not measure scaling.

LLM-assisted text in peer review. Liang et al. (2024) introduced a distributional estimator of the fraction of LLM-modified text and applied it to ICLR 2024, NeurIPS 2023, CoRL 2023, and EMNLP 2023 reviews. They found that reviews with such text are submitted closer to the deadline and cite fewer references. Liang et al. (2025) applied the same estimator to papers across arXiv, bioRxiv, and Nature journals. Russo et al. (2025) used a commercial detector at ICLR 2024 and found that AI-assisted reviews score higher and raise the acceptance chance of borderline papers. Thakkar et al. (2026) randomized an LLM feedback agent over ICLR 2025 reviews, after which reviewers who received feedback wrote longer reviews. These works measure prevalence and effects within a year. Kobak et al. (2025) measured excess vocabulary in millions of abstracts and identified the marker words of 2024, whose share in author replies decreased after 2024 in our data (Figure D.3a). Our fixed text-to-score model adds that the relation between review wording and score declined gradually across the arrival of language models. There is no break in 2024 once the review form is held fixed.

Rebuttals and score changes. Kargaran et al. (2025) categorized ICLR 2024 and 2025 reviews and rebuttals with an LLM and found that initial scores and co-reviewer ratings predict score changes. Louis et al. (2026) extracted content features from about 73,000 reviewer score histories at the same two conferences, showing that the initial review structure bounds the movement. Both use before- and after-rebuttal scores that were saved externally. We do not have before-rebuttal scores, so we measure score changes through reviewer announcements, which extend back to 2017. Their counts calibrate the announcements in Appendix E.

Comparison with science at large in detail. Publication counts across science grow by less than 10 percent a year and double in about a decade (Bornmann & Mutz, 2015), whereas ICLR submissions doubled about every two years, with a fitted doubling time of 1.9 years. Mean team size across science and engineering took half a century to nearly double (Wuchty et al., 2007), whereas authors per paper at ICLR grew by half in 9 years, which is about three times faster per year. Lotka's law describes productivity over a career with an exponent of 2 (Lotka, 1926). The within-year tail at ICLR is a different quantity, estimated at $k_{\min} = 2$ by the discrete maximum likelihood estimate

of Clauset et al. (2009). Its exponent moved from about 3.6 to 2.7, so the tail is heavier each year. It is set next to the career exponent only as a contrast and is not compared with it as if the two were the same quantity. In cities, quantities that depend on social interaction scale with population with exponents between 1.1 and 1.3, and infrastructure scales with exponents near 0.8 (Bettencourt et al., 2007). Total rebuttal words at ICLR, i.e., rebuttal words per main-track paper times main-track papers, grow as $N^{1.30}$, and total review words grow as $N^{1.13}$, by direct fits over 2017 to 2025. These exponents are below the per-paper exponents of Table 1 plus one, because main-track papers grow as $N^{0.90}$ once withdrawals are excluded. Both totals grow faster than the venue, as social interaction grows in cities. The two sides of the same conversation have different exponents, which has no urban analog. Public attention has an exponent near zero, which is lower than the exponent of any urban infrastructure quantity.

B Data Details

Partial exclusion of ICLR 2026. ICLR 2026 was downloaded in full and appears in Table B.1. During its discussion period, OpenReview had a security incident, and in the removal of notes that followed many notes were deleted or restricted. The forums of 2026 therefore hold an incomplete discussion, and every measurement of reviews, replies, follow-ups, tone, and timing would be biased by what is missing. Those measurements stop at 2025. Author lists, titles, abstracts, keywords, and decisions of 2026 were not affected, so measurements that use only submissions, namely the growth, team size, productivity, and vocabulary results, include 2026. The withdrawn share of 2026 enters only as a check of its fit in Appendix C, because the incident may have added withdrawals.

Table B.1: The data. Reviews, author notes, and public notes count non-submission notes by inferred role. ICLR 2015 was not hosted on OpenReview, and ICLR 2016 hosted only a workshop track there. Only submission-based measurements use ICLR 2026, as explained above.

Year	Submissions	Accepted	Withdrawn	Reviews	Author notes	Public notes	Rating scale
2013	67	23	0	204	130	39	none
2014	69	n/a	0	223	186	139	none
2017	490	198	0	1,511	2,764	800	1 to 10
2018	1,018	337	83	2,930	4,003	1,032	1 to 10
2019	1,579	502	160	4,764	8,214	1,449	1 to 10
2020	2,594	687	369	7,775	11,047	932	1, 3, 6, 8
2021	3,014	859	403	11,508	17,324	462	1 to 10
2022	3,422	1,094	779	13,161	22,924	452	1, 3, 5, 6, 8, 10
2023	4,955	1,573	1,145	18,575	33,802	307	1, 3, 5, 6, 8, 10
2024	7,404	2,259	1,661	28,028	56,083	378	1, 3, 5, 6, 8, 10
2025	11,672	3,703	2,991	46,748	118,046	663	1, 3, 5, 6, 8, 10
2026	19,814	5,336	5,203	75,859	164,668	396	0, 2, . . . , 10

Sources. Years up to 2023 are served by version 1 of the OpenReview application programming interface (API v1) and later years by API v2. OpenReview names each type of post by an invitation string. The submission invitations are `ICLR.cc/YYYY/conference/-/submission` for 2013, 2014, and 2017, `ICLR.cc/YYYY/Conference/-/Blind_Submission` for 2018 to 2023, and `ICLR.cc/YYYY/Conference/-/Submission` for 2024 to 2026, which returns every status. For 2018 to 2023 the `Withdrawn_Submission` (2018 on) and `Desk_Rejected_Submission` (2020 on) invitations under the same prefix are fetched as well. Every call requests the `replies` detail, which returns all notes of the forum under the submission in a single call. ICLR 2016 hosted only a workshop track on OpenReview and is excluded. Table B.1 gives the counts.

Roles. A signature ending in `Authors` is an author. A signature matching `AnonReviewer`, `Reviewer_`, or "anonymous reviewer" is a reviewer. `Area_Chair`, `Senior_Area_Chairs`, and `Ethics_Reviewer` count as area chairs, while `Program_Chairs`, `pcs`, and `Conference` are program chairs. A profile ID (`~Name`) or `(anonymous)` is the public. In 2013 and 2014 reviews and replies shared an invitation, so a note is a review only if signed "anonymous reviewer". Because authors of 2013, 2014, and 2017 replied under their own names, a public signature whose normalized name matches an author of the submission is reassigned to the author role. A note's kind (review, comment, meta-review, decision, withdrawal, desk rejection) comes from the last segment of its invitation name.

Decisions. The venue label (e.g., "ICLR 2024 poster", "Submitted to ICLR 2024", "ICLR 2024 Conference Withdrawn Submission") gives the decision where present. Otherwise the decision note (2017, 2018, 2020 on) or the area chair's meta-review recommendation (2019) is used. "Invite to Workshop Track" (2017 and 2018) is a main-track rejection. In 2013, papers accepted to the conference track count as accepted and the workshop track does not. ICLR 2014 recorded no decisions on OpenReview. Our accepted counts agree with the announced acceptance numbers to within a few papers for 2019 to 2025.

Ratings. The rating is the leading number of the rating or recommendation field. Table B.2 lists every scale with the meaning attached to each value in the review form. The 2020 form used a

scale whose values differ from the usual 1-to-10 scale. For every score-based measure its values are therefore mapped to the usual value with the same meaning, as given in the last column of the table. Reject is matched to clear rejection (3), which is the middle of the three rejection values below the borderline, because the 2020 form had a single plain rejection value. Weak reject is matched to marginally below the threshold. A weak reject in 2020 becomes the usual 5, which is marginally below the acceptance threshold, and a weak accept becomes 6. Reviews are positive if the mapped rating is at least 6, and the borderline band is a mean mapped rating of at least 5 and below 6. Confidence is on a 1-to-5 scale and is absent in 2020, whose form asked for an experience assessment with three verbal levels instead.

Table B.2: Rating scales of the review forms by year, with the meaning attached to each value and the value on the usual 1-to-10 scale used for every score-based measure. The 2025 form stored bare numbers, and its meanings, which are taken from the form definition, are those of 2024. The 2026 scale is not listed, because no score-based measure uses 2026.

Years	Value	Meaning in the review form	Mapped
2017 to 2019, 2021	1	Trivial or wrong	1
	2	Strong rejection	2
	3	Clear rejection	3
	4	Ok but not good enough, rejection	4
	5	Marginally below acceptance threshold	5
	6	Marginally above acceptance threshold	6
	7	Good paper, accept	7
	8	Top 50% of accepted papers, clear accept	8
	9	Top 15% of accepted papers, strong accept	9
	10	Top 5% of accepted papers, seminal paper	10
2020	1	Reject	3
	3	Weak reject	5
	6	Weak accept	6
	8	Accept	8
2022 to 2025	1	Strong reject	1
	3	Reject, not good enough	3
	5	Marginally below the acceptance threshold	5
	6	Marginally above the acceptance threshold	6
	8	Accept, good paper	8
	10	Strong accept, should be highlighted at the conference	10

Author keys. Each submission carries an identifier per author, namely an email address in 2017 to 2020 and an OpenReview profile ID from 2021 on. A few 2017 to 2020 submissions carry profile IDs and a few later ones carry only emails. Within a year we identify an author by that identifier, in lower case, and by the normalized name (letters and spaces only) where the identifier is missing. The same person can appear under two email addresses or two profile IDs. This splits a person into two author keys and lowers papers per author, so the ID-based counts are an upper bound on distinct people. Across years an author is keyed by profile ID where present. A slot without an ID takes the single profile ID that carries its name when there is exactly one, the email address otherwise, and the name only when the slot carries neither. Before 2021 about 55% of author slots are bridged to an ID by their name and about 43% are keyed by email, while under 2% are keyed by name in any year. From 2022 on over 94% of slots carry an ID. The share of new and returning authors uses this key. A common name merges people, while a name spelled two ways splits them. Both ways of identifying authors give exponents within 0.02 of each other but different levels of concentration, because a common name merges several people. On names, the count of distinct authors grows from 1,417 to 56,844 with an exponent of 1.01, and papers per author grow from 1.28 to 1.98 with an exponent of 0.12. The tail exponent decreases from 3.46 to 2.54, and the share of author slots held by authors on five or more submissions rises from 6.6% to 31.7%.

C MEASUREMENT DETAILS

Sentiment. Each note is split into sentences at a period, question mark, or exclamation mark followed by white space. Sentences of more than three characters are scored with VADER (Hutto & Gilbert, 2014). VADER is a dictionary of about 7,500 words and emoticons that carry a positive or negative value rated by people, together with rules for negation (e.g., "not good"), intensifiers (e.g., "very"), contrastive conjunctions (e.g., "but"), capitalization, and punctuation. It returns a compound score in $[-1, 1]$ per sentence, and the note's sentiment is the mean over its sentences. Notes with fewer than five words are not scored. The compound score of a whole long text saturates near $+1$ or -1, so scoring each sentence and averaging avoids that. VADER was built for social media, and its absolute values on scientific text have not been validated. Every use in this paper is therefore a comparison between years, between ratings, or between a review and its follow-up, and we do not interpret any absolute level. The dictionary rates words such as thank, appreciate, and please as positive, so a note that opens with thanks scores higher for that reason alone. Every tone measure of reviews and reviewer follow-ups is therefore also computed with the 30 dictionary entries that start with thank, appreciat, please, kindly, or grateful removed (Table C.1). These politeness dictionary entries are a different list from the politeness markers counted per 1,000 words below.

Table C.1: Tone with and without the politeness entries of the dictionary. Follow-up is the tone of reviewer follow-ups after a negative review, as in Figure 4d, and Review is the tone of those reviews. The last three columns are the assessment-field correlation and the fitted tone at ratings 3 and 8 without the politeness entries.

Year	Follow-up	Follow-up, no polite	Review	Review, no polite	Corr. no polite	Tone at 3	Tone at 8
2017	0.111	0.091	0.100	0.098	0.38	0.062	0.237
2018	0.133	0.100	0.082	0.079	0.35	0.047	0.189
2019	0.149	0.108	0.078	0.074	0.37	0.036	0.193
2020	0.150	0.096	0.075	0.071	0.33	0.035	0.187
2021	0.152	0.102	0.088	0.083	0.34	0.049	0.187
2022	0.167	0.109	0.084	0.080	0.32	0.059	0.172
2023	0.154	0.094	0.089	0.086	0.29	0.062	0.186
2024	0.173	0.096	0.093	0.087	0.31	0.076	0.206
2025	0.171	0.099	0.107	0.102	0.29	0.090	0.210

Review forms. The fields of the review form changed as follows. From 2017 to 2021 a review had a single text field. In 2022 it had a paper summary, a main review, and a short summary of the review, which restated the verdict. In 2023 it had a paper summary, a strengths-and-weaknesses field, a field on clarity, quality, novelty, and reproducibility, and the summary of the review. In 2024 and 2025 it had a paper summary, separate strengths, weaknesses, and questions fields, and numeric scores for soundness, presentation, and contribution. The assessment fields used for the comparable tone series are the single field for 2017 to 2021, the main review for 2022, the strengths-and-weaknesses field for 2023, and the strengths and weaknesses fields for 2024 and 2025. The 2020 ratings are mapped to the usual scale before any fit.

Tone-score fit. The tone-score line $s_i \approx \beta_0 + \beta_1 r_i$ is fitted by ordinary least squares within each year on all reviews with a numeric rating, once on the full text and once on the assessment fields. The correlation reported in the text is the Pearson correlation between s_i and r_i. For example, if reviewers give lower scores for the same wording, the points shift along the rating axis and the line through them can stay the same, whereas softening negative reviews while positive ones are unchanged raises the tone at rating 3 and not at 8, so that $\beta_0 + 3\beta_1$ increases and β_1 decreases. For author replies, the same fit and correlation use the rating of the review that the reply answers directly (the reply's parent note). The fitted tone at a rating compares reviews at a fixed value across scale changes, and rating 3 covers a wider band from 2022 on, when the value 4 was removed. Table D.4 therefore also lists the mean tone of the lowest- and highest-rated fifth of each year's reviews, which depends only on the rank of a rating within its year. The lowest fifth decreased from 0.090 in 2017 to 0.056 in 2019 and rose back to 0.090 in 2025, and the highest fifth decreased from 0.222 to 0.183. The assessment fields of 2024 and 2025 concatenate the strengths and weaknesses fields, so a required strengths field adds positive sentences to every review of those years. The single field of 2017 to 2021 required none. The rise of the lowest fifth from 2023 to 2025 coincides with that form and is not separated from it.

Text-to-score model. The assessment text of every review with a numeric rating is turned into counts of single words and word pairs that occur in at least 10 reviews. The counts are weighted by term frequency times inverse document frequency (TF-IDF), which weights down words that occur in most reviews. The vocabulary holds at most 100,000 features with a logarithmic term frequency and is built with scikit-learn on all years, including the held-out year, without using the ratings. The only information from a held-out year that enters the model is therefore document frequencies. A ridge regression of the mapped rating on these features is fitted on a training set and evaluated on every year outside it. The ridge penalty, which controls how strongly the weights are shrunk towards zero, is chosen among 0.1, 0.3, 1, 3, 10, and 30 on a held-out fifth of the training set. Three training sets are used. The first is the single-field reviews of 2017 to 2021, evaluated on 2022 to 2025 and on each year of the period by leaving that year out. The second is the reviews of 2024, and the third is all years except the held-out year. For each held-out year we report the correlation between predicted and actual rating and R^2 over every review of that year while the training set stays fixed (Table C.2). The rating scale had 10 values in 2017 to 2019 and 2021, 4 in 2020, and 6 from 2022 on (Table B.2). The largest step of the decline in correlation, from 0.65 to 0.60, occurs at the 2022 change. A target with fewer distinct values caps a correlation on its own. The correlation is therefore accompanied by two measures that depend on the rating only through its order. These are the Spearman correlation and the probability that the model ranks a review rated 6 or above over a review rated below 6. The whole procedure is also rerun with every year's rating rounded to the nearest value of the 6-value grid used from 2022 on. Table C.3 lists all three for the model trained on 2017 to 2021. Rounded to the grid, the correlation is 0.62 to 0.65 within the training period and 0.59, 0.54, 0.55, and 0.52 in 2022 to 2025. The step at 2022 thus shrinks from 0.05 to about 0.03, and the decline over the decade remains. Within the period the ranking probability lies between 0.84 and 0.87 and the Spearman correlation, i.e., the correlation computed on ranks, between 0.65 and 0.69, and in 2025 they are 0.77 and 0.53. The single review field of 2017 to 2021 holds the verdict sentence of the review, which the assessment fields of 2022 on exclude. The evaluated text thus loses its verdict wording at 2022, where the largest step of every training set occurs. The model trained on 2024, which never saw verdict sentences, also predicts 2017 to 2021 better than 2022 to 2025. The whole procedure is therefore rerun with every sentence that matches recommend, accept, reject, score, rating, or threshold removed from the text of every year (Table C.4). For the model trained on 2017 to 2021 the correlation is then 0.59 to 0.66 within the training period and 0.59, 0.55, 0.55, and 0.53 in 2022 to 2025, so the step at 2022 shrinks from 0.05 to about 0.03 and the decline over the decade remains. The model trained on 2024 still predicts 2017 to 2021 better on average than the later years, so the verdict sentences explain part of the step and not the decline. The form of 2024 and 2025 also separates strengths from weaknesses, which places positive and negative words in every review once the fields are concatenated. The model trained on 2024, i.e., on that segmented form, still predicts the single-field years better than 2022 to 2025 (Table C.2), so the segmentation does not produce the decline either. The model is linear in word and word-pair frequencies, so a signal carried by the order or the structure of a review is outside what it can measure.

Table C.2: The text-to-score model. Correlation between predicted and actual rating and R^2 per held-out year, for the model trained on 2017 to 2021 (each year of that period evaluated with that year left out), on 2024, and on all other years. The model trained on 2024 has no value for its own year.

	Trained 2017 to 2021		Trained 2024		All other years	
Year	Corr.	R^2	Corr.	R^2	Corr.	R^2
2017	0.68	0.43	0.62	0.23	0.69	0.44
2018	0.68	0.45	0.61	0.26	0.69	0.46
2019	0.65	0.42	0.61	0.25	0.68	0.45
2020	0.63	0.40	0.60	0.23	0.66	0.41
2021	0.65	0.41	0.62	0.30	0.68	0.46
2022	0.60	0.34	0.58	0.33	0.63	0.39
2023	0.55	0.29	0.55	0.30	0.59	0.35
2024	0.56	0.28	n/a	n/a	0.60	0.36
2025	0.53	0.26	0.54	0.29	0.57	0.33

Table C.3: The text-to-score model trained on 2017 to 2021 under measures that do not depend on the spacing of the rating scale. Levels is the number of distinct rating values in the year, Spearman is the correlation computed on ranks, and rank is the probability that the model ranks a review rated 6 or above over a review rated below 6. The last three columns rerun the model with every year's rating rounded to the 6-value grid of 2022 on.

Year	Levels	Corr.	Spearman	Rank	Corr. on grid	Spearman on grid	Rank on grid
2017	10	0.68	0.69	0.87	0.65	0.68	0.86
2018	10	0.68	0.69	0.86	0.64	0.67	0.86
2019	10	0.65	0.66	0.84	0.62	0.63	0.83
2020	4	0.63	0.67	0.86	0.62	0.66	0.86
2021	10	0.65	0.65	0.84	0.62	0.63	0.84
2022	6	0.60	0.60	0.80	0.59	0.59	0.80
2023	6	0.55	0.56	0.79	0.54	0.55	0.78
2024	6	0.56	0.57	0.79	0.55	0.56	0.79
2025	6	0.53	0.53	0.77	0.52	0.52	0.76

Table C.4: The text-to-score model with every sentence that names the verdict, the score, or the threshold removed from the text of every year. Correlation between predicted and actual rating per held-out year, next to the correlation of the model trained on 2017 to 2021 with those sentences kept.

Year	With verdict sentences	Trained 2017 to 2021	Trained 2024	All other years
2017	0.68	0.66	0.61	0.67
2018	0.68	0.65	0.59	0.67
2019	0.65	0.63	0.59	0.66
2020	0.63	0.59	0.57	0.62
2021	0.65	0.62	0.59	0.65
2022	0.60	0.59	0.58	0.62
2023	0.55	0.55	0.55	0.59
2024	0.56	0.55	n/a	0.60
2025	0.53	0.53	0.53	0.57

Mood swing. A reviewer's comments are matched to the review with the same signature in the same forum. Only comments posted after the first author reply to that review count as follow-ups. The swing is the tone of the first follow-up minus the tone of the review. The main text reports it after negative reviews (rating below 6) alongside the tone of those reviews. The share of short follow-ups uses the same first follow-ups. Counting every author note in the courtesy measure instead of the first reply would show a decrease that comes only from the later parts of split rebuttals, which do not begin with thanks. The courtesy measure of Figure 4c instead takes the reviewer's first comment matched to the review by signature, whether or not an author reply precedes it. A reviewer who answers before the rebuttal is thus counted. The two sets of first comments give thanks shares that differ by at most 0.08 in a year.

Reply similarity. For each forum with author replies to at least two distinct reviews, the first reply to each review is reduced to its set of word tokens. The similarity of two replies is the number of words that they share divided by the number of distinct words that appear in at least one of the two, i.e., the Jaccard index. The per-forum value is the average over all pairs of replies. The per-year value is the mean over forums.

Lexicons. Hedges are might, may, perhaps, somewhat, seems, seem, unclear, possibly, arguably, appears, likely, suggest, suggests, could. Politeness markers are thank, thanks, appreciate, appreciated, kindly, please, grateful, respectfully, sincerely. Marker words are delve, delves, delving, commendable, meticulous, meticulously, intricate, intricacies, pivotal, underscore, underscores, underscoring, showcase, showcases, showcasing, realm, tapestry, noteworthy, nuanced, holistic, elucidate. Rates are per 1,000 word tokens. A note opens with thanks if its first word, after optional "we" or "I" and an optional adverb, is a form of thank or appreciate. Bullet lines are lines that start with a dash, an asterisk, a bullet, or a number followed by a period or parenthesis.

Score announcements and revision histories. OpenReview keeps every version of a review, but most versions are not publicly accessible. For example, in a 2025 paper with five reviews, the edit history was public for only two, and the public versions are those edited after the reviews were released. The score before the discussion therefore cannot be recovered for most reviews. We therefore measure score changes through what reviewers announce in comments. Announcements

undercount actual changes, as the comparison with external before-and-after scores in Appendix E shows. A reviewer comment announces a raised score if it matches "(raise, increase, bump, improve) my/the (score, rating, recommendation)" or "score/rating has been (raised, increased)" and contains no conditional phrase such as would, will, if, or happy to before the raise. A comment with such a phrase counts as a promised raise, and a match of update, revise, change, or adjust with my/the score counts as an unspecified change. A lowered score uses the raise pattern with lower, decrease, reduce, or downgrade, and a kept score uses keep, maintain, stick to, stand by, or remain followed by score, rating, recommendation, or assessment.

Reviewer concerns. Table C.5 lists the patterns, applied to the lower-cased review text. A review counts once per topic, and a match is a mention and not necessarily a complaint. In the patterns, `\b` marks a word boundary, `?` makes the preceding character optional, and `|` separates alternatives. The 2023 review form contained a field named "clarity, quality, novelty and reproducibility", which inflates the reproducibility and code-release concerns in that year. Those two concerns, matched by `\breproducib` and by a pattern for releasing, providing, or sharing code, are therefore not shown in the figure.

Table C.5: Patterns for reviewer concerns, as regular expressions over lower-cased review text.

Concern	Pattern
ablation	`\bablation`
baselines	`\bbaseline`
benchmarks	`\bbenchmark`
compute cost	`\bcomputational(ly)? (cost|expensive|overhead)|\bcompute (cost| budget)|\btraining (cost|time)|\binference (cost|time|speed)|\bgpu hours?\b|\bflops\b`
real-world	`\breal[- ]world\b`
limitations	`\blimitations?\b`
clarity / typos	`\btypos?\b|\bhard to (follow|read|understand)|\bpoorly written| \bwriting (is|could|needs|should)`
novelty / incremental	`\bincremental\b|\blimited novelty|\bnovelty is (limited|marginal| unclear|low)|\bnot (very |particularly )?novel\b`
theory / proof	`\bproof\b|\btheorem\b|theoretical (analysis|guarantee|justification)`
GPT / LLM mention	`\bchatgpt\b|\bgpt-?[3-5]|\bllms?\b|\blarge language model`
hallucination	`\bhallucinat`
data contamination	`\bcontaminat|\bdata leakage|\btest set leakage`
closed models	`\bclosed[- ]source|\bproprietary (model|api)|\bapi[- ]based`
LLM as judge	`\bllm[- ]as[- ]a?[- ]?judge|\bgpt-4 (as|to) (a )?(judge|evaluat)`

Public comment content. A public comment concerns citation if it matches cite, citation, cited, citing, prior work, our paper, our work, our method, missing reference, missing baseline, missing related, related to, relation to, or arxiv. Several of these are broad, so the share is indicative only.

The clock fit. Let $p_y(h)$ be the share of notes of a role posted in UTC hour h in year y. A daily pattern of activity $\tau(\ell)$ over local hour ℓ, which is specified manually, is low at night and high during the working day and evening. People whose local time is o hours ahead of UTC produce the profile $\tau(h+o)$, with hours taken modulo 24. We solve

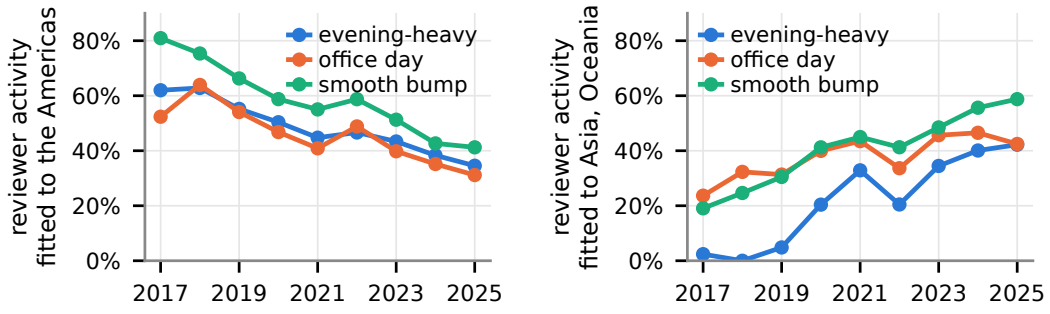
$$\min_{w \geq 0} \sum_{h=0}^{23} \Big(p_y(h) - \sum_{o=-11}^{12} w_o \, \tau(h+o) \Big)^2 \tag{C.1}$$

by non-negative least squares, where w_o is the share of activity attributed to people at offset o. The offsets are then summed for the Americas, for Europe and Africa, and for Asia and Oceania. The fit has 24 hourly observations and 24 unknown shares constrained to be non-negative. The evening-heavy and office-day patterns leave a residual of 1% to 14% of the norm of the profile. The smooth bump, whose 24 shifts span only three dimensions, leaves 8% to 18%. Patterns of different shapes thus fit the profile about equally well and give different shares, so the fit cannot identify the pattern, and the weights are only indicative. The main text uses the shares of the three thirds of the UTC day, which need no model, and Table C.6 lists them by role together with the total variation distance of each year's hourly profile from that of 2017, i.e., half the sum over the 24 hours of the absolute differences in share. The last third, 16:00 to 24:00 UTC, held 49% of reviewer notes in 2017 and 29% in 2025. It decreased in the same way with the busiest day excluded (45% to 32%) and on reviews alone (52% to 28%), so neither the deadline burst nor the shift from reviews to comments produces the shift. The distance from the 2017 profile grew to 0.24 for reviewer notes

and to 0.16 for author notes. Figure C.1 shows the fitted shares under the evening-heavy pattern and two alternatives.

Table C.6: Shares of reviewer and author notes posted in each third of the UTC day, and the total variation distance of the year's hourly profile from that of 2017.

	Reviewer notes				Author notes			
Year	00 to 08	08 to 16	16 to 24	Distance	00 to 08	08 to 16	16 to 24	Distance
2017	27%	24%	49%	0.00	30%	25%	45%	0.00
2018	28%	24%	48%	0.10	26%	26%	47%	0.11
2019	29%	27%	44%	0.08	33%	27%	41%	0.08
2020	32%	30%	38%	0.14	29%	32%	39%	0.10
2021	33%	32%	35%	0.17	33%	32%	35%	0.12
2022	31%	31%	38%	0.14	31%	34%	36%	0.13
2023	34%	32%	34%	0.18	33%	34%	34%	0.14
2024	36%	34%	30%	0.23	32%	36%	32%	0.16
2025	34%	36%	29%	0.24	32%	36%	31%	0.16



(a) Americas share by daily pattern

(b) Asia and Oceania share by daily pattern

Figure C.1: Sensitivity of the clock mixture of (C.1) to the daily pattern (evening-heavy, office day, and smooth bump), for reviewer notes. The direction of change survives every pattern, and the level does not.

Daily activity pattern. The evening-heavy pattern $\tau(\ell)$ over local hour ℓ takes the value 0.6 at 00:00, as activity often continues until midnight, 0.08 from 01:00 to 06:00, 0.3 at 07:00, 0.6 at 08:00, 1.0 from 09:00 to 11:00, 0.8 at 12:00, 1.0 from 13:00 to 17:00, and 0.9 from 18:00 to 23:00, and is normalized to sum to one. The design matrix has a column for each UTC offset from -11 to $+12$, and each column is the pattern shifted by that offset. The fit of (C.1) uses the non-negative least-squares routine of SciPy. Offsets -11 to -3 are the Americas, -2 to $+3$ Europe and Africa, and $+4$ to $+12$ Asia and Oceania. Two alternative patterns show the sensitivity to the pattern. An office-day pattern takes the value 0.05 at night, 0.5 at 08:00, 1.0 from 09:00 to 17:00, 0.3 from 18:00 to 21:00, and 0.1 at 22:00 and 23:00. A smooth-bump pattern is $1.05 + \cos(2\pi(\ell - 15)/24)$, lowest at 03:00 and highest at 15:00. For example, under the evening-heavy pattern the share of reviewer activity fitted to Asia and Oceania in 2018 is 0%, under the office day it is 32%, and under the smooth bump it is 25%. The level thus depends on the pattern. The decrease of the Americas share and the increase of the Asia and Oceania share between 2017 and 2025 appear under all three patterns (Figure C.1).

Busiest day. For each year and role, the busiest day is the UTC calendar day with the most notes, which we take as the review deadline. Every share is a share of all reviews of the year, and Table C.7 lists them. A share of the reviews posted in a window that ends with the busiest day would exclude every review posted after it, which was between 12% and 56% of a year's reviews. It would mix the shape of the deadline with the count of late reviews, so no such window is used. The busiest-day share decreased from 30% to 22% with an exponent of -0.14 and a standard error of 0.05. The share within a day of the busiest day ranged from 44% to 62% with an exponent of -0.05 and a standard error of 0.04. Within a week of it, the share ranged from 86% to 93% with an exponent of -0.01 and a standard error of 0.01. The share posted after the busiest day decreased from 56% to 12%, with 2025 far below every earlier year, and the share posted before it rose from 14% to 66%. A less strictly enforced deadline would raise the share posted after the busiest day, so the

data do not support it, and reviews were finished earlier instead. ICLR sets its review deadline at a fixed local time, and a burst that ends near 12:00 UTC, for example, is split across two calendar days. The single-day share therefore moves with the clock hour of the deadline as well as with how early reviews are finished. Reviewers whose local evening lies in the next UTC day are also counted outside the busiest day. A pool spread over more time zones therefore lowers the single-day share by construction, and neither effect can move the share within a day, which is the share to compare across years. That share has an exponent within its fit standard error of zero, so time zones and earlier completion are not separated by the data.

Table C.7: Reviews around the busiest day of each year, as shares of all reviews of the year.

Year	Busiest day	Reviews	On the day	Within a day	Within a week	After the day
2017	2016-12-16	1,511	30%	54%	91%	56%
2018	2017-11-27	2,930	36%	62%	93%	36%
2019	2018-11-02	4,764	25%	48%	89%	46%
2020	2019-10-23	7,775	27%	52%	86%	42%
2021	2020-10-28	11,508	31%	61%	91%	36%
2022	2021-11-02	13,161	31%	59%	93%	32%
2023	2022-10-24	18,575	24%	55%	89%	50%
2024	2023-10-31	28,028	20%	52%	88%	42%
2025	2024-11-04	46,748	22%	44%	87%	12%

Scaling fits. Equation (1) is fitted by ordinary least squares on the natural logarithms over 2017 to 2026. The withdrawn share is fitted over 2018 to 2025, because the 2026 security incident may have added withdrawals. With 2026 included its exponent is 0.41 with a standard error of 0.08. The data record no withdrawals in 2017 and no withdrawal notes in 2018, and 60% of the withdrawn papers of 2018 carry reviews against at least 91% in every later year. The first points of the series therefore partly measure the completeness of the data. Fitted over 2020 to 2025 the exponent is 0.37 with a standard error of 0.15 and an R^2 of 0.60. Two exponents are compared by fitting the ratio of the two series against $\log N$. The slope of that fit equals the difference of the two exponents. Its standard error comes from the residuals of the ratio series, so it needs no assumption about how the two fits relate. The ratio series has 9 yearly points and the line has 2 fitted parameters, so the standard error rests on 7 residual degrees of freedom, and the difference for words per paper is 2.8 standard errors. The two fits share their years and their venue sizes, and their yearly deviations from the fitted lines move together. The correlation of the deviations is 0.46 for rebuttal against review words per paper and 0.88 for words per note. Treating the two fits as unrelated would therefore overstate the uncertainty. A delta-method error on the ratio of two separately fitted exponents makes that assumption and gives 1.7 ± 0.5 for the ratio of the two exponents, which is why the ratio is not used. Table C.8 lists the differences under each measurement choice. Forums with a public comment are counted among main-track submissions. Rebuttal and review words per paper are totals of word tokens over all notes of a role in main-track forums divided by the number of main-track submissions. This drops withdrawn papers from both the words and the count. The all-track companion divides the words of every track by every submission of the year, so a withdrawal removes nothing. Word tokens include text quoted from the note being answered. The companion with quoting removed therefore drops lines that open with the markdown quotation marker, fenced code, pipe tables, and mathematics between dollar signs before counting.

Table C.8: The effort comparison under each measurement choice. Exponents of rebuttal and review words with the difference from the fit of the ratio series, its standard error, and the correlation of the residuals of the two separate fits (Residual corr.). The difference comes from the ratio fit and can differ from the difference of the rounded exponents by rounding.

Measure	Rebuttal α	Review α	Difference	SE	Residual corr.
Words per paper, main track	0.40	0.24	0.17	0.06	0.46
Words per paper, quoting removed	0.36	0.23	0.12	0.06	0.50
Words per submission, all tracks	0.33	0.23	0.10	0.06	0.43
Words per submission, all tracks, quoting removed	0.29	0.23	0.05	0.06	0.46
Words per note	0.12	0.13	−0.01	0.02	0.88
Notes per paper	0.28	0.11	0.17	0.06	0.64

Tail fit. The discrete maximum likelihood estimate of Clauset et al. (2009) for the tail exponent solves

$$\frac{\zeta'(\hat{\gamma}, k_{\min})}{\zeta(\hat{\gamma}, k_{\min})} = -\frac{1}{n}\sum_{i=1}^{n} \ln k_i, \qquad \text{and } \zeta(\gamma, k_{\min}) = \sum_{k=k_{\min}}^{\infty} k^{-\gamma}, \tag{C.2}$$

where n is the number of authors in the tail, k_i is the count for the i-th of them, and ζ' is the derivative of ζ in γ. It is solved numerically with the Hurwitz zeta function of SciPy and is applied to all author keys with at least $k_{\min}$ submissions in the year. The number of authors in the tail grows from 226 in 2017 to 20,087 in 2026 for $k_{\min} = 2$. The closed-form approximation $\hat{\gamma} \approx 1 + n/\sum_i \ln(k_i/(k_{\min} - 1/2))$ of the same paper is biased downward at small $k_{\min}$. At $k_{\min} = 2$ it gives 3.1 in 2017 and 2.5 in 2026 against 3.6 and 2.7 for the exact estimate. At $k_{\min} = 3$ it gives 3.5 and 2.8 against 3.7 and 2.9. The exact estimate at $k_{\min} = 2$ decreases in every year except 2018 and 2020, as the approximation does. No goodness-of-fit comparison with other heavy-tailed forms was made, so the exponent is a descriptive summary of the heaviness of the tail. We use $k_{\min} = 2$ because it retains enough authors in the tail in the early years, and $k_{\min} = 3$ gives the same downward trend.

Vocabulary. Terms are unigrams and bigrams from titles and abstracts after stop-word removal, and a term's share in a year is the fraction of submissions that mention it. A term is topical if it appears as an author-supplied keyword on at least 15 submissions over all years. It bursts in year t if its share is at least 1% and at least 2.5 times its share two years earlier. Review bursts use the same rule on review text with a 2% share and a factor of 2. Table C.9 lists up to five terms with the largest ratio, after removing terms that come from the names of review-form fields or from mathematical symbols. A year with fewer entries had fewer terms that pass the rule, and phrases of the review process, such as "see weaknesses", are not field names and stay.

Table C.9: Terms that burst in reviews, with the share of reviews mentioning the term and the ratio to the share two years earlier, after removing terms that name review-form fields or mathematical symbols and counting a singular and plural twin once, keeping at most the five terms with the largest ratio, so a year with fewer entries had fewer terms that pass the rule. The share two years earlier has a floor of 0.05%, so a ratio equal to the share divided by that floor means that the term did not occur two years earlier. Terms are copied as they appear in the texts, so abbreviations such as GNN, OOD, ViT, and LoRA are the reviewers' own.

Year	Burst terms (share, ratio)
2019	attacks (4.8%, 97), bounds (3.9%, 79), uncertainty (3.6%, 73), environments (3.6%, 72), ablation study (3.5%, 71)
2020	BERT (3.4%, 67), NeurIPS (2.6%, 52), increase score (2.3%, 46), leaning (2.2%, 43), lean (2.0%, 40)
2021	reasons score (4.6%, 91), NeurIPS (3.5%, 69), BERT (3.1%, 62), during rebuttal (2.8%, 56), rebuttal period (2.7%, 55)
2022	contrastive (5.1%, 5.3), federated (3.0%, 5.2), offline (2.9%, 4.3), OOD (2.6%, 3.0)
2023	prompt (2.4%, 48), ViT (2.1%, 42)
2024	LLMs (10.2%, 205), LLaMA (2.4%, 48), in-context (2.2%, 44), see weaknesses (6.9%, 15.3), prompting (3.1%, 12.2)
2025	LLaMA (4.2%, 84), adaptability (2.6%, 52), would enhance (2.4%, 49), LLM-based (2.0%, 41), LLMs (18.5%, 20.2)

Topic groups. Primary areas exist from 2023, with 14 areas in 2023, 20 in 2024, and 21 in 2025 and 2026, and the 2025 taxonomy is the first with an area for LLMs. Table C.10 lists the 11 groups, each of which coarsens the 2025 areas. LLMs and foundation models is the area of that name, and vision, audio, and language applications is the applications area of those modalities. Safety, fairness, and interpretability joins the alignment area with the interpretability area, reinforcement learning and robotics joins reinforcement learning with the robotics area, and applications to sciences joins the physical sciences with neuroscience. Learning theory and optimization joins the two areas of those names, representation and transfer learning joins the representation-learning area with transfer and meta learning, and generative models, datasets and benchmarks, and graphs are single areas. The remaining areas, namely other topics, probabilistic methods, time series, causal reasoning, neu-

rosymbolic systems, and infrastructure, form the other group. The 2023 and 2024 areas are mapped to the same groups where a counterpart exists.

Two methods assign a group from the keywords, title, and abstract. The classifier turns the concatenated text into TF-IDF counts of words and word pairs that occur in at least three papers, with at most 200,000 features, and fits a multinomial logistic regression with a penalty weight of 5 and a fixed seed in scikit-learn. It is trained on the papers of 2025 and 2026, which share a taxonomy. The labels of 2025 are predicted by the model trained on 2026, those of 2026 by the model trained on 2025, and those of 2017 to 2024 by the model trained on both years, so no year is labeled by a model that saw its areas. The rules of Table C.11 are the comparison. A paper's score for a group is three times the number of matches of the group's pattern in its keywords, plus twice the matches in its title, plus the matches in its abstract. The paper takes the group with the largest score, ties go to the earlier rule, and no match gives the other group. Table C.10 gives precision, recall, and F1 per group against the authors' own area on the held-out papers of 2025 and 2026. The classifier agrees with the authors' area on 62% of the papers (64% for held-out 2025 and 62% for held-out 2026) with a mean F1 of 0.64, and the rules agree on 46% with a mean F1 of 0.45, so the paper uses the classifier. Against the 2024 areas mapped to the groups, the classifier agrees on 58% of the papers, and against the 2023 areas on 49%, where a catch-all area for deep learning holds a third of the papers. The 2024 taxonomy has no LLM area, so the 604 papers that the classifier places in the LLM group in 2024 count as disagreements there. The papers placed in the LLM group before 2020 are language-modeling papers, e.g., on recurrent and transformer language models, which are the predecessors of the group. The authors' choice of area is itself a judgment, e.g., a paper on the safety of LLMs fits two areas, so part of the disagreement is ambiguity between areas and not error.

Table C.10: The topic groups and the agreement of the two assignment methods with the primary area that the authors chose, on the held-out papers of 2025 and 2026. Papers is the number of papers whose area coarsens to the group, and the last row gives the accuracy and the mean F1 over groups.

		Classifier			Rules		
Group	Papers	Precision	Recall	F1	Precision	Recall	F1
LLMs and foundation models	3,957	0.50	0.60	0.54	0.35	0.64	0.45
Vision, audio, and language applications	4,081	0.57	0.56	0.56	0.39	0.58	0.46
Generative models	2,745	0.65	0.61	0.63	0.53	0.46	0.49
Safety, fairness, and interpretability	4,050	0.69	0.73	0.71	0.56	0.44	0.49
Datasets and benchmarks	2,351	0.77	0.77	0.77	0.38	0.43	0.41
Reinforcement learning and robotics	2,582	0.71	0.75	0.73	0.59	0.67	0.63
Applications to sciences	1,998	0.73	0.61	0.66	0.58	0.46	0.51
Learning theory and optimization	2,141	0.66	0.65	0.66	0.50	0.55	0.52
Graphs and geometry	963	0.72	0.64	0.68	0.50	0.76	0.61
Representation and transfer learning	2,979	0.56	0.52	0.54	0.63	0.24	0.34
Other	3,639	0.54	0.51	0.53	0.36	0.04	0.08
All groups (accuracy and mean F1)	31,486	0.62	0.62	0.64	0.46	0.46	0.45

Table C.11: The regular-expression rules, applied to lower-cased keywords, title, and abstract, in the order of precedence used for ties.

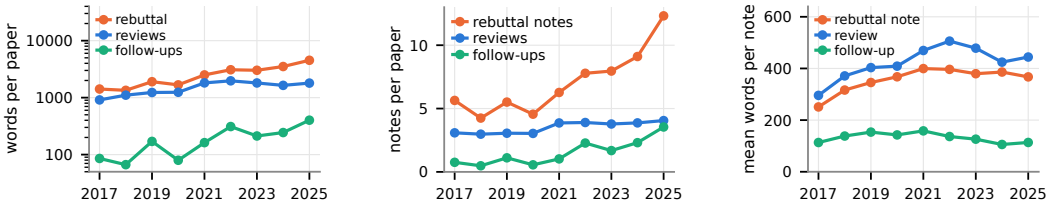
Group	Pattern
LLMs and foundation models	`large language model|\bllms?\b|foundation model|\bgpt|chatgpt|in-context learning|instruction[- ]tun|\brlhf\b|chain[- ]of[- ]thought|retrieval[- ]augmented|prompt(ing)?\b|language model`
Reinforcement learning and robotics	`reinforcement learning|\brl\b|policy gradient|q-learning|markov decision|\bbandits?\b|imitation learning|\bpolicy optimization|reward|robot|actor-critic|\bmdps?\b`
Graphs and geometry	`graph neural|\bgnns?\b|graph (learning|representation|convolution|transformer)|knowledge graph|node classification|link prediction|message passing|geometric deep learning|\bgraphs?\b`
Learning theory and optimization	`learning theory|generalization (bound|error|gap)|sample complexity|convergence|\bregret\b|online learning|\bpac\b|minimax|convex|stochastic gradient|\bsgd\b|optimization|gradient descent|neural tangent|kernel`
Generative models	`generative model|generative adversarial|\bgans?\b|variational autoencoder|\bvaes?\b|diffusion|flow matching|normalizing flow|score-based|text-to-image|image generation|video generation|autoregressive`
Safety, fairness, and interpretability	`interpretab|explainab|fairness|privacy|federated|adversarial (robustness|attack|example|training)|safety|alignment|unlearning|backdoor|watermark|jailbreak|\bbias\b|trustworth`
Datasets and benchmarks	`benchmark|dataset|evaluation`
Applications to sciences	`molecul|protein|\bdrug\b|chemi|physic|biolog|neuroscience|\bbrain\b|medical|clinical|climate|materials|genom|single-cell|health`
Vision, audio, and language applications	`computer vision|image|video|\b3d\b|point cloud|object detection|segmentation|vision transformer|convolutional|\bcnn\b|visual|speech|audio|natural language|\bnlp\b|machine translation|question answering|text (classification|generation)|\bbert\b|dialogue|summarization|multimodal|vision-language`
Representation and transfer learning	`self-supervised|contrastive|representation learning|unsupervised learning|transfer learning|meta-learning|meta learning|continual learning|few-shot|domain adaptation|lifelong|semi-supervised`

D FULL PER-YEAR TABLES AND FURTHER TRENDS

Per-year tables. Table D.1 gives the exponents of Table 1 with the R^2 of each fit against $\log N$ and against the calendar year and the years fitted. Table D.2 lists the rebuttal volume and withdrawal measures, Table D.3 the discussion measures, Table D.4 the tone and lexical measures, Table C.1 the tone measures without the politeness entries, and Table D.5 the score and clock measures per year. Figure D.1 shows the raw per-year series behind the effort exponents.

Table D.1: Exponents α from (1) with fit standard errors (SE), R^2 against $\log N$ and against the calendar year, the years fitted, and the first and last fitted values. Indented rows multiply to the row above.

Quantity	α	SE	R^2	R^2 year	Years	First	Last
Distinct authors	1.03	0.01	1.00	0.98	2017 to 2026	1,458	64,649
Authors per paper	0.13	0.01	0.94	0.96	2017 to 2026	3.7	5.7
Papers per author	0.10	0.01	0.94	0.95	2017 to 2026	1.24	1.74
Top 1% share of author slots	0.16	0.01	0.95	0.94	2017 to 2026	4.8%	8.8%
Withdrawn share of submissions	0.50	0.09	0.84	0.89	2018 to 2025	8.2%	25.6%
Review words per paper	0.24	0.06	0.72	0.74	2017 to 2025	912	1,800
Reviews per paper	0.11	0.03	0.67	0.78	2017 to 2025	3.1	4.1
Words per review	0.13	0.04	0.59	0.54	2017 to 2025	296	444
Rebuttal words per paper	0.40	0.06	0.86	0.94	2017 to 2025	1,416	4,525
Author notes per paper	0.28	0.08	0.66	0.79	2017 to 2025	5.6	12.3
Words per author note	0.12	0.03	0.64	0.54	2017 to 2025	251	367
Follow-up words per paper	0.52	0.14	0.65	0.74	2017 to 2025	86	402
Follow-up notes per paper	0.56	0.16	0.64	0.76	2017 to 2025	0.8	3.5
Words per follow-up	−0.04	0.05	0.08	0.16	2017 to 2025	113	113
Public-comment forums, share	−0.93	0.11	0.91	0.96	2017 to 2025	45%	4.1%
Public-comment forums, count	−0.03	0.12	0.01	0.05	2017 to 2025	221	355
Busiest-day share of reviews	−0.14	0.05	0.54	0.55	2017 to 2025	30%	22%
Bullet lines per 1,000 words	0.34	0.04	0.93	0.96	2017 to 2025	6.5	18.4



(a) Words per paper by side (b) Notes per paper by side (c) Words per note by side

Figure D.1: The raw per-year series behind the effort exponents, for reviews, author replies, and reviewer follow-ups. Discussion words per paper (a), on a logarithmic axis, equal notes per paper (b) times words per note (c).

Table D.2: Rebuttal volume, withdrawals, and public comment content per year. Review, rebuttal, and follow-up words are per main-track paper, and Ratio is rebuttal words over review words. Multi-part is the share of author replies whose title marks a part. Withdrawn is the share of submissions withdrawn, Reviewed the share of withdrawn papers that had reviews, and Rating gap the mean rating of withdrawn papers minus that of rejected papers. Citation is the share of public comments that mention citations or prior work.

Year	Review words	Rebuttal words	Follow-up words	Ratio	Multi-part	Withdrawn	Reviewed	Rating gap	Citation
2017	912	1,416	86	1.55	0.4%	0.0%	n/a	n/a	26%
2018	1,105	1,344	67	1.22	6.2%	8.2%	60%	−0.67	25%
2019	1,231	1,904	170	1.55	13.7%	10.1%	91%	−0.63	26%
2020	1,242	1,675	80	1.35	14.1%	14.2%	94%	−0.53	29%
2021	1,814	2,505	162	1.38	17.9%	13.4%	95%	−0.63	39%
2022	1,972	3,086	311	1.56	24.7%	22.8%	97%	−0.62	40%
2023	1,809	3,022	212	1.67	28.0%	23.1%	98%	−0.57	36%
2024	1,642	3,516	244	2.14	23.8%	22.4%	93%	−0.72	34%
2025	1,800	4,525	402	2.51	21.8%	25.6%	97%	−0.73	40%

Table D.3: Discussion measurements per year. Reply is the share of reviews with an author reply, overall and for negative and positive reviews, and Follow-up is the share of replied-to reviews with a reviewer follow-up. Short is the share of first reviewer follow-ups under 30 words. Thanks is the share of first author replies to a review (auth.) and of first reviewer comments after their review (rev.) that open with thanks. Swing is the follow-up tone minus the review tone, and Public is the share of forums with a public comment. Borderline acceptance is given with and without a reviewer follow-up.

Year	Reply	Reply neg.	Reply pos.	Follow-up	Short	Thanks auth.	Thanks rev.	Swing	Public	Border. w/	Border. w/o
2017	0.63	0.55	0.68	0.17	0.23	0.71	0.30	0.07	0.45	0.13	0.07
2018	0.73	0.62	0.86	0.14	0.17	0.65	0.38	0.12	0.36	0.21	0.13
2019	0.79	0.66	0.93	0.29	0.16	0.67	0.47	0.09	0.32	0.13	0.07
2020	0.80	0.70	0.94	0.15	0.16	0.72	0.59	0.10	0.19	0.12	0.10
2021	0.83	0.73	0.94	0.21	0.13	0.69	0.60	0.10	0.10	0.12	0.09
2022	0.75	0.63	0.91	0.48	0.19	0.66	0.61	0.11	0.10	0.22	0.16
2023	0.74	0.61	0.90	0.37	0.23	0.62	0.61	0.10	0.05	0.20	0.19
2024	0.80	0.68	0.95	0.50	0.30	0.64	0.71	0.12	0.04	0.24	0.18
2025	0.73	0.59	0.91	0.70	0.30	0.64	0.74	0.11	0.04	0.22	0.18

Table D.4: Tone measurements per year. Corr. review is the correlation of tone with rating on all fields of the review, and Corr. reply the correlation for author replies with the rating of the review answered. Tone at 3 and 8 are fitted values of the tone-score line on the assessment fields, as in Figure 6c, and Low fifth and High fifth are the mean assessment-field tone of the lowest- and highest-rated fifth of the year's reviews. Marker is the share of texts with an LLM marker word. Bullets and Hedges are bullet lines and hedges per 1,000 words of reviews. Excl. is exclamation marks per 1,000 words over all author replies, which rose again after 2021, and Excl. first the same rate over first replies to a review, which decreased from 3.1 to 1.3 in 2018 and stayed between 0.7 and 1.2.

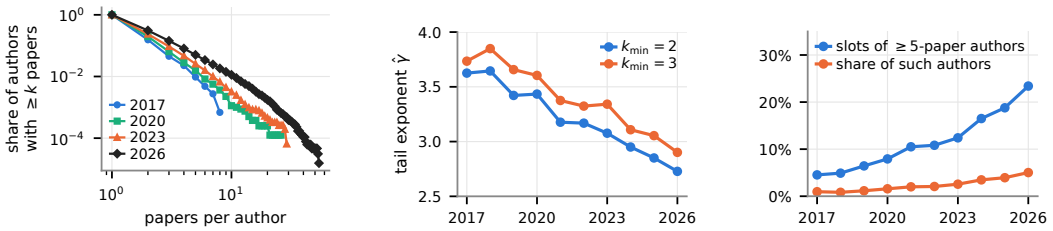
Year	Corr. review	Corr. reply	Tone at 3	Tone at 8	Low fifth	High fifth	Marker review	Marker reply	Bullets	Hedges	Excl.	Excl. first
2017	0.39	0.21	0.064	0.242	0.090	0.222	0.014	0.015	6.5	8.9	4.3	3.1
2018	0.36	0.12	0.050	0.194	0.064	0.176	0.015	0.014	8.1	8.4	2.0	1.3
2019	0.38	0.15	0.039	0.197	0.056	0.180	0.018	0.014	9.2	8.1	1.9	1.0
2020	0.34	0.15	0.038	0.193	0.075	0.156	0.019	0.017	8.6	7.7	1.6	1.2
2021	0.34	0.14	0.054	0.193	0.074	0.176	0.026	0.021	11.5	7.7	1.5	1.0
2022	0.38	0.12	0.063	0.177	0.060	0.152	0.028	0.025	12.2	7.4	2.2	1.0
2023	0.41	0.10	0.065	0.189	0.064	0.162	0.027	0.021	13.7	7.7	2.4	1.0
2024	0.27	0.06	0.079	0.211	0.077	0.180	0.111	0.110	17.3	7.6	2.5	1.1
2025	0.25	0.05	0.093	0.214	0.090	0.183	0.098	0.075	18.4	8.3	2.8	0.7

Table D.5: Scores and decisions per year, on the mapped 1-to-10 scale. Papers is the number of reviewed main-track papers. Std. is the mean within-paper standard deviation of ratings, which jumps at the 2022 scale change and is not comparable across it. Split is the share of papers with reviews on both sides of the acceptance threshold, over all papers and among papers with exactly three and exactly four reviews. The column ≥ 4 rev. is the share of papers with at least four reviews, and Thr. is the mean rating above which more than half of the papers were accepted. Rej. ≥ 6 is the share of rejected papers with mean rating at least 6.

Year	Papers	Std.	Split	Split, 3 rev.	Split, 4 rev.	≥ 4 rev.	Confidence	Thr.	Rej. ≥ 6
2017	490	0.87	0.41	0.40	0.60	9%	3.82	6.07	0.16
2018	922	1.01	0.49	0.49	n/a	2%	3.80	5.82	0.10
2019	1,419	0.94	0.47	0.46	0.62	5%	3.76	5.91	0.10
2020	2,213	0.87	0.51	0.50	0.72	7%	n/a	5.88	0.13
2021	2,594	1.01	0.62	0.50	0.64	80%	3.71	5.93	0.12
2022	2,617	1.18	0.58	0.47	0.60	79%	3.67	5.75	0.10
2023	3,792	1.22	0.58	0.47	0.61	69%	3.58	5.74	0.07
2024	5,689	1.20	0.61	0.51	0.63	75%	3.60	5.68	0.10
2025	8,609	1.17	0.61	0.48	0.62	85%	3.62	5.68	0.09

Templated replies. Author replies did not become more similar. The mean word overlap between a team's first replies to different reviewers stayed between 0.19 and 0.23 in every year. The similarity measure is defined in Appendix C.

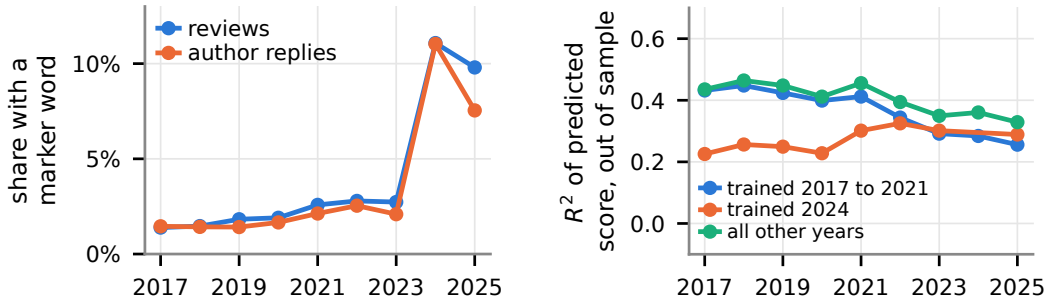
The tail of papers per author. The distribution of papers per author within a year grew heavier at the top in every year except 2018 and 2020 (Figure D.2a). The tail exponent decreased from 3.6 to 2.7 (Figure D.2b). The share of author slots held by authors with five or more submissions rose from 4.5% to 23% (Figure D.2c).



(a) Papers per author in a year (b) Tail exponent (c) Prolific-author share

Figure D.2: Papers per author within a year, with authors keyed by profile ID or email. Subfigure (a) shows the share of authors with at least k papers, on logarithmic axes, for four years. Subfigure (b) gives the tail exponent of (C.2) for $k_{\min}$ of 2 and 3, and subfigure (c) the share of author slots held by authors with five or more submissions and the share of such authors.

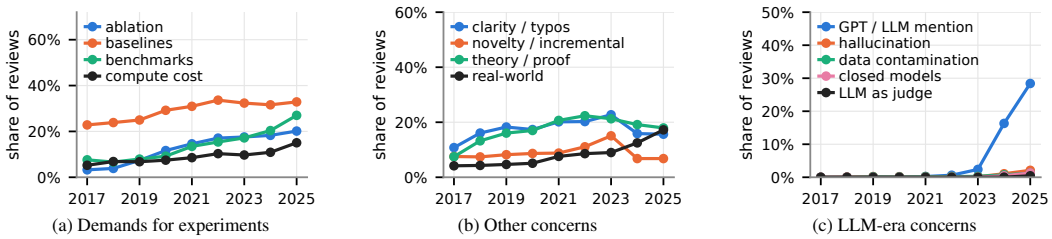
Marker words and the text-to-score model. A marker word appears in 1.4% of reviews in 2017, 11% in 2024, and 10% in 2025. In author replies the share rose from 1.5% to 11% and then decreased to 7.5% (Figure D.3a). The R^2 of the text-to-score model on years not used for training decreases from about 0.43 within 2017 to 2021 to 0.26 in 2025 for the model trained on 2017 to 2021. With every other year used for training it decreases from 0.44 to 0.33 (Figure D.3b and Table C.2).



(a) Marker word share (b) Text-to-score predictor, R^2

Figure D.3: Subfigure (a) is the share of reviews and author replies with a marker word, and subfigure (b) the R^2 of the text-to-score model on years not used for training, by training set.

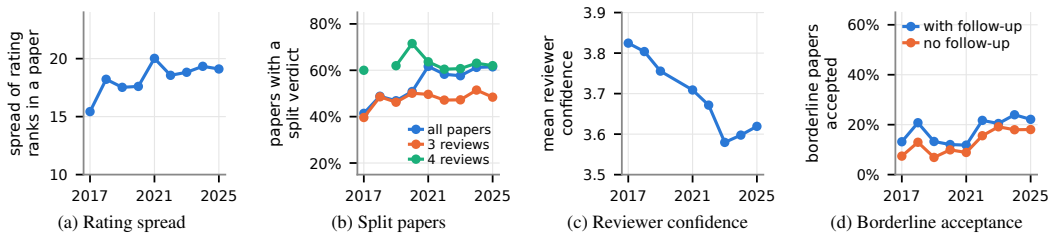
Reviews mention experiments more and clarity less, and a form change is a likely explanation. Mentions of clarity and typos doubled up to 2023 and decreased by a third in 2024 (Figure D.4b). That year the form added separate numeric scores for soundness, presentation, and contribution. Over the same period mentions of ablations rose 6× and mentions of benchmarks and compute cost tripled (Figure D.4a). *We conjecture two causes for this. Submissions may now be edited by the same language-model tools that edit reviews, so that presentation faults became rarer. The 2024 form also moved presentation faults from prose into a number.* The paper text was not analyzed, so the first cause is not checked, while the timing supports the second. Topics that did not exist before 2023, namely hallucination, data contamination, closed models, and evaluation by a language model, now appear in 0.5% to 2% of reviews each (Figure D.4c). Mentions of GPT or language models rose from almost none to over a quarter of reviews in 2025. From 2024 the phrase "see weaknesses" appears in about 7% of reviews, because the new questions field is often answered with it. For authors the review has become a request for evidence more than a critique of presentation.



(a) Demands for experiments (b) Other concerns (c) LLM-era concerns

Figure D.4: Share of reviews that mention a topic, by year, for experimental demands, other recurring topics, and topics that arrived with language models. Each series is the share of reviews whose text matches the pattern for that topic, so the counts are mentions and not complaints. Topics that the text does not discuss are shown for completeness.

Scores and confidence. Reviewers disagree more than they did, and rate their own confidence slightly lower. The rating scale changed in 2020 and 2022, so we compare years on the rank of each rating within its year, i.e., the share of that year's reviews rated lower. The spread of ranks within a paper, i.e., the standard deviation of the percentile ranks of a paper's ratings, rose from 15 to 19 percentile points. Almost all of that rise came by 2021, and the spread has been flat since (Figure D.5a). Table D.5 lists the standard deviation of the raw ratings instead. A split paper has a review at or above the acceptance threshold of 6 and a review below it. The share of split papers rose from 41% to 61% with a jump in 2021 and no trend after (Figure D.5b). The jump is the change of the panel size, because the chance of a split rises with the number of reviews. The share of papers with at least four reviews went from under a tenth to four fifths in 2021 (Table D.5). Among papers with exactly three reviews the split share rose from 40% to 50% by 2020 and stayed between 47% and 52% after. Papers with exactly four reviews had a split share between 60% and 72% in every year with at least 30 of them. Mean reviewer confidence, on a scale from 1 to 5, decreased from 3.82 to 3.58 by 2023 and was flat after (Figure D.5c). *We conjecture that a reviewer pool growing with submissions is drawn increasingly from newer members, while papers come from ever more prolific groups, so the gap in experience widens.* OpenReview holds no reviewer seniority data, so this is not checked.



(a) Rating spread (b) Split papers (c) Reviewer confidence (d) Borderline acceptance

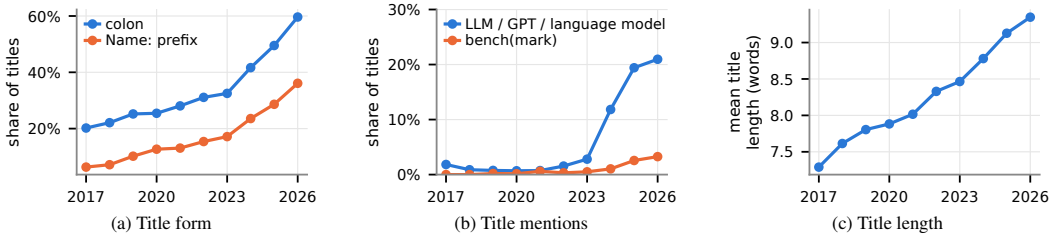
Figure D.5: Scores and decisions. Subfigure (a) is the spread of rating ranks within a paper, subfigure (b) the share of papers with reviews on both sides of the threshold, subfigure (c) mean confidence, with no 2020 value because that form had no confidence field, and subfigure (d) borderline acceptance by reviewer follow-up.

Engagement and borderline outcomes. Among papers whose mean rating lies in the borderline band of their year's scale, those with a reviewer follow-up were accepted more often than those without in every year (Table D.3 and Figure D.5d). The direction agrees with Kargaran et al. (2025), Louis et al. (2026), and Jung et al. (2025). Our data do not separate whether a follow-up raises the chance of acceptance or whether reviewers follow up on papers that they already favor.

Topics and titles. Vocabulary bursts in titles and abstracts trace the topics of each year (Table D.6). Titles also changed shape (Figure D.6). The share of titles with a colon tripled to three fifths, which is far above the long-run rate reported for science journals (Lewison & Hartley, 2005), and the share of titles of the form "Name: description" rose 6×. Titles and abstracts grew longer, and a fifth of 2026 titles mention LLMs or language models.

Table D.6: Terms that burst in titles and abstracts in each year, with the share of submissions mentioning the term and the ratio of that share to the share two years earlier. Only terms used as author keywords on at least 15 submissions over all years are kept, a singular and plural twin count once, and the 6 terms with the largest ratio are listed. The share two years earlier has a floor of 0.01%, so a ratio equal to the share divided by that floor means that the term did not occur two years earlier. Terms are copied as they appear in the texts, so abbreviations such as GNN, CLIP, RAG, and MLLM are the authors' own.

Year	Burst terms (share, ratio)
2019	adversarial attacks (4.2%, 424), defense (1.8%, 184), Wasserstein (1.4%, 139), transport (1.3%, 127), adversarial robustness (1.2%, 120), maximization (1.2%, 120)
2020	BERT (2.3%, 227), self-supervised (2.0%, 18.5), edge (1.6%, 7.7), GNNs (1.4%, 6.9), adversarial robustness (1.4%, 6.7), lightweight (1.2%, 6.0)
2021	BERT (1.8%, 24.4), contrastive (3.8%, 8.5), self-supervised (4.1%, 6.3), GNNs (2.7%, 5.9), fairness (1.1%, 5.6), federated (1.8%, 4.6)
2022	vision transformers (1.4%, 137), client (1.1%, 8.4), contrastive (4.7%, 7.6), distribution shift (1.2%, 7.1), diffusion (1.2%, 6.1), federated (3.3%, 5.7)
2023	vision transformers (1.9%, 190), prompts (1.7%, 39.7), CLIP (1.6%, 37.4), prompting (1.2%, 27.1), ViT (1.1%, 25.7), diffusion (3.2%, 5.2)
2024	LLMs (10.9%, 159), LLaMA (1.3%, 134), in-context (2.3%, 33.7), prompting (2.6%, 20.4), vision-language (2.7%, 17.5), instruction (2.5%, 13.7)
2025	LLaMA (2.4%, 236), MLLMs (1.8%, 183), LLM-based (1.6%, 164), preference optimization (1.5%, 148), hallucinations (1.4%, 141), instruction tuning (1.2%, 117)
2026	agentic (2.3%, 228), group relative (1.1%, 107), relative policy (1.1%, 46), RAG (1.6%, 24.5), retrieval-augmented generation (1.5%, 23.8), llm reasoning (1.4%, 21.4)



(a) Title form (b) Title mentions (c) Title length

Figure D.6: Title style, 2017 to 2026. Subfigure (a) is the share of titles with a colon and with a "Name: description" form, subfigure (b) the share mentioning LLMs or benchmarks, and subfigure (c) the mean title length in words.

E Further Analyses and Interpretations

Scaling across areas. From 2023 each submission carries a primary area chosen by the authors. In each of 2023, 2024, and 2025 we keep every area with at least 40 main-track submissions. This gives 12, 18, and 21 areas with between about 60 and 1,100 papers, as no area holds between 40 and 60. Within each area we compute the mean review words, rebuttal words, and follow-up words per paper, the author reply rate, the reviewer follow-up rate, the share of forums with a public comment, and the mean team size. Each quantity is fitted against the number of submissions of the area by (1), with a standard error, and Table E.1 lists the exponents. All of them except the public share of 2025 lie between -0.04 and 0.06. The follow-up rate of 2024 is the largest in units of its standard error, at 1.8 standard errors before rounding. The public share of 2025 is 0.14 with a standard error of 0.11. Review words per paper has an exponent of -0.03 with a standard error of 0.02 in each of the three years. Larger areas thus receive slightly shorter reviews, with a stable sign and a small size. This fit holds the era fixed, but the areas of a year also share the reviewer pool, the review form, the load per reviewer, the deadline, and the rules. The comparison can therefore show whether the number of papers alone changes effort per paper, but not whether venue size changes it through the pool, the form, the load, or the deadline. Those are the same for every area of a year. Areas also differ in research practice, e.g., theory against applications. The result thus shows that effort per paper does not depend on the size of an area, and it does not identify what venue size changes along the decade. A fit of the same quantities across venues within a year would vary the pool and the load at a fixed era, and it is left for future work.

Table E.1: Exponents of each per-paper quantity against the number of main-track submissions of a primary area within a year, with standard errors in parentheses. Figure 3c shows the rows for review words, rebuttal words, and the reviewer follow-up rate.

Quantity	2023 (12 areas)	2024 (18 areas)	2025 (21 areas)
Review words per paper	-0.03 (0.02)	-0.03 (0.02)	-0.03 (0.02)
Rebuttal words per paper	0.00 (0.04)	-0.01 (0.03)	0.00 (0.02)
Reviewer follow-up words per paper	0.00 (0.06)	-0.04 (0.05)	-0.03 (0.04)
Forums with a public comment, share	0.02 (0.15)	0.02 (0.12)	0.14 (0.11)
Author reply rate	0.00 (0.02)	0.00 (0.01)	0.01 (0.01)
Reviewer follow-up rate	-0.01 (0.03)	-0.03 (0.01)	0.01 (0.01)
Authors per paper	0.06 (0.04)	0.00 (0.05)	0.03 (0.06)

Topic groups over the decade. Table E.2 gives the number of submissions per group and year, the first year in which the group has at least 30 papers under review, from which its fits start, and its growth since then. Every group grew, from $7\times$ for graphs to $84\times$ for LLMs over five years and $55\times$ for vision, audio, and language applications over the decade, against $40\times$ for the venue. The LLM group held about 1% of submissions up to 2019 and a sixth in 2026, benchmarks rose from 2% to 8%, representation and transfer learning decreased from a third to 8%, and theory decreased from 14% to 6%. Tables E.3 and E.4 give each quantity in the group's first fitted year and in 2025. In every group, rebuttal words per paper grew by a factor between 1.7 and 3.7, author notes per paper by a factor between 1.8 and 2.7, and the follow-up rate rose to between 68% and 75%. Words per review rose in every group that starts in 2017 to 2019 and decreased slightly in the two that start in 2021, whose first year lies at the 2021 to 2022 peak of review length. Reviews are longest in reinforcement learning and in the sciences and shortest in vision, audio, and language applications and in generative models, and the safety group has the least positive assessment tone in every year. In 2023 to 2025 the mean rating of the highest group exceeds that of the lowest by 0.45 to 0.61 points, with generative models and LLMs highest and the other, graphs, and safety groups lowest, and the acceptance rate of generative models exceeds that of graphs by 16 percentage points in 2025. The withdrawn share of 2025 is highest in vision, audio, and language applications (38%) and lowest in theory (17%), and the public-comment share converged to between 3.5% and 5.4% in every group.

Table E.2: Submissions per topic group and year, the first year with at least 30 papers under review (First), and the growth of submissions from that year to 2026.

Group	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	First	Growth
LLMs and foundation models	5	9	9	39	39	60	136	604	1,498	3,284	2021	84×
Vision, audio, and language	47	82	112	190	179	239	458	720	1,405	2,576	2017	55×
Generative models	15	51	83	113	99	104	170	518	914	1,644	2018	32×
Safety and interpretability	21	67	145	268	330	400	587	927	1,572	2,712	2018	40×
Datasets and benchmarks	9	6	13	22	34	32	76	223	753	1,579	2021	46×
RL and robotics	46	131	223	355	395	464	587	799	944	1,808	2017	39×
Sciences	6	30	38	72	93	130	243	387	710	941	2019	25×
Theory and optimization	67	146	233	362	412	457	554	656	972	1,119	2017	17×
Graphs	7	26	61	134	163	186	265	382	397	455	2019	7×
Representation and transfer	160	266	375	629	741	787	1,136	1,245	1,158	1,593	2017	10×
Other	107	204	287	410	529	563	743	943	1,349	2,103	2017	20×
All submissions	490	1,018	1,579	2,594	3,014	3,422	4,955	7,404	11,672	19,814	2017	40×

Table E.3: Effort and engagement per topic group, in the group's first fitted year (First in Table E.2) and in 2025. Words and notes are per main-track paper, and the follow-up rate is the share of replied-to reviews with a reviewer follow-up.

Group	Words per review	Rebuttal words per paper	Author notes per paper	Follow-up rate	Authors per paper
LLMs and foundation models	459 to 430	2,630 to 4,752	6.3 to 12.9	24% to 70%	4.6 to 6.1
Vision, audio, and language	305 to 412	1,484 to 4,459	6.0 to 12.8	16% to 71%	3.9 to 6.0
Generative models	360 to 417	1,376 to 4,340	4.5 to 12.3	12% to 72%	3.8 to 5.6
Safety and interpretability	341 to 456	1,463 to 4,440	4.6 to 12.1	10% to 69%	3.9 to 4.9
Datasets and benchmarks	523 to 457	2,837 to 4,858	7.0 to 12.8	25% to 68%	5.8 to 8.0
RL and robotics	327 to 471	1,500 to 4,767	5.3 to 12.7	14% to 75%	4.3 to 5.0
Sciences	382 to 496	1,867 to 5,102	4.8 to 13.1	32% to 74%	4.3 to 5.4
Theory and optimization	286 to 448	1,360 to 3,851	5.8 to 10.5	22% to 72%	3.3 to 3.8
Graphs	375 to 425	2,184 to 4,777	6.3 to 12.8	33% to 74%	3.9 to 4.6
Representation and transfer	291 to 424	1,524 to 4,368	5.8 to 12.2	15% to 69%	3.6 to 4.6
Other	289 to 459	1,199 to 4,436	5.2 to 11.9	17% to 72%	3.8 to 4.6
All submissions	296 to 444	1,416 to 4,525	5.6 to 12.3	17% to 71%	3.7 to 5.3

Table E.4: Tone and outcomes per topic group, in the group's first fitted year and in 2025. Tone is the mean assessment-field tone of reviews, the mean rating is on the mapped scale over main-track papers, Accepted is the share of main-track papers accepted, Withdrawn the share of all submissions withdrawn, and Public comment the share of main-track forums with a public comment.

Group	Tone	Mean rating	Accepted	Withdrawn	Public comment
LLMs and foundation models	0.150 to 0.183	5.53 to 5.59	31% to 46%	8% to 20%	3% to 5%
Vision, audio, and language	0.152 to 0.160	5.68 to 5.63	38% to 49%	0% to 38%	53% to 4%
Generative models	0.129 to 0.160	5.47 to 5.77	42% to 54%	2% to 27%	34% to 5%
Safety and interpretability	0.063 to 0.121	5.58 to 5.43	41% to 39%	6% to 25%	54% to 4%
Datasets and benchmarks	0.144 to 0.152	5.49 to 5.58	39% to 43%	9% to 26%	16% to 4%
RL and robotics	0.208 to 0.176	5.93 to 5.50	54% to 43%	0% to 19%	46% to 4%
Sciences	0.128 to 0.155	5.41 to 5.46	40% to 43%	8% to 23%	20% to 5%
Theory and optimization	0.164 to 0.137	5.70 to 5.59	42% to 44%	0% to 17%	43% to 4%
Graphs	0.120 to 0.143	5.60 to 5.38	41% to 38%	8% to 23%	41% to 4%
Representation and transfer	0.149 to 0.153	5.57 to 5.41	38% to 39%	0% to 31%	43% to 4%
Other	0.161 to 0.146	5.64 to 5.31	36% to 36%	0% to 27%	43% to 3%
All submissions	0.159 to 0.153	5.69 to 5.51	40% to 43%	0% to 26%	45% to 4%

Table E.5: Reviewing per topic group, in the group's first fitted year and in 2025. Confidence is the mean reviewer confidence on the 1-to-5 scale (absent in 2020), Rated 8 or above the share of reviews with a rating of 8 or more on the mapped scale, Split verdict the share of main-track papers with a review at or above 6 and a review below it, Author reply tone the mean tone of author notes, and Follow-up rate the share of replied-to reviews with a reviewer follow-up.

Group	Confidence	Rated 8 or above	Split verdict	Author reply tone	Follow-up rate
LLMs and foundation models	3.92 to 3.64	6% to 18%	67% to 63%	0.189 to 0.271	24% to 70%
Vision, audio, and language	3.94 to 3.92	10% to 18%	45% to 59%	0.154 to 0.274	16% to 71%
Generative models	3.80 to 3.66	9% to 19%	56% to 56%	0.181 to 0.260	12% to 72%
Safety and interpretability	3.61 to 3.56	9% to 17%	43% to 63%	0.109 to 0.228	10% to 69%
Datasets and benchmarks	3.85 to 3.75	6% to 18%	61% to 65%	0.201 to 0.249	25% to 68%
RL and robotics	3.91 to 3.54	17% to 19%	39% to 62%	0.195 to 0.262	14% to 75%
Sciences	3.83 to 3.68	15% to 20%	49% to 59%	0.192 to 0.246	32% to 74%
Theory and optimization	3.79 to 3.36	15% to 20%	43% to 62%	0.186 to 0.219	22% to 72%
Graphs	3.98 to 3.58	7% to 14%	46% to 62%	0.170 to 0.243	33% to 74%
Representation and transfer	3.87 to 3.66	10% to 15%	42% to 61%	0.162 to 0.252	15% to 69%
Other	3.74 to 3.50	9% to 16%	41% to 63%	0.187 to 0.232	17% to 72%
All submissions	3.82 to 3.62	12% to 18%	41% to 61%	0.174 to 0.249	17% to 71%

Table E.5 gives reviewer confidence, the share of reviews rated 8 or above, the share of split verdicts, the tone of author replies, and the follow-up rate. Mean reviewer confidence is highest in vision, audio, and language applications in every year with a confidence field except 2022 and lowest in theory in every year since 2021, with a gap of 0.56 points in 2025, and the mean over the groups decreased by 0.2 points over the decade. The safety group has the least positive author replies in 7 of the 9 years, as it has the least positive reviews in 8 of them, which fits a vocabulary effect that acts on both sides of the discussion. Generative models have the fewest split verdicts in 4 of the 5 years since 2021, the highest mean rating in 2022, 2024, and 2025, and the highest acceptance rate in 2024 to 2026, whereas the graphs group has the lowest acceptance rate in 2024 and 2026. The share of reviews rated 8 or above is highest in the LLM group in 2022 to 2024 and in the sciences in 2025, and the follow-up rate is highest in reinforcement learning in 5 of the 9 years and lowest in benchmarks in 2023 to 2025, by a tenth of the venue value in 2025. Authors per paper are largest in benchmarks in 2021 and in every year since 2023 and in LLMs in 2022, at more than twice the theory value in 2025 and 2026, and smallest in theory in every year.

Table E.6 lists the exponent of each quantity in the number of the group's submissions, with fit standard errors, fitted from the group's first fitted year to 2025 (2026 for authors per paper). The rebuttal exponent exceeds the review exponent in every group, by 0.10 to 0.24, which is at least 1.6 fit standard errors in every group and at least 3 in 7 of the 11 groups. The exponents of words per review, rebuttal words, and the follow-up rate are smallest for the two groups that grew fastest, LLMs and benchmarks, at -0.05 and -0.04 for words per review and 0.21 and 0.20 for the follow-up rate against 0.13 and 0.50 for the venue, because their quantities moved with the year like those of every other group while their size grew $46\times$ to $84\times$ in five years. The exponent of authors per paper is lowest in theory, RL, and LLMs, at about half of the venue-wide value or less, and highest in representation learning and applications. The public share has an exponent near zero in the LLM group (-0.05) against -0.93 for the venue, i.e., the share of LLM forums with a public comment did not decrease as the group grew, and it was the highest of all groups in 2022 to 2024. Table E.7 gives the two-way fit on the group-year panel, $\log q_{gy} \approx a_g + \alpha \log n_{gy} + \beta(y - 2017)$, where q_{gy} is the quantity and n_{gy} the number of submissions of group g in year y, over the group-years with at least 30 papers under review and without the other group, next to the fit with the size term alone and the venue-wide exponent. With a term per group and no year term, the size exponent lies between half of the venue-wide exponent and the venue-wide exponent. With the year held fixed, the size exponents of words per review, review and rebuttal words per paper, author notes per paper, and the follow-up rate lie between -0.13 and -0.02, and the year term carries the trend, e.g., 0.16 per year for rebuttal words, i.e., a factor of 3.7 over 8 years against an observed factor of 3.2. Authors per paper keeps a size exponent of 0.06 at a fixed year, and the public share has a size exponent of 0.24 at a fixed year, as in the area fit of 2025, and a year coefficient of -0.44. The within-group correlation of log size with the year is 0.92, so the panel separates the two only partly, and the year term also absorbs the review form, the tools, and the reviewer pool, which every group of a year shares.

Table E.6: Exponents of each quantity in the number of submissions of the group, with fit standard errors in parentheses, fitted from the group's first fitted year to 2025, or to 2026 for authors per paper. Rebuttal minus review is the slope of the ratio of rebuttal to review words per paper, and Public share is the share of main-track forums with a public comment. The last row fits the venue as a whole in the same way.

Group	Words/review	Rebuttal words	Rebuttal − review	Follow-up rate	Public share	Authors/paper
LLMs and foundation models	−0.05 (0.03)	0.15 (0.02)	0.20 (0.02)	0.21 (0.09)	−0.05 (0.26)	0.04 (0.03)
Vision, audio, and language	0.06 (0.04)	0.36 (0.07)	0.19 (0.06)	0.47 (0.18)	−0.94 (0.15)	0.13 (0.02)
Generative models	−0.01 (0.05)	0.33 (0.10)	0.24 (0.06)	0.54 (0.17)	−0.64 (0.15)	0.13 (0.01)
Safety and interpretability	0.09 (0.03)	0.36 (0.04)	0.16 (0.04)	0.57 (0.14)	−1.07 (0.21)	0.08 (0.01)
Datasets and benchmarks	−0.04 (0.03)	0.17 (0.03)	0.20 (0.02)	0.20 (0.12)	−0.35 (0.19)	0.10 (0.01)
RL and robotics	0.13 (0.04)	0.39 (0.08)	0.15 (0.08)	0.55 (0.10)	−0.99 (0.16)	0.07 (0.02)
Sciences	0.06 (0.04)	0.35 (0.05)	0.19 (0.05)	0.43 (0.19)	−0.46 (0.13)	0.09 (0.01)
Theory and optimization	0.18 (0.04)	0.41 (0.07)	0.10 (0.06)	0.49 (0.17)	−1.12 (0.18)	0.07 (0.01)
Graphs	0.08 (0.07)	0.44 (0.15)	0.20 (0.12)	0.55 (0.36)	−1.40 (0.22)	0.10 (0.02)
Representation and transfer	0.18 (0.04)	0.53 (0.11)	0.19 (0.11)	0.64 (0.23)	−1.28 (0.18)	0.14 (0.01)
Other	0.18 (0.04)	0.55 (0.07)	0.23 (0.07)	0.62 (0.17)	−1.20 (0.15)	0.09 (0.01)
All submissions	0.13 (0.04)	0.40 (0.06)	0.17 (0.06)	0.50 (0.13)	−0.93 (0.11)	0.13 (0.01)

Table E.7: The two-way fit on the group-year panel. Venue is the exponent in venue size over the years, Size alone the exponent in group size with a term per group and no year term, and the next two columns the size exponent and the per-year coefficient of the fit with both terms, each with its fit standard error. Corr. is the within-group correlation of log size with the year.

Quantity	Group-years	Venue	Size over years	Size within a year	Year, size fixed	Corr.
Words per review	84	0.13 (0.04)	0.07 (0.01)	−0.06 (0.03)	0.055 (0.013)	0.92
Review words per paper	84	0.24 (0.06)	0.15 (0.02)	−0.13 (0.04)	0.123 (0.017)	0.92
Rebuttal words per paper	84	0.40 (0.06)	0.34 (0.02)	−0.04 (0.04)	0.163 (0.014)	0.92
Author notes per paper	84	0.28 (0.08)	0.28 (0.02)	−0.02 (0.05)	0.129 (0.019)	0.92
Follow-up rate	84	0.50 (0.13)	0.46 (0.05)	−0.08 (0.10)	0.234 (0.040)	0.92
Authors per paper	95	0.13 (0.01)	0.10 (0.01)	0.06 (0.01)	0.018 (0.005)	0.92
Public share	84	−0.93 (0.11)	−0.77 (0.07)	0.24 (0.11)	−0.435 (0.043)	0.92

Collinearity of the venue-wide regressors. Table E.8 gives the correlations over 2017 to 2025 of the calendar year, $\log N$, the review-form period (a single field up to 2021, then the 2022, the 2023, and the 2024 form), and the submission shares of three groups, together with the variance inflation factor of each of the first four when regressed on the other three, i.e., $1/(1 - R^2)$. Year and $\log N$ are correlated at 0.98, both are correlated with the form period at 0.84 to 0.90 and with the LLM share at 0.77 to 0.80, and the theory share is correlated at -0.97 with the form period. The variance inflation factors of year and $\log N$ are 84 and 55, so a fit that included both would have standard errors about 9 and 7 times those of a fit with a single regressor, on 9 points. The venue-wide fits therefore report the line in $\log N$ and the line in the year side by side and identify neither, and the RL share, which is nearly flat, is the only regressor that the others do not predict. The area fit varies size at a fixed year, form, pool, and mix, so it identifies the effect of the number of papers alone. The group panel varies size and year separately, because the groups grew at different rates, and identifies size against everything that changes with the year, but not the year against the form, the tools, or the mix, which change together with it.

Table E.8: Correlations of the venue-wide regressors over 2017 to 2025 and the variance inflation factor (VIF) of each of the first four regressed on the other three.

Regressor	Year	$\log N$	Form period	LLM share	RL share	Theory share	VIF
Year	1.00	0.98	0.90	0.80	−0.35	−0.86	84
$\log N$	0.98	1.00	0.84	0.77	−0.25	−0.81	55
Form period	0.90	0.84	1.00	0.88	−0.61	−0.97	16
LLM share	0.80	0.77	0.88	1.00	−0.74	−0.94	6
RL share	−0.35	−0.25	−0.61	−0.74	1.00	0.72	n/a
Theory share	−0.86	−0.81	−0.97	−0.94	0.72	1.00	n/a

Cumulative advantage. Authors are keyed across years by profile ID where a submission carries one, by the single ID that carries the name otherwise, and by email address where neither applies, as described in Appendix B. For each year t from 2017 to 2025 and each author with k papers in year

t, we count the papers of the same author in year $t + 1$. An absent author counts as zero. The mean of this count by k, in bins of 1, 2, 3, 4 to 5, 6 to 10, and over 10 papers, rises roughly in proportion to k in every year. Newman (2001) used the same construction on co-authorship networks, where such a rise marks preferential attachment, i.e., new links attach to the nodes that already have many (Barabási & Albert, 1999). Fixed differences in productivity between authors would also raise the mean with k, because an author with a fixed high rate has a high count in both years. To separate the two, we keep the authors present in both years $t - 1$ and t and regress the count k_{t+1} on k_t and k_{t-1} by least squares. The standard errors allow the variance of a count to grow with its mean, as the counts are Poisson-like. If every author had a fixed rate and no feedback, the two past counts would be equally informative about that rate, so their coefficients would be about equal. Under cumulative advantage the current count instead carries extra weight. That premise holds exactly if the fixed rates follow a gamma distribution, because the expected value of k_{t+1} given the two counts is then a straight line in $k_{t-1} + k_t$. Keeping only the authors present in both earlier years does not change that expected value. Under other rate distributions it is only approximate, and the selection can tilt a least-squares fit towards either count. We therefore check the premise by simulation in 6 steps. (1) Each selected author's rate is estimated from the two observed counts, each divided by the growth of the venue since year $t - 1$. (2) The pool of authors is weighted by the inverse of the probability that an author of that rate has at least a single paper in both years. Authors who passed the selection by chance then appear at their unconditional frequency. (3) Three years of counts are drawn from those rates as independent Poisson variables, i.e., random counts with the given average and no feedback, under the growth of each year. (4) The selection is imposed again. (5) The regression is refitted. (6) The steps are repeated 200 times per year, and the share of simulations whose gap between the two coefficients reaches the observed gap is recorded. The comparison rejects fixed rates, and not every model without feedback, because a rate that drifts smoothly over a career would also make the recent count more informative than the earlier one. Table E.9 lists the coefficients, the observed gap between them, the standard deviation of the gap over the simulations, and the share of simulations whose gap reaches the observed gap. The simulated gap centers within 0.04 of zero in every year. In 2018, with 467 authors, the observed gap is 0.25 against a standard deviation of 0.22 under fixed rates, and 8% of the simulations reach it. From 2019 on the observed gap is at least 0.25 against a standard deviation of at most 0.15, and none of the 200 simulations reaches it, so the share is below 0.5%. Pooled over 2018 to 2025 the coefficient on k_t is 0.85 with a standard error of 0.02, and on k_{t-1} it is 0.29 with a standard error of 0.02. In every single year the coefficient on k_t is at least 1.4 times that on k_{t-1}. Figure 2b shows the same pooled data as means by bin, for k_t of 1, 2, 3, and 4 or more at each of k_{t-1} of 1, 2, and 3 or more. For example, among authors with a single paper in year $t - 1$, the mean next-year count is 0.90, 1.56, 2.26, and 4.11 for k_t of 1, 2, 3, and 4 or more. For authors with three or more papers in year $t - 1$ it is 1.68, 2.64, 3.42, and 7.92. The positive coefficient on k_{t-1} shows that fixed differences exist as well, so the result supports state dependence on top of them and not instead of them.

Table E.9: Regression of next year's paper count on this year's count k_t and the count of the year before k_{t-1}, among authors present in both earlier years, with standard errors in parentheses, which allow the variance of a count to grow with its mean, and the coefficient on this year's count alone. The last three columns give the gap between the two coefficients, the standard deviation of the gap under fixed rates, and the share of the 200 simulations reaching the observed gap.

Year t	Authors	Coefficient on k_t	Coefficient on k_{t-1}	On k_t alone	Gap	SD under fixed rates	Share $\geq$ gap
2018	467	0.82 (0.15)	0.57 (0.10)	1.12	0.25	0.22	8%
2019	931	0.72 (0.10)	0.29 (0.08)	0.84	0.43	0.15	<0.5%
2020	1,760	0.57 (0.06)	0.32 (0.08)	0.73	0.25	0.09	<0.5%
2021	1,975	0.66 (0.06)	0.25 (0.05)	0.79	0.41	0.07	<0.5%
2022	2,996	0.77 (0.06)	0.23 (0.05)	0.90	0.55	0.07	<0.5%
2023	4,130	0.73 (0.04)	0.37 (0.05)	0.94	0.36	0.06	<0.5%
2024	6,380	0.81 (0.03)	0.34 (0.04)	0.97	0.47	0.05	<0.5%
2025	10,799	0.93 (0.03)	0.28 (0.03)	1.07	0.65	0.04	<0.5%
pooled	29,438	0.85 (0.02)	0.29 (0.02)	n/a	n/a	n/a	n/a

The mean next-year count for $k = 1$ took the values 0.40, 0.39, 0.49, 0.29, 0.38, 0.48, 0.54, 0.64, and 0.69 for the transitions from 2017 to 2018 through 2025 to 2026. These means cover all single-paper authors, whereas the value of 0.90 in Figure 2b covers only those also present in year $t - 1$, which is why the two differ. The venue grew by a factor between 1.1 and 2.1 each year, which inflates every next-year count, and the count divided by that factor rose from 0.19 to 0.41. The share of single-

paper authors with any submission the next year rose from 27% to 38%. Before 2021 about 55% of author slots are bridged to a profile ID by a name that exactly one ID carries, and about 43% are keyed by email address. A name shared by several people merges them through the bridge, while an author who changed email address is split, so the counts of the early years carry both biases, which the later years, keyed almost entirely by ID, do not. Among the authors of 2021 to 2026 the share of names that cover more than a single profile ID rose from 1.7% to 6.9%, which bounds how often the bridge of an earlier year would merge people.

Public commenters. A public note is a comment whose signature belongs to none of the author, reviewer, and chair roles of its forum. Signed notes carry an OpenReview profile ID, and anonymous public comments were possible up to 2019. Distinct commenters are the distinct profile IDs among signed notes of a year, and a returning commenter is an ID that also commented in the previous year. Table E.10 lists the counts. Anonymous public comments ended in 2020, so the series of all public notes joins a period with anonymous notes and a period without. Table E.11 gives the fits against submissions on both sides of that break and on signed notes alone. Over 2017 to 2019 all public notes and forums with a comment grew with the venue, on three points, while signed notes did not. Every count over 2020 to 2025 has an exponent within 0.12 of zero, with standard errors of 0.26 to 0.34 that cannot separate zero from one half. Over those years the count of signed notes decreased from 932 to 663 while the venue grew $4\times$, which any exponent near one would contradict. The 2020 split is taken from the venue's change to signed public comments and not from the series. We also place the break at each of 2020 to 2023 in turn and select the break year with the smallest residual sum of squares of the two-segment fit. The best-fitting break is 2021 for all public notes and for forums with a comment, and it is 2023 for signed notes and for distinct commenters. After a 2023 break, signed notes and distinct commenters have exponents of 0.91 with a standard error of 0.21 and 1.07 with 0.17 over 2023 to 2025. That is growth in proportion to the venue, on three points. The notes rose from 307 to 663 and the commenters from 187 to 464 over those years. The flat count is therefore a statement about the decade, and the last three years may mark a reversal. Returning commenters in 2018 to 2020 are biased down because an anonymous commenter of the previous year cannot be matched.

Table E.10: Public comments per year, the commenters behind them, and how many of them commented in the previous year. The column Forums with a comment counts every track, whereas Table 1 counts main-track forums.

Year	Public notes	Signed	Distinct commenters	Returning	Forums with a comment	Notes per commenter
2017	800	526	188	0	221	2.8
2018	1,032	336	221	13	344	1.5
2019	1,449	376	217	18	490	1.7
2020	932	932	549	39	463	1.7
2021	462	462	301	25	264	1.5
2022	452	452	320	25	266	1.4
2023	307	307	187	19	207	1.6
2024	378	378	252	9	233	1.5
2025	663	663	464	11	370	1.4

Table E.11: Exponents of public-comment counts against submissions, with standard errors, over the whole period and on each side of the end of anonymous commenting.

Quantity	2017 to 2025	2017 to 2019	2020 to 2025
All public notes	−0.30 (0.16)	0.49 (0.11)	−0.12 (0.34)
Signed notes	0.03 (0.13)	−0.32 (0.23)	−0.12 (0.34)
Distinct commenters	0.19 (0.13)	0.13 (0.07)	−0.06 (0.34)
Forums with a comment	−0.01 (0.12)	0.67 (0.05)	−0.06 (0.26)

Rebuttal response. Consider each main-track paper with at least a single review and a single author reply. Let w be the total words of its reviews, m the number of its reviews, and v the total words of its author replies. Within each year we fit

$$\log v \approx b_0 + b_1 \log w + b_2 \log m \tag{E.1}$$

by least squares. The coefficient b_1 is the percentage change in rebuttal length per percentage change in review length at a fixed venue size, and b_2 is the same for the number of reviews. For example,

a value $b_1 = 0.5$ means that a review 10% longer corresponds to a rebuttal 5% longer. Figure 3b plots both coefficients by year, each fitted on every main-track paper of the year. The coefficient on review length is 0.47 to 0.64 from 2017 to 2023, with a mean of 0.53. In 2024 and 2025 it is 0.38 and 0.40, so it decreased from about 0.5 to 0.4, or by about a quarter. With quoted review text removed from the rebuttal words the coefficient changes by at most 0.02 in any year. The same regression on the raw counts gives the number of rebuttal words added per review word. It lies between 0.70 and 1.06 from 2017 to 2025 and shows no trend. The proportional response decreased because the median ratio of rebuttal words to review words rose from 1.6 to 2.7 over the same years. Rebuttals thus gained words that do not vary with the review, in line with the conjecture of Section 3.1. The paper summary field of 2022 on and the questions field of 2024 on are part of the review words. The flattened table keeps no per-field counts, so the response to the critique fields alone was not fitted. The fit explains at most a tenth of the variance in every year (R^2 of 0.06 to 0.10), so the coefficients describe the average response and not any single paper.

Withdrawal after a low rating. Among reviewed papers with at least two ratings, consider those with a mean rating below 5 on the mapped scale. The share of them that were withdrawn was 0.12, 0.21, 0.29, 0.31, 0.43, 0.44, 0.44, and 0.51 in 2018 to 2025. Among the rest the share was 0.01, 0.02, 0.07, 0.04, 0.08, 0.06, 0.05, and 0.06 in the same years. These shares are computed from the per-paper score table and use no withdrawal note.

Announced and actual score changes. Announcements are counted per reviewer comment in the main text. Per review, i.e., whether any later comment by the reviewer of that review announces a change, 6.4% of reviews carried an announced raise in 2024 and 9.7% in 2025. Among reviews with an author reply the shares are 8.1% and 13.2%. A lowering was announced for 0.3% and 0.5% of reviews. Kargaran et al. (2025) report, from before-and-after scores saved outside OpenReview, that 17% of reviews were raised and 1% lowered in 2024, and 23% raised and 1% lowered in 2025. Louis et al. (2026) report 21.0% raised and 1.4% lowered over both years, and 27.0% raised among reviews with an author reply. Announcements thus capture about two fifths of the raises and about a third of the lowerings. Promised raises, which include definite announcements of a future raise such as "I will raise my score", are excluded from the announced raises, and they concern 3.7% and 5.8% of reviews. With them included, 10.2% and 15.5% of reviews carry an announced or promised raise, about three fifths and two thirds of the external counts. The populations differ slightly, as the external collections cover the papers that stayed under review, so the two fractions are approximate. Announcements are therefore an undercount of actual changes and not an upper bound. The rates of Figure 6d are per reviewer comment, so the $4\times$ rise in follow-ups does not raise them by itself. A reviewer who now posts a routine follow-up may still state a change that an earlier reviewer made without a comment. No before-and-after scores exist for the early years to separate the two.

Social-science theories. Table E.12 matches each finding to an established theory of collective behavior. The matching is an interpretation. The data do not distinguish any of these theories from an alternative, so the labels add an interpretation of the findings and not evidence for it.

Table E.12: Findings, the mechanisms proposed here, and the social-science theories that offer an interpretation.

Finding	Mechanism here	Theory
Submissions concentrate on a few prolific authors	Advantage compounds through students and co-authors, checked through next-year submissions at a fixed earlier count	Cumulative advantage (Simon, 1955; Price, 1976), i.e., those who have more gain more, also named the Matthew effect (Merton, 1968), with the pattern of Newman (2001)
Public comments show no relation with size	One-time visitors with a private reason supply a public good	Logic of collective action (Olson, 1965), i.e., a good that benefits everyone is undersupplied because each person gains without contributing, and attention scarcity (Simon, 1971), i.e., attention is the resource that information consumes
Withdrawals and rebuttal effort grow (two findings of Section 3.1)	Low-scored papers exit, more of them each year, and the rest use voice	Exit, voice, and loyalty (Hirschman, 1970), i.e., members of a declining organization either leave (exit) or speak up (voice)
Follow-ups adopt thanks, scores announced up	A reply to a person returns courtesy and protects face	Norm of reciprocity (Gouldner, 1960), i.e., a favor calls for a return, and politeness theory (Brown & Levinson, 1987), i.e., speakers protect each other's public self-image, called face
Required follow-ups are short and routine	A rule is met by a formality	Ceremonial conformity (Meyer & Rowan, 1977), i.e., an organization meets a rule by a formality that shows compliance

F THE SIGNAL SHARE OF A RATING: DERIVATIONS, ESTIMATOR, AND PROJECTION

Reliability of the panel mean. By A1 the variance of a rating is $\mathrm{Var}(q_i) + \mathrm{Var}(e_{ij}) = \sigma_q^2 + \sigma_e^2$, where the cross term vanishes by the independence of q_i and e_{ij}, and for two reviewers $j \neq k$ of the same paper expanding the product and using the independence of e_{ij}, e_{ik}, and q_i gives $\mathrm{Cov}(r_{ij}, r_{ik}) = \mathrm{Cov}(q_i + e_{ij}, q_i + e_{ik}) = \mathrm{Var}(q_i) = \sigma_q^2$, so the correlation between two ratings of a paper is $\sigma_q^2/(\sigma_q^2 + \sigma_e^2) = \rho$. Substituting A1 into the panel mean gives $\bar{r}_i = q_i + \bar{e}_i$ with $\bar{e}_i = \frac{1}{m}\sum_{j=1}^{m} e_{ij}$, whose variance is σ_e^2/m because the m noise terms are independent, so that $\mathrm{Var}(\bar{r}_i) = \sigma_q^2 + \sigma_e^2/m$ and $\mathrm{Cov}(\bar{r}_i, q_i) = \mathrm{Cov}(q_i + \bar{e}_i, q_i) = \sigma_q^2$. Squaring the ratio of the covariance to the product of the standard deviations, then multiplying the numerator and the denominator by m/σ_q^2, and finally dividing both by $\sigma_q^2 + \sigma_e^2$ and using $\sigma_q^2/(\sigma_q^2 + \sigma_e^2) = \rho$ and $\sigma_e^2/(\sigma_q^2 + \sigma_e^2) = 1 - \rho$ gives

$$R_m = \frac{\sigma_q^4}{\sigma_q^2(\sigma_q^2 + \sigma_e^2/m)} = \frac{m\sigma_q^2}{m\sigma_q^2 + \sigma_e^2} = \frac{m\rho}{m\rho + 1 - \rho} = \frac{m\rho}{1 + (m-1)\rho}, \tag{F.1}$$

which is the reliability of (2), with $\mathrm{Corr}(\bar{r}_i, q_i) = \sqrt{R_m}$. For any $\rho > 0$, dividing the numerator and the denominator by m gives $R_m = \rho/(\rho + (1-\rho)/m) \to 1$ as $m \to \infty$, so enough reviews recover the quality whenever a single review carries any signal. For a small share, $(m-1)\rho \ll 1$, the denominator of (F.1) is close to one, and the first-order expansion gives $R_m \approx m\rho$.

What the decision recovers. Under A1 and A2 the pair $(q_i, \bar{r}_i)$ is jointly Gaussian, so the conditional mean of q_i given $\bar{r}_i$ is linear with the coefficient $\mathrm{Cov}(\bar{r}_i, q_i)/\mathrm{Var}(\bar{r}_i) = \sigma_q^2/(\sigma_q^2 + \sigma_e^2/m) = R_m$ by the second form of (F.1), i.e., $\mathbb{E}[q_i \mid \bar{r}_i] = \mu + R_m(\bar{r}_i - \mu)$. Let $Z_i = (\bar{r}_i - \mu)/\sqrt{\mathrm{Var}(\bar{r}_i)}$ denote the standardized panel mean, which is standard Gaussian. Substituting $\mathrm{Var}(\bar{r}_i) = \sigma_q^2/R_m$, which is the second form of (F.1) rearranged, gives $\bar{r}_i - \mu = \sigma_q Z_i/\sqrt{R_m}$ and hence $\mathbb{E}[q_i \mid \bar{r}_i] = \mu + \sqrt{R_m}\,\sigma_q Z_i$. By A3 paper i is accepted if and only if $Z_i > z_a$, which has probability $1 - \Phi(z_a) = a$, and taking the expectation of the conditional mean over the accepted papers gives

$$\mathbb{E}[q_i \mid \text{accepted}] - \mu = \sqrt{R_m}\,\sigma_q\,\mathbb{E}[Z_i \mid Z_i > z_a] = \sqrt{R_m}\,\sigma_q\,\frac{1}{a}\int_{z_a}^{\infty} u\,\phi(u)\,du = \sqrt{R_m}\,\sigma_q\,\frac{\phi(z_a)}{a}, \tag{F.2}$$

where the last equality integrates $u\phi(u) = -\phi'(u)$ from z_a to infinity. Random choice accepts papers whose expected quality is μ, so the left side is the selection gain over random choice, and with perfect information, i.e., $\rho = 1$ and $R_m = 1$, the gain is $\sigma_q\phi(z_a)/a$, so the ratio of the actual gain to the perfect-information gain is $\sqrt{R_m}$, which by the small-share expansion is $\sqrt{m\rho}$ when $(m-1)\rho \ll 1$. Let $Z_i^q = (q_i - \mu)/\sigma_q$ denote the standardized quality, so that (Z_i^q, Z_i) is a standard bivariate Gaussian pair with correlation $\sqrt{R_m}$ by (F.1). The probability that a paper in the true top share a is accepted is

$$\Pr(Z_i > z_a \mid Z_i^q > z_a) = \frac{\Pr(Z_i > z_a,\ Z_i^q > z_a)}{a}, \tag{F.3}$$

which is a bivariate Gaussian orthant probability, evaluated numerically, that equals a under random choice and one under perfect information. For the gain to be halfway to random choice at four reviews, i.e., $\sqrt{R_4} = 0.41$, solving (F.1) for ρ gives $\rho = R_4/(4 - 3R_4) = 0.048$, which is about a seventh of the 2025 share.

Coin-flip decisions and the two-panel disagreement rate. Under A1 and A2, given q_i the panel mean $\bar{r}_i = q_i + \bar{e}_i$ is Gaussian with mean q_i and variance σ_e^2/m, and by A3 paper i is accepted if and only if $\bar{r}_i > c$, where the threshold c is the $(1-a)$ quantile of $\bar{r}_i$, i.e., $c = \mu + z_a\sqrt{\sigma_q^2 + \sigma_e^2/m}$ by $\mathrm{Var}(\bar{r}_i) = \sigma_q^2 + \sigma_e^2/m$. Standardizing $\bar{r}_i - q_i$ by $\sigma_e/\sqrt{m}$ gives the acceptance probability of a paper of quality q_i,

$$p(q_i) = \Pr(\bar{r}_i > c \mid q_i) = \Pr\left(\frac{\bar{e}_i}{\sigma_e/\sqrt{m}} > \frac{c - q_i}{\sigma_e/\sqrt{m}}\right) = \Phi\left(\frac{(q_i - c)\sqrt{m}}{\sigma_e}\right), \tag{F.4}$$

which is increasing in q_i, so $p(q_i) \in [0.25, 0.75]$ if and only if $(q_i - c)\sqrt{m}/\sigma_e \in [-z_{0.75}, z_{0.75}]$, i.e., $|q_i - c| \leq z_{0.75}\,\sigma_e/\sqrt{m}$. By A2, $Z_i^q = (q_i - \mu)/\sigma_q$ is standard Gaussian, and dividing the band by

σ_q gives its center $(c-\mu)/\sigma_q = z_a\sqrt{1+\sigma_e^2/(m\sigma_q^2)} = z_a\sqrt{1+\tau^2}$ and its half-width $z_{0.75}\tau$ with $\tau = \sigma_e/(\sqrt{m}\,\sigma_q)$. Dividing the numerator and the denominator of the second form of (F.1) by σ_q^2 gives $R_m = 1/(1+\tau^2)$, so $\sqrt{1+\tau^2} = 1/\sqrt{R_m}$ and $\tau = \sqrt{(1-R_m)/R_m}$, and the share of the quality distribution in the band is the difference of Φ at its two ends,

$$\kappa = \Phi\left(\frac{z_a}{\sqrt{R_m}} + z_{0.75}\,\tau\right) - \Phi\left(\frac{z_a}{\sqrt{R_m}} - z_{0.75}\,\tau\right). \tag{F.5}$$

At $\rho = 0$, $\tau \to \infty$ and every paper has $p(q_i) = a$, so $\kappa = 1$ whenever $a \in [0.25, 0.75]$, and at $\rho = 1$, $\tau = 0$ and $\kappa = 0$. Two independent panels of the same paper accept it independently given q_i, each with probability $p(q_i)$, so they disagree with probability $2p(q_i)(1-p(q_i))$, and taking the expectation over q_i gives the disagreement rate $2\,\mathbb{E}[p(q_i)] - 2\,\mathbb{E}[p(q_i)^2] = 2\,(a - \Pr(\text{both accept}))$, because $\mathbb{E}[p(q_i)] = a$ and $\mathbb{E}[p(q_i)^2]$ is the probability that both panels accept. Two panel means share q_i and have covariance σ_q^2 and variance $\sigma_q^2 + \sigma_e^2/m$, whose ratio is R_m by the second form of (F.1), so $\Pr(\text{both accept})$ is the orthant probability of (F.3) with correlation R_m in place of $\sqrt{R_m}$, and at $\rho = 0$ the rate is $2(a-a^2) = 2a(1-a)$. At the 2025 values the rate is 27% at $a = 0.43$ and 21% at $a = 0.25$.

Link to growth. Substituting A4, i.e., $\sigma_e^2 = \sigma_0^2/\theta$ with σ_0^2 the noise variance of a single check, into the definition of ρ gives the signal share as a function of attention,

$$\rho = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_0^2/\theta} = \frac{\sigma_q^2\,\theta}{\sigma_q^2\,\theta + \sigma_0^2}. \tag{F.6}$$

Let $\theta = \theta_0 N^{-\delta}$ with $\delta \geq 0$. Taking the logarithm of (F.6) and differentiating in $\log\theta$ gives

$$\frac{d\log\rho}{d\log\theta} = 1 - \frac{\sigma_q^2\theta}{\sigma_q^2\theta + \sigma_0^2} = 1-\rho, \qquad \text{so} \qquad \frac{d\log\rho}{d\log N} = -\delta\,(1-\rho). \tag{F.7}$$

In the regime $\sigma_q^2\theta \ll \sigma_0^2$, (F.6) gives $\rho \approx (\sigma_q^2\theta_0/\sigma_0^2)N^{-\delta}$, so by the small-share expansion the gain scales as $\sqrt{m}\,N^{-\delta/2}$, and with reviews per paper $m \propto N^{0.11}$ it decreases with the venue only if $\delta > 0.11$. Let P denote the size of the reviewer pool and $L = mN/P$ the mean number of reviews per reviewer. Reviewers are drawn from the authors, and distinct authors grow as $N^{1.03}$, so under a pool proportional to the authors $L \propto mN/N^{1.03} \propto N^{0.08}$, which rose by about a third over the decade. If the hours that a reviewer gives to reviewing are fixed, attention per review is $\theta \propto 1/L \propto N^{-0.08}$, i.e., $\delta \approx 0.08$, and (F.7) predicts a slope of $\log\rho$ against $\log N$ of about $-0.08\,(1-\rho) \approx -0.05$ at $\rho \approx 0.35$. Words per review rose as $N^{0.13}$, which is output and not attention, so it does not show a decrease of θ. The text-to-score correlation concerns the map from the words of a review to the reviewer's own rating r_{ij} and not the map from r_{ij} to q_i, so it can decrease while ρ stays fixed, e.g., through a change of the form or of the drafting tools, and it can also decrease together with ρ if a rating gains a component that neither the words nor the paper determine. The split-verdict share at a fixed panel of three reviews rose from 40% to 50% by 2020 and was flat after (Table D.5), and under A1 a split at fixed m becomes more likely as ρ decreases, so the model predicts that ρ decreased up to 2020 and was about flat after, which Table F.1 shows.

Estimator. Let n denote the number of papers of a year with at least two ratings, m_i the number of ratings of paper i, $M = \sum_{i=1}^n m_i$, $\bar r$ the mean of all ratings, $\mathrm{MS}_b = \frac{1}{n-1}\sum_{i=1}^n m_i(\bar r_i - \bar r)^2$, $\mathrm{MS}_w = \frac{1}{M-n}\sum_{i=1}^n\sum_{j=1}^{m_i}(r_{ij}-\bar r_i)^2$, and $m_0 = \frac{1}{n-1}\big(M - \frac{1}{M}\sum_{i=1}^n m_i^2\big)$. Under A1 each inner sum of MS_w is $m_i - 1$ times the sample variance of the noise of paper i, so $\mathbb{E}[\mathrm{MS}_w] = \sigma_e^2$, and $\mathbb{E}[\mathrm{MS}_b] = \sigma_e^2 + m_0\sigma_q^2$ with the effective panel size m_0, which equals m when every paper has m ratings. Solving the two expectations gives $\hat\sigma_e^2 = \mathrm{MS}_w$ and $\hat\sigma_q^2 = (\mathrm{MS}_b - \mathrm{MS}_w)/m_0$, and dividing $\hat\sigma_q^2$ by $\hat\sigma_q^2 + \hat\sigma_e^2$ and multiplying the numerator and the denominator by m_0 gives

$$\hat\rho = \frac{\mathrm{MS}_b - \mathrm{MS}_w}{\mathrm{MS}_b + (m_0-1)\,\mathrm{MS}_w}. \tag{F.8}$$

The population is every main-track paper of the year with at least two ratings, as in Table D.5, so the estimate uses the complete record of the year and carries no sampling error, and the 2020 ratings are mapped to the common scale by Table B.2. For each year, (F.1) is evaluated at the year's mean panel size m, (F.2) in rating points with $\hat\sigma_q$, and (F.3), (F.5), and the disagreement rate at the year's acceptance rate a among main-track papers with a decision.

Alternative populations and the slope. Table F.1 lists the estimates for three populations, together with the rank-based estimate, i.e., (F.8) applied to the percentile rank of each rating within its year, which does not depend on the spacing of the scale, and the estimated standard deviations of quality and noise in rating points. The rank-based estimate follows the main-track estimate within 0.03 in every year. With the withdrawn papers included the share is higher by 0.02 to 0.05 from 2021 on and reaches 0.38 in 2025, because withdrawn papers carry low ratings and their exclusion narrows the between-paper variance, so part of the decrease after 2018 is the growing removal of low-rated papers from the main track and not a change of the reviewers. Among the papers with no author reply, whose ratings the discussion did not move, $\hat{\sigma}_e$ lies within 0.09 of the main-track value in 2022 to 2025, but that population is selected on low and unanimous ratings, so the comparison does not settle the size of the convergence effect. The slope of $\log \hat{\rho}$ against $\log N$ is -0.15 with a standard error of 0.03 over 2017 to 2025 and -0.10 with a standard error of 0.04 over 2018 to 2025, and with the withdrawn papers included it is -0.12 and -0.06 with the same standard errors. Under A4 and (F.7) at $\rho \approx 0.35$ these slopes correspond to δ between about 0.1 and 0.25, and the value of 0.08 that the load per reviewer implies lies at the low end of that range. With three reviews at the 2025 share, (F.1) gives $R_3 = 0.59$ and a correlation of 0.77, so the move from three to four reviews in 2021 restored about the reliability of 2018 and 2019. The coin-flip share of (F.5) depends on the acceptance rate as well as on the reliability, and at the 2025 acceptance rate in every year it rises from 0.26 in 2017 to 0.36 in 2025, within 0.04 of the values at each year's own rate, so the rise is mostly the loss of reliability. Four further limits apply. External counts put raised scores at 17% to 23% of reviews in 2024 and 2025, and no before-rebuttal ratings exist for the early years, so the size of the convergence effect on $\hat{\rho}$ is not known. A decision that uses the text of the reviews or the area chair's own examination recovers more of the quality than (F.2) states, so the gain is a lower bound under A1 to A3 in that respect. The rating scale changed in 2020 and 2022, and $\hat{\sigma}_q$ and $\hat{\sigma}_e$ both jump at 2022, so only $\hat{\rho}$ and its rank-based form are compared across the change. The year's mean panel size in (F.1) overstates the correlation $\sqrt{R_m}$ of a mix of panel sizes by under 0.01 in every year.

Table F.1: The signal share of a rating per year for three populations, with the rank-based estimate, the estimated standard deviations of quality and noise in rating points, the mean panel size m, the reliability of (F.1), the acceptance rate a, the selection gain of (F.2) in rating points next to the perfect-information gain, the share of the true top a accepted from (F.3), the coin-flip share of (F.5), and the two-panel disagreement rate. Papers with at least two ratings.

Population	Year	Papers	$\hat{\rho}$	Rank $\hat{\rho}$	$\hat{\sigma}_q$	$\hat{\sigma}_e$	m	R_m	a	Gain / perfect	Top a	Coin flip	Disagree
Main track	2017	490	0.563	0.561	1.19	1.05	3.08	0.80	0.40	1.02 / 1.14	0.82	0.26	0.20
	2018	922	0.396	0.402	0.97	1.20	3.02	0.66	0.37	0.82 / 1.00	0.75	0.34	0.25
	2019	1,419	0.450	0.455	1.00	1.11	3.05	0.71	0.35	0.89 / 1.05	0.76	0.30	0.23
	2020	2,213	0.346	0.371	0.79	1.09	3.04	0.62	0.31	0.71 / 0.90	0.70	0.34	0.25
	2021	2,594	0.339	0.347	0.82	1.15	3.86	0.66	0.33	0.73 / 0.90	0.73	0.32	0.24
	2022	2,617	0.388	0.398	1.08	1.35	3.90	0.71	0.42	0.85 / 1.01	0.79	0.32	0.24
	2023	3,792	0.366	0.380	1.06	1.40	3.78	0.69	0.41	0.83 / 1.00	0.78	0.34	0.25
	2024	5,689	0.330	0.346	0.96	1.37	3.87	0.66	0.40	0.76 / 0.93	0.76	0.36	0.26
	2025	8,609	0.326	0.346	0.94	1.35	4.05	0.66	0.43	0.70 / 0.86	0.77	0.36	0.27
Main track and withdrawn	2017	490	0.563	0.561	1.19	1.05	3.08	0.80	0.40	1.02 / 1.14	0.82	0.26	0.20
	2018	970	0.413	0.421	1.00	1.19	3.02	0.68	0.37	0.85 / 1.03	0.75	0.33	0.25
	2019	1,564	0.464	0.471	1.03	1.10	3.05	0.72	0.35	0.92 / 1.08	0.77	0.29	0.22
	2020	2,560	0.362	0.393	0.82	1.09	3.04	0.63	0.31	0.74 / 0.93	0.71	0.33	0.25
	2021	2,978	0.364	0.374	0.86	1.14	3.86	0.69	0.33	0.78 / 0.94	0.74	0.31	0.23
	2022	3,375	0.422	0.433	1.15	1.35	3.90	0.74	0.42	0.92 / 1.07	0.80	0.30	0.23
	2023	4,916	0.385	0.405	1.11	1.40	3.78	0.70	0.41	0.87 / 1.04	0.78	0.33	0.25
	2024	7,242	0.363	0.382	1.04	1.37	3.86	0.69	0.40	0.84 / 1.01	0.77	0.34	0.25
	2025	11,510	0.376	0.402	1.05	1.35	4.06	0.71	0.43	0.81 / 0.96	0.79	0.33	0.24
No author reply	2017	48	0.652	0.648	1.31	0.96	3.17	0.86	0.40	1.16 / 1.26	0.85	0.21	0.17
	2018	166	0.383	0.383	0.83	1.06	3.02	0.65	0.37	0.69 / 0.86	0.74	0.35	0.26
	2019	233	0.289	0.248	0.64	1.00	3.03	0.55	0.35	0.50 / 0.67	0.69	0.41	0.29
	2020	349	0.234	0.231	0.57	1.02	3.02	0.48	0.31	0.45 / 0.64	0.64	0.42	0.30
	2021	406	0.271	0.262	0.65	1.06	3.89	0.59	0.33	0.54 / 0.71	0.70	0.37	0.27
	2022	704	0.317	0.299	0.86	1.27	3.88	0.64	0.42	0.65 / 0.81	0.76	0.37	0.27
	2023	1,122	0.251	0.257	0.78	1.35	3.75	0.56	0.41	0.55 / 0.73	0.73	0.44	0.30
	2024	1,329	0.079	0.029	0.39	1.34	3.84	0.25	0.40	0.19 / 0.38	0.59	0.70	0.40
	2025	2,815	0.191	0.200	0.64	1.32	4.07	0.49	0.43	0.41 / 0.59	0.71	0.49	0.33

Projection. Let N_y denote the submissions of year y, with $y_0 = 2025$ the last observed year, and let b denote the least-squares slope of $\log N_y$ on y over 2017 to 2026, which gives a doubling time

of $\log 2/b = 1.9$ years. The projection sets $N_y = N_{y_0}\, e^{b(y-y_0)}$ for $y = 2026$ to 2035 and

$$\log \rho_y = \log \hat{\rho}_{y_0} + \beta\,(\log N_y - \log N_{y_0}), \tag{F.9}$$

where β is the slope of $\log \hat{\rho}$ on $\log N$ fitted over 2018 to 2025 in the central scenario and over 2017 to 2025 in the second scenario, each with its standard error, so that a band from $\beta + \text{SE}$ to $\beta - \text{SE}$ covers the flattest and the steepest decline within a standard error. The panel size is either fixed at its 2025 mean or grows as $m_y = m_{y_0}(N_y/N_{y_0})^{0.11}$ with the exponent of reviews per paper in Table 1, and the acceptance rate stays at its 2025 value. For each year and scenario, (F.1), (F.3), (F.5), and the disagreement rate follow from ρ_y, m_y, and a, and Table F.2 lists them for 2026, 2030, and 2035. The projection assumes that the fitted slope, the acceptance rate, and the panel-size rule continue, and it describes the trajectory under the observed trend. Under the central scenario with a fixed panel, the coin-flip share passes 0.40 in 2030 and reaches 0.46 in 2035, with a band from 0.42 to 0.49, while the reliability decreases from 0.66 to 0.54, whereas the growing panel, which reaches 6 reviews in 2035, holds the reliability at 0.64 and the coin-flip share at 0.38. Under the slope fitted from 2017 with a fixed panel, the coin-flip share reaches one half in 2035, with a band from 0.47 to 0.54.

Table F.2: The projection of (F.9) by slope scenario and panel-size rule, with the band from the slope's standard error in parentheses, submissions N in thousands, and the true top a accepted, the coin-flip share, and the two-panel disagreement rate (whose band is within 0.02) at the 2025 acceptance rate.

Slope	Panel	Year	N	m	ρ (band)	R_m (band)	Top a	Coin flip (band)	Disagree
-0.10 ± 0.04 (from 2018)	fixed	2026	17k	4.05	0.31 (0.31 to 0.32)	0.65 (0.64 to 0.65)	0.77	0.37 (0.37 to 0.37)	0.27
		2030	72k	4.05	0.27 (0.25 to 0.29)	0.60 (0.58 to 0.62)	0.75	0.41 (0.39 to 0.43)	0.29
		2035	448k	4.05	0.22 (0.20 to 0.26)	0.54 (0.50 to 0.58)	0.73	0.46 (0.42 to 0.49)	0.31
	$\propto N^{0.11}$	2026	17k	4.22	0.31 (0.31 to 0.32)	0.66 (0.65 to 0.66)	0.77	0.36 (0.36 to 0.37)	0.27
		2030	72k	4.95	0.27 (0.25 to 0.29)	0.65 (0.63 to 0.67)	0.77	0.37 (0.36 to 0.39)	0.27
		2035	448k	6.05	0.22 (0.20 to 0.26)	0.64 (0.59 to 0.68)	0.76	0.38 (0.35 to 0.41)	0.28
-0.15 ± 0.03 (from 2017)	fixed	2026	17k	4.05	0.31 (0.30 to 0.31)	0.64 (0.64 to 0.65)	0.77	0.38 (0.37 to 0.38)	0.27
		2030	72k	4.05	0.25 (0.23 to 0.26)	0.57 (0.55 to 0.59)	0.74	0.43 (0.42 to 0.45)	0.30
		2035	448k	4.05	0.19 (0.16 to 0.21)	0.48 (0.44 to 0.52)	0.71	0.50 (0.47 to 0.54)	0.33
	$\propto N^{0.11}$	2026	17k	4.22	0.31 (0.30 to 0.31)	0.65 (0.65 to 0.66)	0.77	0.37 (0.37 to 0.37)	0.27
		2030	72k	4.95	0.25 (0.23 to 0.26)	0.62 (0.60 to 0.64)	0.76	0.40 (0.38 to 0.41)	0.28
		2035	448k	6.05	0.19 (0.16 to 0.21)	0.58 (0.54 to 0.62)	0.74	0.42 (0.40 to 0.45)	0.30