Atlas of Judgment: A Structured Analysis of Reviewer Reasoning in ICLR, 2018–2026

Shiro Takagi
shiro.takagi@unktok.com
Unktok
https://atlas-of-judgment.pages.dev/

Abstract

Peer-review scores summarize judgments but discard much of the stated reasoning behind them. We develop a model-assisted instrument for studying the written reasoning in nine years of public ICLR reviews and discussions. For each local evaluation it records the inspected object, the observation, the reasoning standard, and the conclusion separately, so that the interpreted link from observation to judgment can be inspected. A record of 52,460 submissions and 792,703 messages supports two overlapping readings: 410,586 units from 74,380 individual ICLR 2026 reviews, and 1,009,592 units from 50,861 discussion forums across 2018–2026. An induced taxonomy organizes twelve objects of scrutiny and twelve reasoning standards. Across the analyses, within-review sequencing and observation–rule–remedy coupling are weak under relevant null models. Object–standard associations recur, and novelty criticism stands out for its strong association with lower scores. Written judgment updates are uncommon, and their estimated frequency depends heavily on whether the denominator is a post-response unit, an exchange, or an original objection. Attention shifts toward empirical and computational scrutiny over time, while coarse novelty rules remain comparatively stable. The prevalence of a preselected vocabulary associated with LLM writing peaks in 2024, and co-review wording converges within the constant-form 2024–2026 window without a corresponding trend in size-adjusted object overlap. We report validation exercises, extraction failures, rejected interpretations, and reproducibility limits alongside the results. The findings are model-mediated readings of public text and observational associations; they do not reveal private reasoning, establish causal effects, or attribute AI use to individuals.

Preprint.

1 Introduction

Scientific communities rely on peer review to assess research claims and select work for publication. Reviewers are asked to judge whether a contribution is original, whether its evidence supports its claims, and whether its results matter. These demands do not reduce to a single test. A study can be technically convincing yet narrow in scope, or original yet insufficiently supported. Reviewers must decide which properties matter in the particular case and how competing considerations bear on an overall evaluation. Peer review is therefore both an institutional selection process and a setting in which scientific standards are articulated and applied.

Research on peer review examines both the reliability of its outcomes and the content of these evaluative practices. Repeated-committee experiments ask whether independently assigned reviewers would reach the same publication decisions [2, 8]. Content studies investigate the objects reviewers assess, the criteria they apply, and how positive and negative evaluations relate to ratings and editorial decisions [32, 34]. Computational studies represent local review propositions and induce taxonomies of evaluative aspects from text [13, 30]. These approaches make the reasons expressed in reviews an empirical subject in their own right. They also raise the question of how the content of an evaluation relates to agreement, disagreement, and later deliberation.

Agreement has several meanings in this setting. Two reviewers can assign the same score while criticizing different parts of a paper. They can discuss the same experiment but apply different standards to it, or agree that an experiment is missing while disagreeing about whether its absence undermines the contribution. Conversely, reviewers who disagree numerically can largely agree about the paper's strengths and weaknesses. An overall score does not separate these possibilities. Nor does a count of negative sentences tell us whether a criticism concerns an unsupported observation, an inappropriate comparison, or a different expectation of what the paper should contribute. To understand such differences we need to distinguish what is inspected, what is observed, which standard is applied, and what judgment follows.

Discussion adds another complication. An author may supply information about an experiment without changing a reviewer's view of its relevance, and a reviewer may acknowledge a response without withdrawing the original criticism. A final panel summary can emphasize a concern that was rare in the individual reviews. To study these exchanges, we need to connect a local evaluation with the review in which it appears, the replies that follow, and the final assessment. We also need to specify what counts as revisiting a concern or changing a judgment at each level. Scores, textual similarity, shared attention, and explicit revision each capture a different part of that process.

In this study we examine the evaluative logic expressed in reviews: how an observation about a paper is treated as grounds for a judgment, under which standard that connection is made, and how the resulting evaluation is defended or revised in discussion. We develop Atlas of Judgment, a model-assisted framework for representing these written justifications. The central measurement task is to make the connection from observation through standard to judgment inspectable within each local evaluation, including the premises that the instrument itself supplies. The inspected object locates what the evaluation concerns. Categorization, score comparisons, and longitudinal summaries all build on this representation, and the extracted connections remain model-mediated interpretations throughout.

To make these questions observable, we represent a locally interpretable evaluation as a *unit of written evaluation*. Figure 1 places the review passage in the center and its four interpreted components in numbered callouts: inspected object, observation, standard, and judgment. Solid leaders end on the corresponding wording; the dashed leader ends in empty space because the standard is supplied by the interpretation. The negative tag gives the judgment's valence. The figure is a schematic decomposition and has no numerical axes. In the passage, also discussed in Section 3, the reviewer says that a proposed mechanism resembles existing beam/pruning and adaptive-consistency techniques and describes the contribution as incremental. Here the *object* is the mechanism's novelty, the *observation* is its resemblance to existing techniques, the *standard* requires a new mechanism rather than repackaging, and the *judgment* is a negative evaluation of the contribution. The four are separate components of one evaluation. The standard is the instrument's interpretation of the link between observation and judgment, and the reviewer need not have written it as a sentence. Recording that distinction keeps an inferred premise visible as an inference instead of presenting it as explicit evidence.

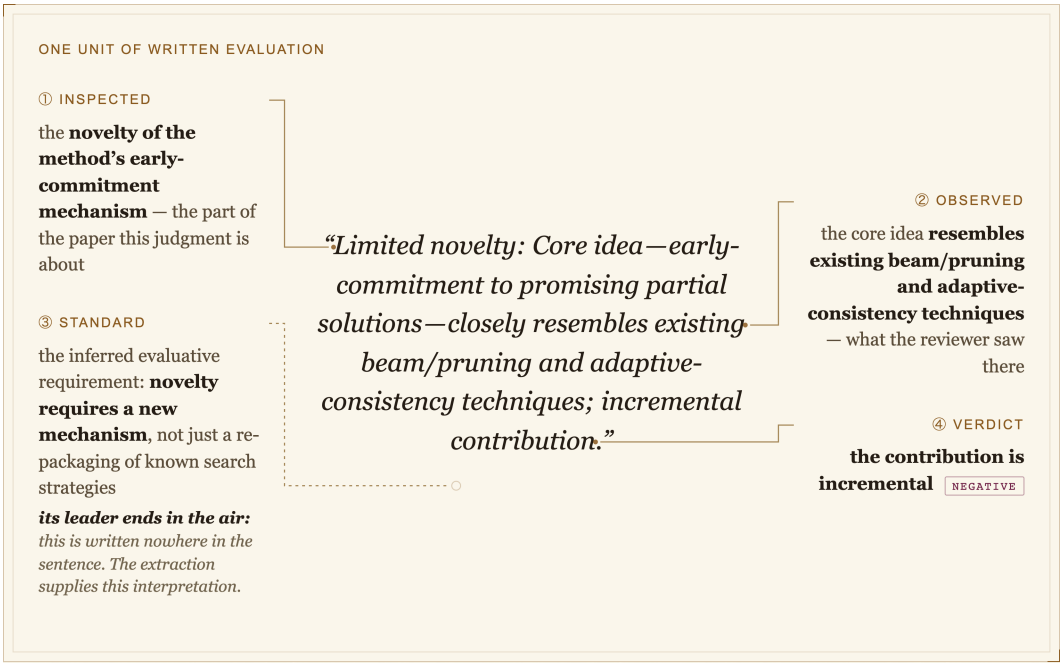


Figure 1: One review passage represented as a unit of written evaluation: inspected object, observation, standard, and judgment ("verdict"). Solid leaders connect components to wording in the passage. The dashed leader ends outside the quotation because the standard is inferred, not explicitly stated. Adapted from the project's introductory specimen; the passage is from an official ICLR 2026 review. The decomposition illustrates how the instrument interprets the passage; it does not validate that interpretation.

The project turns such readings into a basis for comparison across a corpus. It first collects reviews, author replies, discussion messages, and area-chair meta-reviews, which synthesize the reviewing panel and justify a recommendation or decision. Language models then write analytic memos and convert them into structured units that preserve the inspected object, observation, standard, judgment, and supporting information. Object descriptions and reasoning descriptions are grouped separately into an induced taxonomy of twelve objects and twelve standards. Each induced category summarizes recurring descriptions found through extraction and grouping, rather than a heading stipulated by the reviewing form. The resulting records let us count evaluations, compare reviewers, and follow groups of concerns through a discussion while retaining links to the underlying public material. The models analyze existing reviews; they are never asked to review the submitted paper afresh.

Several levels of analysis follow from this representation. Within a review, we examine which objects and standards occur together and whether the ordering of evaluations carries structure beyond the review's overall composition. Across reviewers of the same paper, we distinguish shared attention from agreement in stance and relate the content of criticism to differences in ratings. In discussion, we group an original reviewer's evaluations of the same object within a paper and follow the subsequent responses, which lets us tell apart a concern that was addressed, a written judgment that changed, and a reviewer who said nothing further. A group can contain several objections, so the grouping is not a validated one-to-one alignment of each original argument with its reply. At the panel and area-chair levels, we ask which concerns are shared, revisited, or emphasized in the final synthesis. Across years, we examine changes in attention and wording while checking how changing review forms, review length, and extraction affect those comparisons. These questions are connected, but they call for different observational units and comparison groups.

The empirical setting is the International Conference on Learning Representations (ICLR), a machine-learning conference whose OpenReview records link reviews to discussion and decisions. Because reviews of rejected as well as accepted submissions are public, criticism can be studied without restricting the corpus to accepted work. We use the 2018–2026 record, with a detailed review-level reading of 2026 and a forum-level reading of the full period; a forum is the message thread attached to one submission. The two readings trade finer within-review detail against a longer

historical window. They overlap in source material and are not independent samples. Section 3 gives the data counts, eligibility rules, and extraction procedures for each reading.

The analyses separate persistent associations from patterns that weaken under an appropriate comparison. Objects and standards are associated, and negative novelty evaluations have a particularly strong relationship with ratings and decisions. Within-review sequencing, by contrast, is weak when compared with rearrangements that preserve each review's composition, and observations, novelty rules, and suggested remedies do not form a tightly coupled chain. Visible judgment updates in discussion are uncommon, but their estimated frequency changes markedly depending on whether we count later units, exchanges, or original objections. Over time, attention shifts toward empirical and computational scrutiny, while coarse novelty rules remain comparatively stable. Wording converges among co-reviewers without a corresponding trend in object overlap once set size is accounted for. These results show why a change in language, a shared concern, a score association, and a revised judgment need to be measured separately.

The central contribution is an attempt to represent the evaluative logic within written reviews and make it available for empirical analysis. Keeping observations, standards, and judgments together allows us to ask how similar observations are evaluated under different requirements and how assessments are visibly revisited. The object and standard taxonomies give ways to aggregate these readings, and the rating and decision associations show how the readings relate to numerical and institutional summaries. The framework also connects local evaluation to the review, the exchange, the panel, and the meta-review. Section 2 distinguishes this representational task from prior analyses of criteria, polarity, and their statistical associations with outcomes.

How the findings should be read depends on the instrument. Throughout, "reasoning" means the reasoning represented in public text after model-mediated extraction; the instrument has no access to a reviewer's private thought. The taxonomy is one useful organization of those readings rather than a uniquely correct division of scientific judgment [36]. A richer representation can expose distinctions, but it also introduces extraction and categorization errors. We therefore report validation exercises, failed interpretations, null comparisons, and corrections alongside the substantive findings. The report proceeds from related work to data and measurement, then examines the organization of evaluations, the substance of criticism, associations with outcomes, discussion, ratings as measurements, and change over time. Section 10 documents the development record and public artifacts so that the reported reading can be inspected and alternative readings pursued.

2 Related work

2.1 Evaluation criteria and the content of judgment

Research on peer review has long treated the substance of evaluation as an empirical object. To find out which criteria were actually applied to qualitative articles, Savall et al. [32] content-analyzed referee reports and organized evaluative statements by nineteen critical points and ten quality criteria. Their cross-classification distinguishes what is assessed from the perspective used to assess it, a direct precedent for our object–standard distinction. Laudel [25] compares evaluations in plant biology and medieval history, connecting arguments to research practices, assessment objects, criteria, and evaluative content. The settings there span manuscript review, book review, and informal evaluation, and the differences among those institutions matter for interpretation. Both studies show the value of analyzing actual evaluations and the degree to which quality judgments depend on disciplinary practice.

Related studies distinguish agreement about a criterion from agreement in its application. Using interviews with fellowship evaluators, Guetzkow et al. [12] shows that originality can concern different features of research and carry different disciplinary meanings; the evidence there consists of evaluators' retrospective accounts rather than a corpus of referee reports. Hug and Ochsner [16] identifies heterogeneous patterns of agreement about grant-assessment criteria using a survey and latent-class analysis. Hug [15] instead models how criterion-specific ratings on actual grant-review forms relate to overall judgments, comparing logistic regression with simple decision trees. Endorsing a criterion, mentioning it in a report, and giving it weight in a particular judgment are three different things, and keeping them apart is essential when interpreting Atlas's category frequencies and rating associations.

Content analysis has also connected local evaluations to outcomes and participant roles. Bornmann et al. [3] codes the domain and polarity of comments on rejected chemistry manuscripts and relates them to where the manuscripts were later published. Bornmann et al. [4] uses a common coding scheme to compare formal referees with voluntary participants in public review. These designs anticipate our comparisons between evaluative content, institutional outcomes, and reviewing roles. Their selected journal samples and publication-based outcomes differ from our conference record, and in neither setting does the outcome, whether later journal placement or conference acceptance, directly measure scientific correctness.

Smith et al. [34] examines how reviewers commensurate epistemic criteria and how editors respond to divergent valuations. In their main measurement, review sentences receive criterion and polarity labels, which are aggregated into review-level proportions and related to ratings and editorial alignment using manuscript fixed-effects models. They find that salience differs from outcome association, with negative novelty and accuracy judgments particularly prominent in the latter. That design asks which criterion-specific evaluations accompany overall judgments, but it does not encode the observation, the applied standard, and the conclusion as connected components of each local argument. Atlas attempts that representational task, interpreted premises included, before classifying units or relating them to scores. The two studies therefore overlap in their content–outcome analyses while differing in what they try to represent within an evaluation. Because the ICLR archives overlap, similar novelty associations in the two studies are related observations and should not be read as independent replications on disjoint evidence.

2.2 Review propositions, aspects, and induced taxonomies

Computational research offers several representations of review content below the level of the whole review. PeerRead collects public reviews and studies relationships between aspect-specific assessments and recommendations [17]. An *aspect* here is a property of the paper under evaluation, such as novelty or methodological soundness. AMPERE segments reviews into local propositions and annotates their functions, including evaluation, request, fact, and reference, which shows why one sentence can hold more than one analytic unit [13]. Chakraborty et al. [5] combines aspect and sentiment classification to analyze ICLR reviews, including differences across decisions and disagreements with reviewers or chairs. Peer Review Analyze annotates paper-section correspondence, aspect, communicative purpose, and significance in separate layers [10]. Local units and multiple annotation axes are thus well-established measurement choices that this study inherits.

Lu et al. [30] offers a close precedent for our taxonomy construction. They use language-model-assisted extraction and expert organization to induce a bottom-up, multilevel taxonomy, then compare review focus across research communities. Their aspects include both objects such as methodology and evaluative properties such as novelty and significance. Their categories thus include evaluative criteria as well as subjects of evaluation. Atlas builds on this work by storing the inspected object, observation, inferential standard, and judgment as separate components, with object and reasoning descriptions grouped separately. The unit tries to retain how an observation is used to justify a judgment under a standard, and the categories then support comparisons across those interpreted links. This arrangement does not amount to a complete conceptual separation of object and criterion, and we make no claim of superior extraction accuracy. Human annotation in the prior corpora also supplies validation evidence that Atlas's model-based audits cannot replace.

2.3 Disagreement, weighting, and decision reproducibility

Content disagreement and decision inconsistency call for different measurements. ContraSciView identifies explicit contradictions between reviewers, starting from aspect and sentiment candidates and using human annotation to distinguish incompatible evaluations [22]. RevCI extends contradiction analysis to full-review context, evidence spans, explanations, and graded intensity [23]. Atlas's overlap analyses address a related but different possibility: reviewers may raise different concerns without contradicting one another. The comparisons ask different questions: whether reviewers inspect the same object, invoke similar standards, take the same stance, or make the same criticism. A low overlap score on its own tells us neither that the reviewers complemented one another productively nor that they disagreed in scientific judgment.

Work on grant review likewise separates the content of critiques from how reviewers translate it into scores. Pier et al. [31] combines ratings with coded strengths and weaknesses in a mock NIH review setting, distinguishing consistency within a reviewer from agreement across reviewers assessing the same applications. At the committee level, the NeurIPS consistency experiments independently assign submissions to two reviewing committees and compare their decisions [2, 8]. Those experiments observe alternative committee outcomes directly, whereas Atlas's rating variance decomposition and panel redraw estimate what a different panel might do under explicit statistical assumptions. Agreement in broad numerical scale with an experiment provides context, but it does not validate the simulation or the extracted reasoning representation.

2.4 Responses, discussion, and change over time

Review–response research already examines how a criticism is interpreted and addressed. APE annotates argument spans and links criticisms to responding arguments [7]. DISAPERE combines review discourse, aspects, and polarity with rebuttal actions, stance, and sentence-level correspondence [19]. These resources show that an evaluation's surface form and its interpretation as a request or question need not coincide. Atlas takes up the same need to represent such relations and adds a descriptive question: whether the recorded stance toward an original objection is later maintained, strengthened, weakened, or reversed. In the objection-lifecycle analysis the tracking key is the forum, reviewer, and object, and several complaints can share one key. That grouping is coarser than human annotation of specific argument spans and their response links. Throughout, we treat the existence of a response, the addressing of a concern, and a reviewer's explicit change of judgment as distinct outcomes.

Studies of numerical updates also analyze response content and interaction. Gao et al. [9] combines annotated weaknesses and response properties with prediction of post-rebuttal scores; the authors' additional note stresses that movement toward other reviewers does not distinguish useful learning from conformity. Huang et al. [14] studies rebuttals through score changes and social-network structure. More recent preprints classify ICLR criticism and response strategies in relation to score revisions [18], or induce exchange features and compare the predictive information in initial reviews, responses, and panel discussion [29]. Using language models to structure human review exchanges is therefore itself a close precedent. Atlas builds on these studies through its chosen representation and its tracking of original objections and written judgment through the exchange.

A complementary line of work uses interventions to study proposed mechanisms. Randomized studies test whether prompting the initially most positive or most negative reviewer to begin discussion induces herding [35], and whether exposure to an initially flawed manuscript anchors later evaluation after correction [28]. Neither finds significant evidence for the targeted effect in its setting. These results do not rule out all social influence or persistence, but they illustrate why an observational pattern of convergence or non-updating does not identify its cause. For longitudinal context, Paper Copilot analyzes multiple years of ICLR decisions and ratings and tracks discussion-period updates [44]. A multiyear archive alone does not distinguish Atlas from this prior work. Atlas's historical comparisons concern the content assigned to the structured units, and they are subject to changes in forms, source visibility, and extraction.

2.5 The position and limits of the present study

These literatures establish several dimensions of evaluative practice, including criteria, stance, argument function, disagreement, and response. Atlas focuses on one representational task: recording an observation about an object, the standard under which that observation is treated as grounds for an assessment, and the judgment connected to it. Its taxonomy and outcome comparisons operate on that interpreted evaluative link. The same representation is then used at the level of the unit, the review, the reviewer–object objection group, the exchange, the panel, the meta-review, and the year, with the matching resolution specified for each analysis. The contribution is an attempt to make written evaluative logic inspectable and analyzable, together with a documented empirical account that keeps weak results and corrections in view. Representing a link does not establish that it was extracted accurately, that the reviewer's inference is scientifically valid, or that private cognition has been observed. Alternative representations and independent validation are still needed to assess which distinctions the instrument preserves and which it loses.

3 Data and measurement

3.1 Public record, coverage, and institutional context

The empirical record is organized around a submission and its discussion thread, which OpenReview calls a forum. A forum can contain initial reviews, author replies, subsequent reviewer comments, and an area chair's meta-review summarizing the assessment. These messages play different roles: a review expresses an individual evaluation, while a meta-review helps explain the panel's final recommendation. A submission, a message, and an extracted evaluation are therefore different counting units. This section first establishes which public messages are available and then describes how their text becomes a representation of evaluation that can be counted and compared. The venue census below uses the public metadata directly; the content analyses that follow depend on the additional reading and structuring steps described here.

The collection contains 52,460 ICLR submissions and 792,703 messages across 2018–2026. Message normalization identifies 199,031 official reviews, 507,543 official comments, 40,821 decisions, 30,108 meta-reviews, 8,235 withdrawals, 5,563 public comments, 980 desk rejections, and 422 other comments, leaving no unclassified messages in the recorded snapshot. Different analyses use different subsets of this record, since a collected submission need not have an eligible review, a successful extraction, or an observable decision.

Collection queries OpenReview API v2 for each venue configuration and falls back to v1 when the submission identifier is unavailable. In this run, 2018–2023 resolve to v1 and 2024–2026 to v2. Cursor pagination retrieves replies together with submissions, persists the cursor so that a run can resume, and retries failures up to seven times with exponential backoff. The raw SQLite layer is written during collection and thereafter opened read-only. Normalization adds message types, roles, timestamps, and provenance without interpreting the review text. Notes and public signatures are serialized as provided; the collection pipeline does not mask them.

Figure 2 summarizes how the size of the venue and the reviewing process changed across years. Every horizontal axis runs from 2018 to 2026. Bar height in the upper-left panel is the number of submissions; the upper-middle line is the mean number of official reviews per paper; the upper-right line is acceptance as a percentage of decided, non-withdrawn papers; and the lower-left line is median official-review length in words. The panels have separate vertical scales, and the absent 2019 decision value reflects missing data. The recorded submission counts are 935, 1,419, 2,213, 2,594, 2,617, 3,792, 7,404, 11,672, and 19,814, respectively. Growth on this scale changes both statistical precision and the institutional context of reviewing. Median review length peaks at 532 words in 2022, falls to 410 in 2025, and is 412 in 2026. Reviews per paper rise from approximately three to four around 2021. Acceptance rates use decided, non-withdrawn papers. The public snapshot has no recoverable 2019 decisions, and withdrawal counts are comparable only from 2025 onward, because earlier years delisted withdrawals.

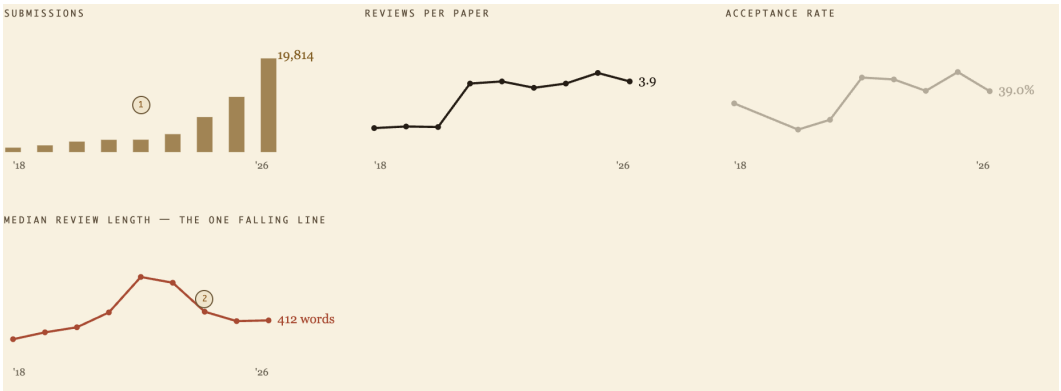


Figure 2: Venue census from the public record: submissions, acceptance among decided non-withdrawn papers, reviews per paper, and review length. These counts require no language-model reading. Missing 2019 decisions and changes in withdrawal visibility limit comparisons.

Figure 3 shows the sequence of public submission, review, discussion, withdrawal, and decision events during the ICLR 2026 cycle. The horizontal axis is calendar date, from 1 September 2025 to 8 March 2026, with month labels as ticks. Five vertically separated rows show submissions, official reviews, official discussion comments, withdrawals, and decisions, in that order. Each series counts records created on each UTC day: the submission row counts papers, and the other rows count the corresponding OpenReview notes. Normalized height above a row's baseline is $\sqrt{n_{kd}/\max_d n_{kd}}$, where n_{kd} is that event type's daily count. Each row therefore has its own zero and peak, and a peak is a daily maximum rather than a cumulative total; the review peak, for example, is 15,147 notes on 1 November 2025. The plotted daily arrays sum to 19,814 submissions, 75,859 reviews, 180,142 discussion comments, 5,140 withdrawals, and 14,175 decisions. The full-snapshot totals reported below include four additional discussion comments and 76 additional withdrawals that fall outside the plotted window.

In the ICLR 2026 cycle, submissions crest in September 2025; 75,859 reviews then arrive in a concentrated burst, 180,146 discussion comments follow, 5,216 submissions are withdrawn, and 14,175 decisions are recorded. Withdrawals rise after reviews become available and before decisions. The ordering alone does not tell us why authors withdrew.

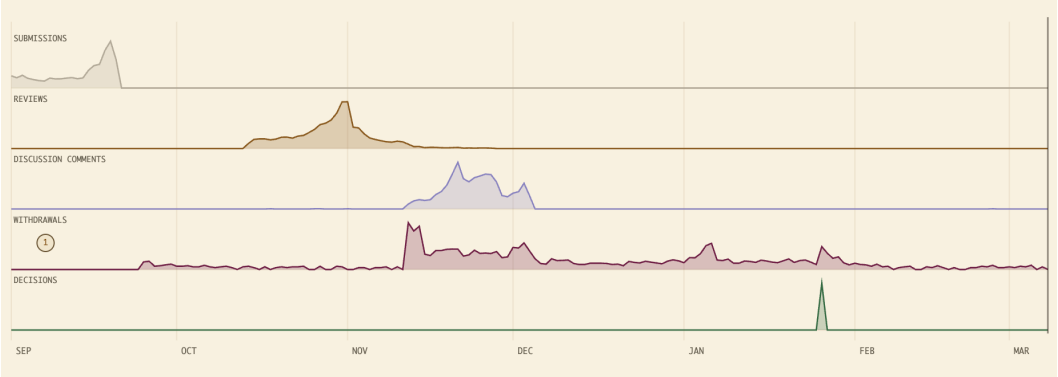


Figure 3: The ICLR 2026 public-event calendar. Each event series is square-root scaled to its own peak, so heights are not comparable across series as absolute counts. The timing of withdrawals is descriptive only.

Collection completeness is checked against counts returned by the same collection run rather than by an independent later reconciliation. Edits or removals made after collection cannot be recovered from this snapshot. The original analysis deliberately excludes 2017 because its official venue configuration was not verified in the same way.

3.2 A unit of written evaluation

The content analysis does not assign one label to an entire review. Instead it extracts local evaluations that can be understood in context, so that one review can contribute several units about different parts of a paper. An object is the aspect being inspected, such as the scope of the experiments. An observation says what the reviewer takes to be present or absent. A standard specifies why that observation counts for or against the work, and a judgment records the resulting assessment. Keeping these components separate lets the same object be assessed under different standards, instead of treating a topic label as a complete explanation of a criticism.

Each unit records four components: the object inspected, the observation made about it, the reasoning standard connecting observation to evaluation, and the judgment. Additional fields preserve valence, suggested improvement, evidence references, support status, confidence, and missing inferential links. Valence has five values: negative, positive, mixed, conditional, and uncertain.

Returning to the passage in Figure 1, the review states that an early-commitment mechanism resembles existing beam/pruning and adaptive-consistency techniques and calls the contribution incremental. The inspected object is the mechanism's novelty; the observation is the similarity to existing techniques; the inferred standard requires a new mechanism rather than repackaging; and the judgment is negative. The quoted sentence never states the standard. Representing it is an interpretive act by the instrument, and the support-status field exists to record that fact.

Table 1 compares the two extraction tracks in its columns: one reads individual reviews and the other whole forums. Rows give eligible input documents, completed and excluded outputs, completion percentages, and extracted evaluation units. Completion is calculated as completed outputs divided by eligible inputs within the same column. Both tracks produce multiple evaluation units from one document, so document counts and unit counts describe different stages, and "review-level" does not mean one unit per review.

Table 1: Two extraction tracks. The eligible-input denominator differs from the collected-record denominator.

	Review-level, 2026	Forum-level, 2018–26
Inputs to structuring	75,859 reviews	51,813 forums
Strict-complete outputs	74,380	50,861
Excluded outputs	1,479	952
Completion	98.05%	98.16%
Extracted units	410,586	1,009,592

The two tracks answer different questions. The review-level track supplies line anchors and fine within-document positions. The forum-level track reads a whole discussion and records temporal position, update trigger, and judgment change. Neither is a subset of the other, and their absolute levels are not interchangeable measurements. The 1,420,178 total units sum two readings of overlapping source material, so the figure does not count that many independent reviews or independently observed judgments.

3.3 Three layers and two memo protocols

Separating the pipeline into layers makes the transformation of evidence visible. The raw layer preserves what was publicly written. The memo layer asks a language model to interpret the evaluations in that record and explain their connections in prose. The structured layer then converts those memos into the common fields defined above. This division gives later analyses a consistent representation, but it also opens a route for error propagation: an observation omitted or reinterpreted in a memo cannot be recovered by making the final output conform to a schema. A protocol, in what follows, is the set of reading instructions and context supplied to the model; it introduces no new source of reviews.

Figure 4 is a processing diagram in which rectangles are saved data or processing stages and arrows give the direction of transformation. The two middle columns separate the Layered review reading from the Direct forum reading, which rejoin at the common taxonomy and aggregation stages. Numbers inside boxes give input or completed-output counts, and dollar amounts give recorded API charges. Neither horizontal spacing nor box area measures a statistical quantity. Layer II uses `deepseek-v4-flash`, with thinking disabled, temperature 0.4, and no output seed. A persistent budget ledger, request records, and file locking support resumable runs. The prompts frame the task as metascientific analysis, require evidence line identifiers, and instruct the model neither to re-review the submission nor to invent a taxonomy.

For 2026, the Layered protocol produces initial-blind, trajectory, and paper-synthesis memos, 151,193 memos in all. The isolated-review reading withholds other reviews, author replies, and a separately supplied final decision; its information limits are specified below. The nine-year Direct protocol produces one memo per whole forum, 51,813 memos in all. A blind same-model comparison on 100 forums preferred Layered 69–31. Direct was nevertheless retained for the longer historical window on cost and error-propagation grounds. That choice is a pragmatic trade-off and should not be read as evidence that the two protocols are equally accurate.

Only completed `initial_blind` memos feed the Compact/Qwen review-level track, with one review memo per request; trajectory and paper-synthesis memos are not inputs to that track. The memo total above covers all stages of the Full Layered run and is therefore larger than the input count for structuring. The name `initial_blind` refers to an isolated-review protocol and carries no guarantee that the public review is an untouched historical pre-rebuttal version.

Layer III uses `qwen3.7-flash` through the DashScope realtime interface, with thinking disabled and strict JSON-schema output. It reads only the Layer II memo, never the raw review. Neither



Figure 4: Measurement pipeline: collection and provenance normalization; Layered and Direct analytic memos; schema-constrained structuring; extraction into unit databases; and shared taxonomy assignment. The two branches share source material and must not be pooled as independent observations.

temperature nor an output seed is explicitly set, so provider defaults apply. Manifests pin protocol and schema hashes. The review-level schema contains the unit fields described above. The forum schema additionally records an `initial`, `post_author_response`, `meta_assessment`, or `cross_reviewer` position, together with a strengthened, weakened, reversed, clarified, or unobserved judgment change.

The separation of outcomes from extraction has a limited, stage-specific meaning. Layer III receives memo packets and identifiers without separately supplied rating or decision fields, and its instructions prohibit reproducing those outcomes in semantic fields. That restriction leaves open whether the upstream reading was free of outcome information. The Layer II isolated-review prompt receives the normalized public review-form response: the normalizer renders all scalar form fields, and the prompt explicitly warns that numerical ratings may be embedded in that text. The Direct protocol reads the whole public forum, including later discussion and the reasoning in meta-reviews or decisions. The public code also warns that complete edit histories are unavailable [38]. These analyses therefore describe the public written record and its association with ratings and decisions. They are not prospectively blinded predictions from a verified pre-rebuttal snapshot.

3.4 Validation, retry policy, and excluded records

Successful extraction has an operational meaning. The output must contain the required fields and refer to evidence in the permitted input, and an interrupted or malformed response cannot enter the unit database as if it were complete. Retrying with a larger token budget gives an output more room

to finish. These checks concern the usability and internal consistency of a record. Whether the model found every evaluation or interpreted its substance correctly is a separate question, taken up by the instrument checks below.

Completion requires local schema and evidence-reference checks, not merely a successful provider response. The Direct first pass leaves 7,166 truncated outputs. A 24,000-token retry targets 7,270 eligible rows and recovers 6,223; a 32,000-token retry of 959 rows recovers another 812. These selections are operational retry cohorts rather than a single nesting of identically defined failure counts. The final strict-complete coverage is 50,861 of 51,813 forums.

Table 2: The final 952 excluded forum-level records, by local or provider failure.

Failure	Rows	Provider JSON available
Invalid evidence reference	709	Yes
Schema validation	65	Yes
Other strict/local validation	39	Yes
Parse or truncation	107	No
Provider policy refusal	32	No

Table 2 lists failure types as rows, excluded-forum counts in the middle column, and availability of the provider's JSON response in the final column. The counts add up to the excluded-forum denominator. "Yes" means a response was available, not that it passed the evidence or schema check. Semantic-validation failures with valid JSON are excluded from API retry selection, and a larger budget leaves the 32 policy refusals unchanged. The merge process checks manifest ancestry, refuses unfinished retry runs, snapshots the target database before modification, and preserves first-attempt provider evidence. A weaker JSON-available coverage of 99.73% exists but is not the denominator used for the strict unit database. The 1,479 failed review-level inputs and 952 failed forums may be selected with respect to content; we did not measure that dependence.

3.5 Inducing and assigning the taxonomy

The unit descriptions start as free text. To compare evaluations across reviews and years, the analysis needs recurring categories, but without asking the reader model to choose those categories in advance. An embedding represents a description as a numerical vector, so that similarly worded or semantically related descriptions can lie near one another. Object descriptions and reasoning descriptions are processed separately because the question under discussion and the standard used to judge it need not coincide. UMAP reduces the dimensionality of those vectors for clustering, and HDBSCAN then looks for dense groups of similar descriptions, leaving other descriptions ungrouped as noise. The resulting groups are candidates for interpretation rather than already named scientific constructs.

Object descriptions and reasoning descriptions are embedded separately with normalized `BAAI/bge-small-en-v1.5` vectors. A 12,000-unit sample from the 2026 track, seed 7, is reduced with UMAP: 15 dimensions, 30 neighbors, minimum distance 0, cosine metric, random state 7. HDBSCAN uses minimum cluster size 60, minimum samples 10, and EOM extraction. This yields 36 object clusters with 37.6% noise and 17 reasoning clusters with 61.5% noise.

Figure 5 uses boxes and arrows to show sampling, separate object and reasoning embeddings, dimensionality reduction, clustering, manual merging, and nearest-centroid assignment. The object and reasoning noise percentages refer to the clustering sample, whereas assignment applies the resulting categories to both complete unit tracks. The diagram's rows are processing steps, not positions on coordinate axes. Cluster exemplars are merged and named into twelve objects and twelve standards; two heterogeneous clusters are excluded when the category centroids are built. Each centroid is the normalized mean of its member-cluster embeddings. Every unit in both tracks is assigned to its nearest category by cosine, and the similarity is retained. In the 2026 track, 94.5% of object assignments and 87.9% of standard assignments exceed 0.75. Re-running the seeded local pipeline reproduces the sampled cluster memberships, although this local reproducibility does not make the earlier language-model readings deterministic.

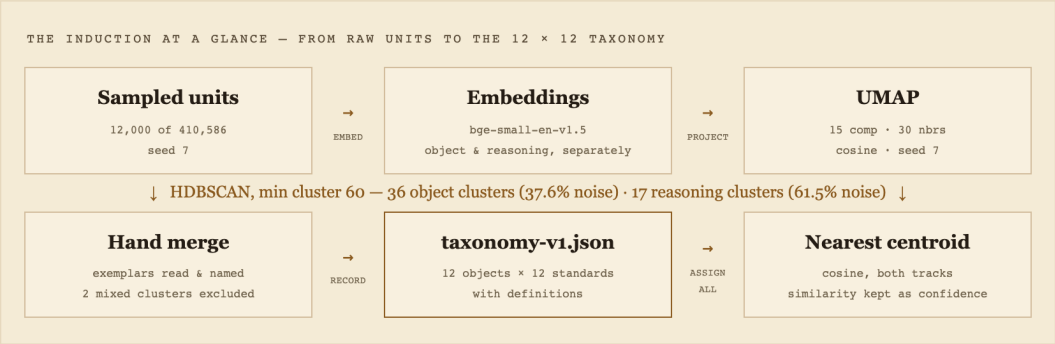


Figure 5: Taxonomy construction and assignment. Clusterable examples supply the category skeleton, after which all units receive a nearest-centroid assignment. The substantial clustering-noise fractions are part of the instrument's limitations.

Table 3: Objects of scrutiny in the induced taxonomy. Codes are used only to keep the count matrix legible.

Code	Category	Working definition
O1	Empirical scope & generalizability	Breadth and representativeness of empirical evaluation: dataset scale, benchmark coverage, generalization to other models/settings, robustness scope, attack/threat coverage.
O2	Baselines & ablations	Choice, fairness, and completeness of baseline comparisons and ablation studies.
O3	Theory, assumptions & guarantees	Theoretical assumptions, formal guarantees, proofs, and the loss/objective's formal properties.
O4	Method design rationale	Justification of design choices: architecture, reward/objective design, training procedure, mechanism claims.
O5	Compute cost & scalability	Training/inference computational cost, latency, scalability analysis.
O6	Clarity & presentation	Writing quality, figures, notation, exposition and structure.
O7	Novelty & contribution	Novelty of the contribution and its articulation relative to existing methods.
O8	Related work & citations	Coverage and accuracy of citations and positioning within prior literature.
O9	Statistical rigor & metrics	Statistical reporting, significance, choice and validity of evaluation metrics including LLM-as-judge setups.
O10	Robustness & sensitivity	Hyperparameter sensitivity, noise robustness, failure-mode analysis.
O11	Reproducibility & code	Code availability and sufficiency of detail for independent replication.
O12	Problem framing & motivation	Framing, motivation, and significance of the problem being addressed.

Table 4: Reasoning standards in the induced taxonomy. Codes are used only to keep the count matrix legible.

Code	Category	Working definition
S1	Novelty standard	Contribution must show conceptual/epistemic difference from prior art, not mere recombination.
S2	Claim-evidence scope matching	Generality of claims must be matched by breadth of empirical demonstration.
S3	Fair comparison norm	Baselines must be comparable in tuning, compute, and recency for the comparison to be meaningful.

Code	Category	Working definition
S4	Statistical identifiability	Without variance/significance reporting, signal cannot be distinguished from noise.
S5	Confound / alternative explanation	Results could be explained by leakage, memorization, weak protocols, or other artifacts.
S6	Design justification demand	Design choices need mechanistic grounding or comparison against plausible alternatives.
S7	Robustness norm	A practical method should be stable across hyperparameters and settings.
S8	Practical cost-benefit	Gains must be weighed against computational/engineering cost for practical utility.
S9	Measurement construct validity	Metrics must actually measure what they claim to measure.
S10	Reproducibility norm	Enough detail must be disclosed for independent reimplementation and verification.
S11	Presentation as trust signal	Polish signals care; presentation flaws undermine trust in the work's rigor.
S12	Merit recognition	Positive appraisal: breadth/soundness of evidence or contribution acknowledged as convincing.

Tables 3 and 4 each have one category per row and three columns: its short code, name, and working definition. The first table defines the subject inspected and the second the evaluative requirement. The codes identify the row and column categories in Table 8; they are identifiers, not scores or ranks. Unlike the clustering step, nearest-centroid assignment always chooses a category, even for descriptions that would not form a dense cluster on their own. Giving every unit a label does not resolve the uncertainty represented by clustering noise. Retaining the similarity keeps weak matches inspectable, while the fixed categories make comparisons across years possible.

The same centroids are transferred from the 2026 review-level protocol to the historical forum protocol, and mean assignment similarity stays near 0.81 across years. This is a useful transfer diagnostic, but it is neither a direct accuracy estimate nor proof that model error is constant across eras. A high cosine similarity measures proximity to the chosen representation; it is not calibrated confidence that the category is correct.

3.6 Instrument checks and provenance of interpretation

No single check validates every part of this pipeline. Evidence linkage concerns the connection to the record; support status concerns whether the reviewer stated a connection or the memo supplied it; and construct association concerns whether extracted categories correspond to independently recorded distinctions on the review form. For the separate argument-form classifier, accuracy measures overall agreement with its reference labels, while macro-F1 balances label-specific precision and recall so that the most frequent form does not dominate the summary. Each check answers a narrower question than whether the instrument has recovered a reviewer's reasoning completely.

The measurement checks ask two questions: how much of the represented reasoning is inferred by the pipeline, and whether the resulting object categories align with distinctions made on the venue's own review form. We first tabulate support-status flags separately by extraction track, object, and year. The external construct comparison then holds overall rating fixed when contrasting sub-scores, so that a general tendency for unfavorable reviews to score everything lower cannot by itself produce the apparent category alignment. Both checks address recorded inference and construct association; neither measures complete extraction accuracy.

Figure 6 shows how the share of reasoning inferred by the memo varies across objects and years. Rows are the twelve objects of scrutiny and columns are years from 2018 to 2026. Each cell is the fraction of forum-track units in that object–year group whose support status is memo-inferred. Darker purple indicates a larger fraction, with the same color rule across the matrix, so a cell is a within-group proportion rather than a count or a model-accuracy estimate. Support-status fields distinguish a reviewer-explicit statement from a memo inference or a mixture of the two. In the review-level track these shares are 19.8%, 77.1%, and 3.1%; in the forum track they are 57.4%,

40.8%, and 1.7%, with rounding. Layered synthesis and direct evidence-citation prompts thus produce sharply different levels, and much of the apparent reasoning is an explicit interpretation of the record by the pipeline.

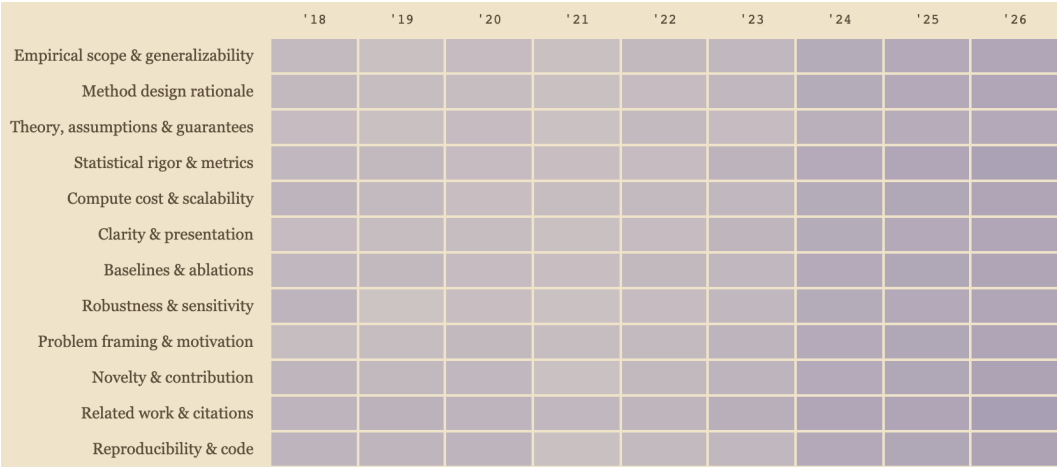


Figure 6: Memo-inferred share by object and year in the forum-level track. Differences across cells qualify the interpretation of downstream comparisons: a single pipeline does not imply uniform extraction error.

Figure 7 compares the objects identified in criticism with the venue's soundness, presentation, and contribution sub-scores. Objects of criticism are rows and the soundness, presentation, and contribution sub-scores are columns. A cell is the rating-stratified mean sub-score with criticism minus that without it, measured in points on the 1–4 sub-score scale. Purple marks negative differences and green positive ones, with stronger color for larger magnitude, and the printed number gives the signed value. The comparison is made within overall-rating strata before aggregation. This external construct check joins 74,380 rated 2026 reviews to ICLR soundness, presentation, and contribution sub-scores. At each overall rating, mean sub-scores are compared between reviews with and without a negative unit on the object; strata with fewer than 30 reviews on either side are dropped, and the differences are prevalence-weighted. Clarity criticism is associated most strongly with lower presentation (-0.35 on the 1–4 sub-score scale), and novelty criticism with lower contribution (-0.12). Theory criticism is associated with a larger presentation difference (-0.07) than soundness difference (-0.02). The check supports alignment with the venue's distinctions, but it cannot separate a reviewer's categorization habits from extraction conventions, and it does not validate every inferred reasoning field.

The six-form rhetoric classifier has its own interpretive codebook. Surface markers guide sampling; an analyst LLM reads 600 units and develops eight codes, later merged into six. The classifier combines local embeddings and marker features in a logistic model and reaches five-fold cross-validated accuracy 0.685 and macro-F1 0.666. Precedent has the weakest F1, 0.53. Marker suggestions agree with the assigned labels in only 61% of flagged examples; a cheaper second annotator agrees in 42% and is excluded. These labels are reference annotations for scaling one reading, not independent human ground truth.

Three independent machine readers additionally audit 60 raw novelty denials each without receiving the taxonomy. They recover the main clause families and identify finer false-priority and unlocated-contribution readings. This supports the intelligibility of the categories, but it remains a model-based audit and cannot replace a representative human validation of extraction recall, category correctness, or temporal-change detection.

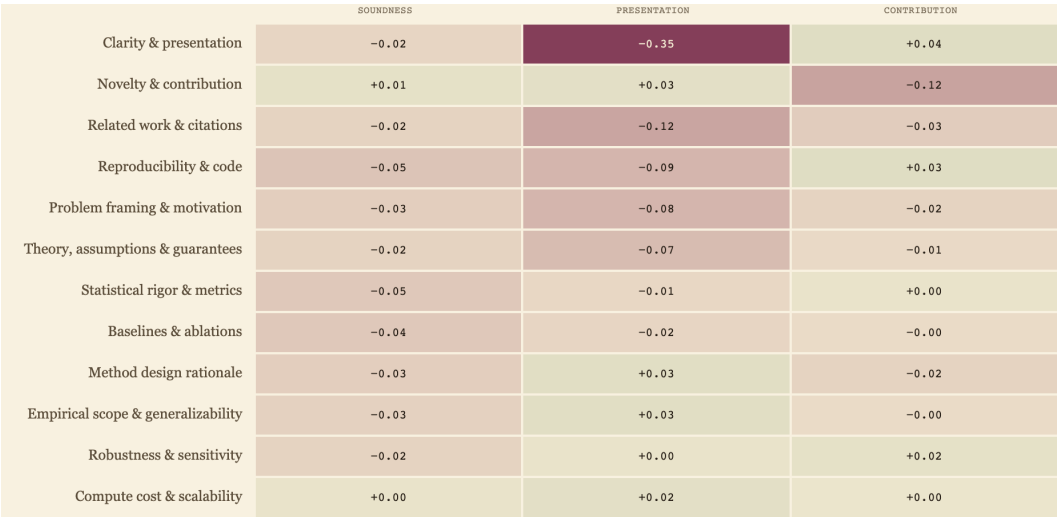


Figure 7: Rating-adjusted construct check against ICLR 2026 sub-scores. Cells are differences between reviews criticizing an object and reviews without that criticism at the same overall rating. The differences are associations conditional on rating, not causal effects of mentioning an objection.

4 The organization of written evaluation

4.1 Object, standard, and valence are related but distinct

We begin with the composition of evaluation itself: the subjects a reviewer addresses, the grounds on which those subjects are judged, and the direction of the judgment. These descriptive distributions supply the reference frequencies for later comparisons. A common object will occur in many combinations even without any special affinity for a particular standard, and a common standard will likewise appear across many objects. The denominators therefore have to stay visible before a large flow can be read as a distinctive relationship. Here the counting unit is an extracted evaluation, so several evaluations from one review can contribute to the same cell.

Before any association with ratings or decisions is examined, we need to describe what the instrument represents. Counting objects alone would identify the subjects of attention but would not distinguish the standard applied or the judgment reached, so we cross-tabulate the 2026 units by object, standard, and valence. Marginal counts measure the prevalence of each label; counts divided by each object–standard cell's total measure how often that particular pairing ends negatively. The flow diagram and the severity map are two summaries of this one joint table.

The 2026 review-level track contains 410,586 units. Its largest object is empirical scope (21.8%), followed by method design (14.6%) and theory (13.9%). Negative units account for 71.8%; the remaining mix is positive 13.7%, conditional 9.4%, uncertain 4.3%, and mixed 0.7%, with rounding. These percentages describe the extracted units. They cannot be read directly as the sentiment distribution of raw reviews, especially since extraction favors explicit criticism and can omit brief praise.

Figure 8 shows how extracted evaluations connect objects of scrutiny, reasoning standards, and valences. It has three categorical columns: objects on the left, reasoning standards in the middle, and valences on the right. Ribbons connect the labels assigned to the same extracted units, and ribbon and node widths are proportional to unit counts; adjacent percentages are shares of all 410,586 units. The vertical ordering is a layout choice and carries no numerical meaning. Valence colors distinguish negative, positive, conditional, uncertain, and mixed evaluations. The largest flow is empirical scope evaluated through design justification (4.8% of all units). Novelty is concentrated: its leading standard accounts for 67% of novelty units, and its top two account for 80%. Statistical rigor is spread across many standards; its leading one holds only 14%, and eight are needed to reach 80%. Related work frequently invokes the novelty standard (50% of its units; 3.3 times that standard's base share), whereas framing disproportionately invokes merit recognition. Within this

representation, novelty is governed by a relatively concentrated set of requirements, while statistical criticism draws on a wider repertoire.

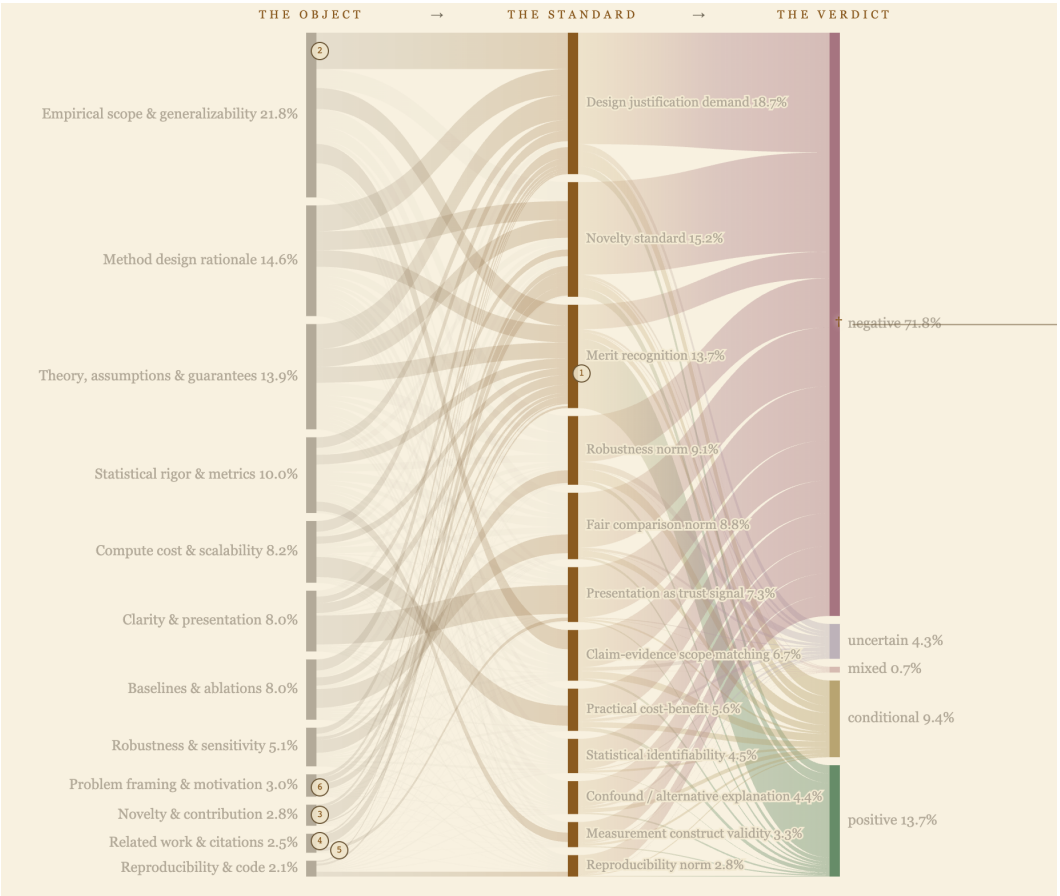


Figure 8: Object–standard–valence flows in the 2026 review-level corpus. Width represents unit count. The diagram describes the instrument's extracted evaluations; the negative share is sensitive to unequal recall of criticism and brief praise.

Figure 9 shows how the prevalence of negative evaluation differs among object–standard combinations. It places objects in rows and standards in columns. Color is the percentage of units in that particular object–standard cell that are negative, with darker purple meaning a higher negative share. The hatched, low-count cell is withheld from interpretation. Unlike the preceding flow widths, cell color is normalized by the cell's own total, so equal colors can represent very different numbers of units. Positive evaluations concentrate in merit recognition: 66.9% of all positive units use that standard. Outside its column, all 131 object–standard cells with at least 40 units are majority-negative. The single low-count cell, novelty under reproducibility, contains 36 units and is not interpreted. Related work remains 55% negative even under merit recognition; other objects in that column are 14–35% negative. The highest negative shares include related work under presentation-as-trust (95.9%), clarity under the same standard (94.7%), and reproducibility under statistical identifiability (94.4%). In the extracted record, then, positive judgment is concentrated in recognizing a contribution's merits, whereas criticism is spread across many more evaluative standards. Applying one extractor uniformly does not by itself guarantee equal error across these cells.

To ask whether the standard label adds a distinction beyond the object, we next normalize the object–standard counts within each object. If the labels were independent, every row would reproduce the same corpus-wide standard mix, so departures identify which standards are particularly associated with each object. The historical comparison then asks whether this conditional organization changes, which is a different question from whether the objects receiving attention change. It subtracts the early-era conditional shares from the late-era shares and compares those differences with relabeling noise. The answer also bears on the later analyses that pool years.

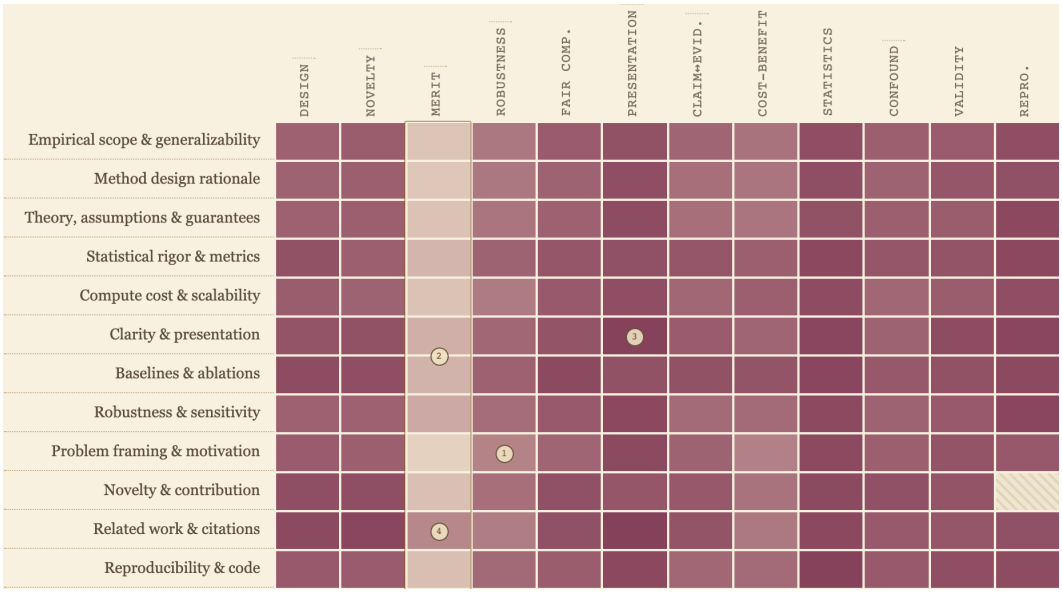


Figure 9: Negative share for each object–standard pairing. The merit-recognition column differs sharply from other standards. Levels and between-cell differences remain subject to extraction bias.

Figure 10 shows which reasoning standards are used for each object of scrutiny. Rows again identify objects and columns identify standards, but color now represents the percentage assigned to a standard among all units on that row's object. Every row sums to 100%; darker brown marks a larger conditional share. The matrix thus compares $P(\text{standard} \mid \text{object})$, rather than negative share or absolute cell count. Five objects are most often evaluated by their namesake standards: novelty, clarity, robustness, compute cost, and reproducibility. Design justification leads for empirical scope, method design, theory, and statistical rigor; novelty leads for related work; merit recognition leads for framing. Some of the diagonal concentration is definitional, because object and standard names overlap. The off-diagonal mass is the stronger evidence that the standard adds information beyond the object.

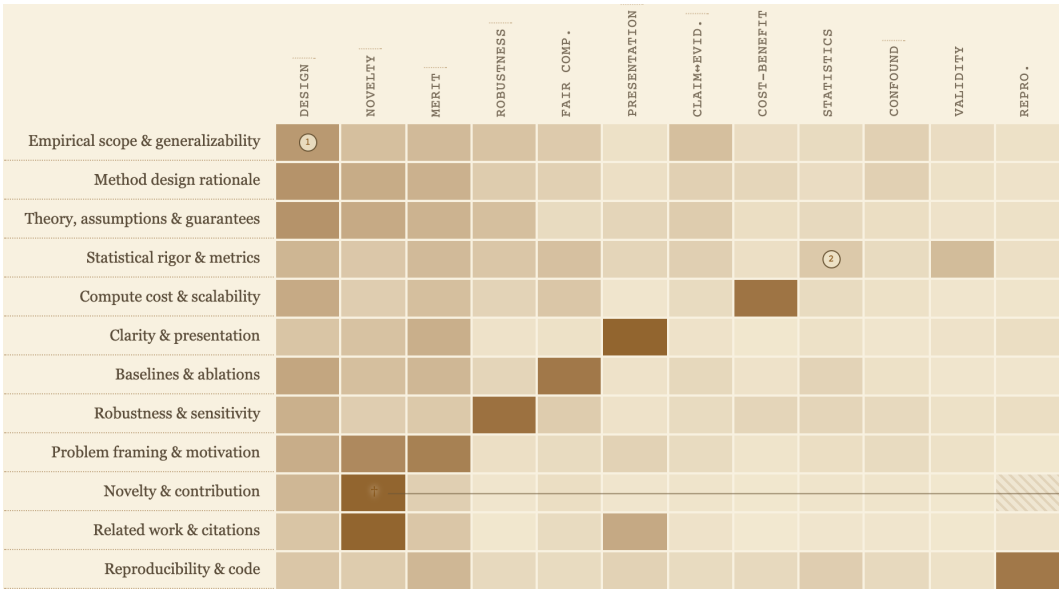


Figure 10: Conditional distribution of standards within each object. The matrix is row-normalized, so it measures the organization of evaluation rather than each object's total prevalence.

For the era comparison, the shuffle removes the association between era and standard while keeping object membership fixed. The maximum change across the entire matrix is recorded for each shuffle, so the reference is the largest fluctuation anywhere in the matrix rather than a separate threshold for each cell. A cell that stays below this reference has shown no conspicuous change under the procedure, which still leaves room for smaller shifts and for changes that the categories or the reading protocol cannot resolve.

Figure 11 compares the standards used for each object between the early and late periods. It retains object rows and standard columns and subtracts the 2018–2019 conditional share from the 2025–2026 share in each cell. Values are percentage-point changes: brown denotes an increase and violet a decrease. Cells inside the maximum-change permutation threshold are deemphasized. The colors therefore encode historical differences in within-object shares, not the late-era levels themselves. The conditional matrix changes little between 2018–2019 and 2025–2026 in the forum track. The null relabels era within object; across 50 shuffles, the 95th percentile of the largest absolute cell change is 4.05 percentage points, and 138 of the 144 cells remain within this threshold. The six larger changes include cost-benefit gaining within compute cost (25.9% to 34.6%), reproducibility's own norm gaining (28.9% to 33.8%), and novelty receding within framing and related work. Merit recognition rises for framing, partly because positive framing units became more common. Valence-stratified checks retain a $+6.3\,\mathrm{pp}$ gain among praise units and the retreat of novelty among both critical and positive units. Fifty permutations and model-mediated labels limit the resolution of this stability claim, which is not an equivalence test against all possible changes.

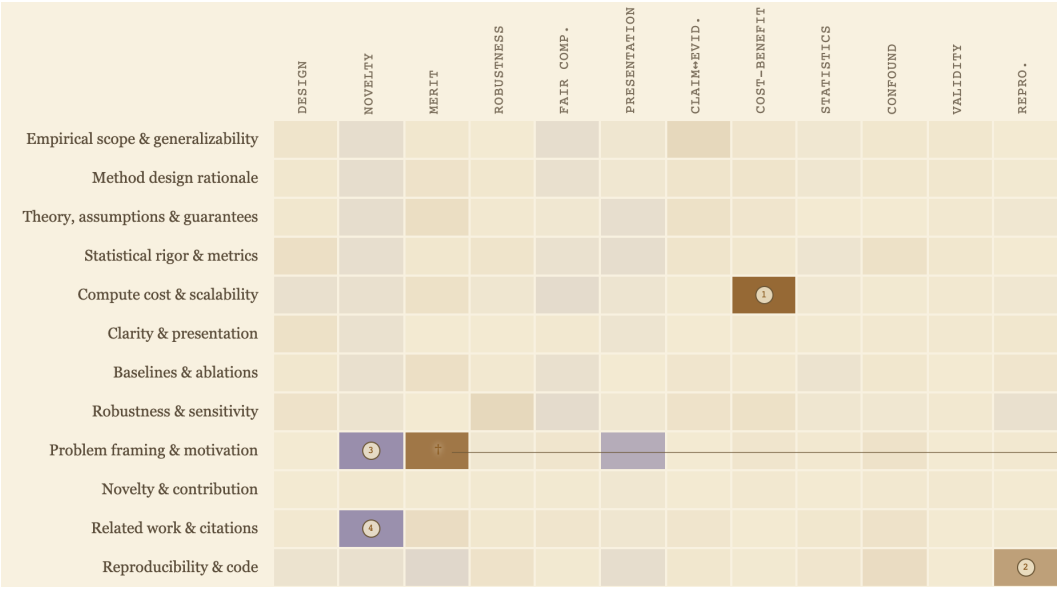


Figure 11: Historical changes in $P(\text{standard} \mid \text{object})$. Cells within the maximum-change shuffle threshold are deemphasized. Both eras use the forum-level protocol; levels should not be directly equated with the separate review-level matrix.

4.2 Argument forms distinguish different evaluative acts

Object and standard do not specify how an argument is made: a reviewer can invoke a requirement, identify a failed check, or balance strengths against weaknesses. To distinguish those acts, the classifier described in Section 3 assigns a form to each forum-level unit. We tabulate form shares within each standard, compare them with the overall form mix, and then tabulate valence within forms. The two tabulations ask whether standards have distinctive argumentative expressions and whether praise is expressed differently from criticism.

The argument-form labels concern the move a passage makes. Norm invocation appeals to what a paper should provide; inability to verify says that the available information does not permit a check; precedent compares the work with an earlier example; scope limitation restricts where a claim holds; rival explanation offers another account of the apparent result; and on-balance weighing brings

considerations together into an assessment. These labels describe the representation of an argument. They do not imply that a reviewer consciously chose one of the forms, and a form label describes a different feature from the object or the direction of the evaluation.

Figure 12 compares the argument forms through which different reasoning standards are expressed. It has one row per reasoning standard plus a pooled-corpus reference row. Each horizontal stacked bar is normalized to all units using that standard, and its six colored segments give the percentages assigned to the six argument forms. From left to right, the segments are norm invocation, inability to verify, appeal to precedent, scope limitation, rival explanation, and on-balance weighing. Segment length measures a within-row share, and the right-hand labels identify the leading form and its percentage; bar length says nothing about how often the standard occurs overall. Across these six forms, the corpus shares are approximately 30%, 6%, 8%, 13%, 6%, and 37%, respectively. Reproducibility is 69% norm invocation. Statistical identifiability is 43% inability to verify, more than seven times that form's overall share. The novelty standard uses precedent in 29% of its units, although on-balance weighing remains its largest form.

Valence is also differentiated. The five forms other than on-balance weighing are 88–97% negative. On-balance units are 42% positive, and 78.8% of all praise takes that form. The represented standards thus differ in the argumentative acts through which they are expressed: demanding that something be supplied and reporting that it cannot be verified are distinct patterns of criticism. These distinctions rest on a codebook inferred from the material, with the classification error reported in Section 3, so exact cell proportions should not be treated as error-free measurements.

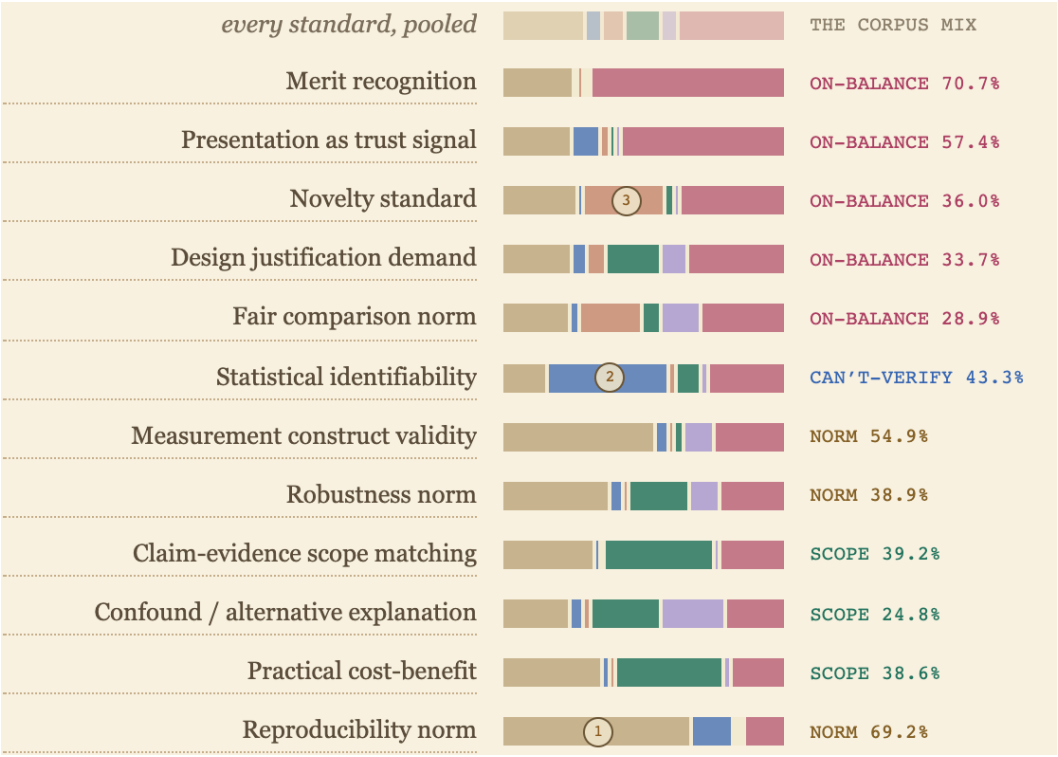


Figure 12: Argument-form composition by reasoning standard in all 1,009,592 forum-level units. Departures from the corpus composition describe the relation between a standard and how its argument is expressed. Classifier macro-F1 is 0.666.

4.3 Sequential structure is weak after conditioning on composition

The preceding tables discard order. To examine whether written evaluation has a sequential organization, we compare the observed arrangement of units with arrangements that keep their labels but remove their order. This separates adjacency beyond composition from the simpler fact that a review contains many units on the same subject. Position within the whole document and position within

its weaknesses field address different questions, because the review form can impose an order of its own.

For these sequence comparisons, a transition is a pair of neighboring units in their written order. A shuffle keeps the units in a review but rearranges them, giving a reference for the adjacency expected from that review's own contents. Transition lift divides the observed frequency by this reference frequency: values above unity indicate more adjacency than expected, and values below unity indicate less. The comparison is needed because repeated attention to theory, for example, can produce theory-to-theory transitions even when unit order is otherwise arbitrary. The analysis concerns the written sequence after extraction; it does not reconstruct the chronology of the reviewing process.

Three position analyses use different denominators. Argument-form transitions compare neighboring units against a shuffle within each review, preserving its composition. Whole-review arcs describe the order of fields and units in the document. The weakness-list analysis restricts position to line-anchored criticism within one field, reducing the mechanical effect of the form's section order. None directly observes the temporal order in which a reviewer thought or decided.

Figure 13 shows how argument forms follow one another and which forms open or close reviews. The six central rows give the current argument form and the six central columns give the following form. A cell's printed lift is observed transition frequency divided by its within-review shuffle expectation; reddish cells exceed unity and greenish cells fall below it. The outside OPENS and CLOSES columns instead give percentages of reviews whose first or last unit has that form. Those marginal percentages and the central dimensionless ratios use different scales. Across 183,717 initial reviews and 829,684 units, form-transition lifts stay between 0.80 and 1.18. Rival explanation and scope each follow themselves at 1.18; other forms modestly avoid being followed by on-balance. On-balance opens 38.5% of reviews despite a 30.5% share in this initial-review subset. Closing shares are close to the subset's baselines. Review sections and summary conventions could plausibly account for much of the residual structure. Given the small effects and the classifier error, the result is best read as a failure to detect strong sequential regularity rather than as evidence that reviewer thought lacks structure.

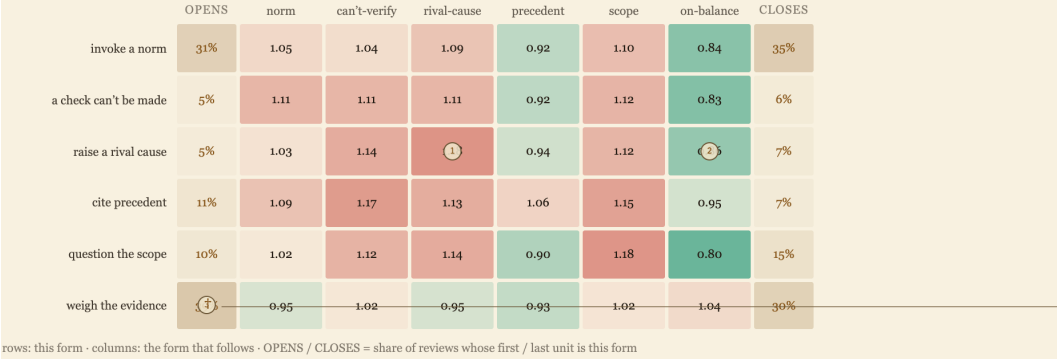


Figure 13: Argument-form transition lift relative to shuffling each review's own units. The appropriate baseline conditions on within-review composition; an unconditional corpus baseline would exaggerate sequence effects.

For the document arc, we tabulate the objects in the first and last unit of each eligible review and compare their frequencies with the overall object mix. Separately, valence shares are tabulated by relative position in the review. The analysis after that is restricted to units anchored inside the weaknesses field, which removes the mechanically ordered strengths and questions sections from the position comparison.

Figure 14 traces which objects open and close reviews and how evaluative valence changes through the written document. The left panel has objects in rows and a logarithmic horizontal axis for opening or closing frequency divided by the object's base frequency. Circles mark openings, squares closings, and the vertical unity line marks the base rate. The right panel divides a review's ordered units into fifths on the horizontal axis and plots the percentage of units in each fifth that are negative, positive, or hedged; the hedged series combines the remaining valences. Positions measure document order, not elapsed reviewing time. Whole reviews open disproportionately on framing (1.86 times its base rate) and novelty (1.57), and close on clarity (1.81), related work, and reproducibil-

ity. The negative share reaches approximately 82% in the middle. The form's order of fields, with strengths first, weaknesses in the middle, and questions last, strongly shapes this arc. The arc uses the 72,828 reviews with at least three units.

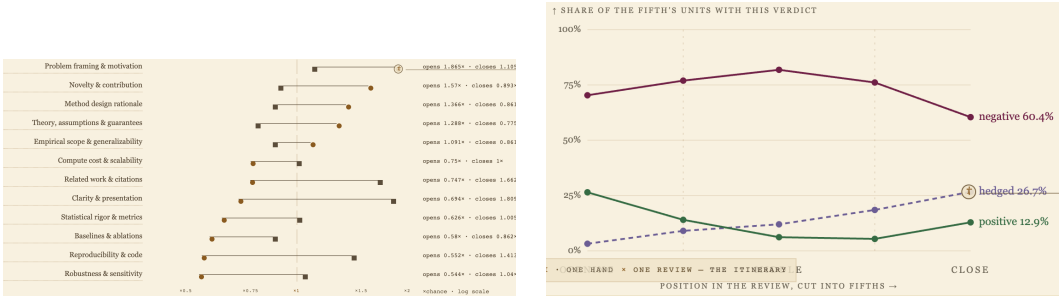


Figure 14: Left: Objects in opening and closing positions, expressed relative to their base rates. **Right:** Valence through the review. This describes the order of written sections, which can impose the observed arc.

Figure 15 compares where criticism of different objects appears within the weaknesses field and whether it is also raised as a question. It has one row per object. Within each row, the horizontal ridge runs from the beginning to the end of the weaknesses field and its height is the distribution of that object's faults across normalized depth; the vertical mark gives median position. The next column reports first-position frequency relative to chance, and the final column gives the percentage raised in the questions field. Ridges are normalized within object, so their heights do not compare absolute fault counts across objects. The line-anchoring stage identifies 192,085 units across all review-form fields. Position and valence summaries then use units anchored in the weaknesses field, whereas the lead-off comparison is restricted to the 30,869 reviews with at least two such units. A fault is defined by its location in that field rather than by negative valence alone, although 93.4% are negative. The published display reports approximately 3.9 faults per review and a 25% first-position share, but these approximate annotations do not establish a reconciled mean for the lead-off eligibility subset. Every nonempty list has a first position, so the useful comparison is which object takes that slot beyond its base share: novelty at 1.6 times chance, framing 1.4, and method design 1.2, against approximately 0.7 for statistics, robustness, and reproducibility. Novelty is raised in the questions field in 9% of cases, robustness in 29%. Even within one form field, then, the subjects of criticism are ordered unevenly: novelty and framing are more likely to lead than statistical or reproducibility concerns. The pattern may reflect the paper's own ordering as well as reviewing conventions, and it should not be read as a measured hierarchy of importance.

Figure 16 compares the adjacency and repetition of objects within written reviews. In the left panel, nodes are objects arranged around a circle, node size reflects unit frequency, connecting curves show elevated adjacency between different objects, and extra rings indicate self-repetition above the corpus baseline. Circular position is a layout choice. The right panel puts objects in rows and self-transition lift on a horizontal axis: open circles use the corpus baseline and filled dots use the within-review shuffle, with unity as reference. A connecting segment shows the change of baseline, not an uncertainty interval. Across 336,206 adjacent pairs, a corpus baseline gives theory a dwell lift of 1.9 and novelty 0.72. Clarity–related-work, compute–robustness, and statistics–reproducibility pairs are also elevated. Under within-review shuffling, dwell lifts compress to 0.94–1.29; novelty reverses to 1.20, and the clearest surviving adjacency is baselines and ablations. Empirical scope, despite being the largest object, has no strong specific companion, because its pairings lie near its own large base rate. At review level, 41% of reviews mentioning theory return to it, while 96.5% of those mentioning novelty mention it only once. The ring mainly shows which subjects share a review; it does not describe a grammar of their order.

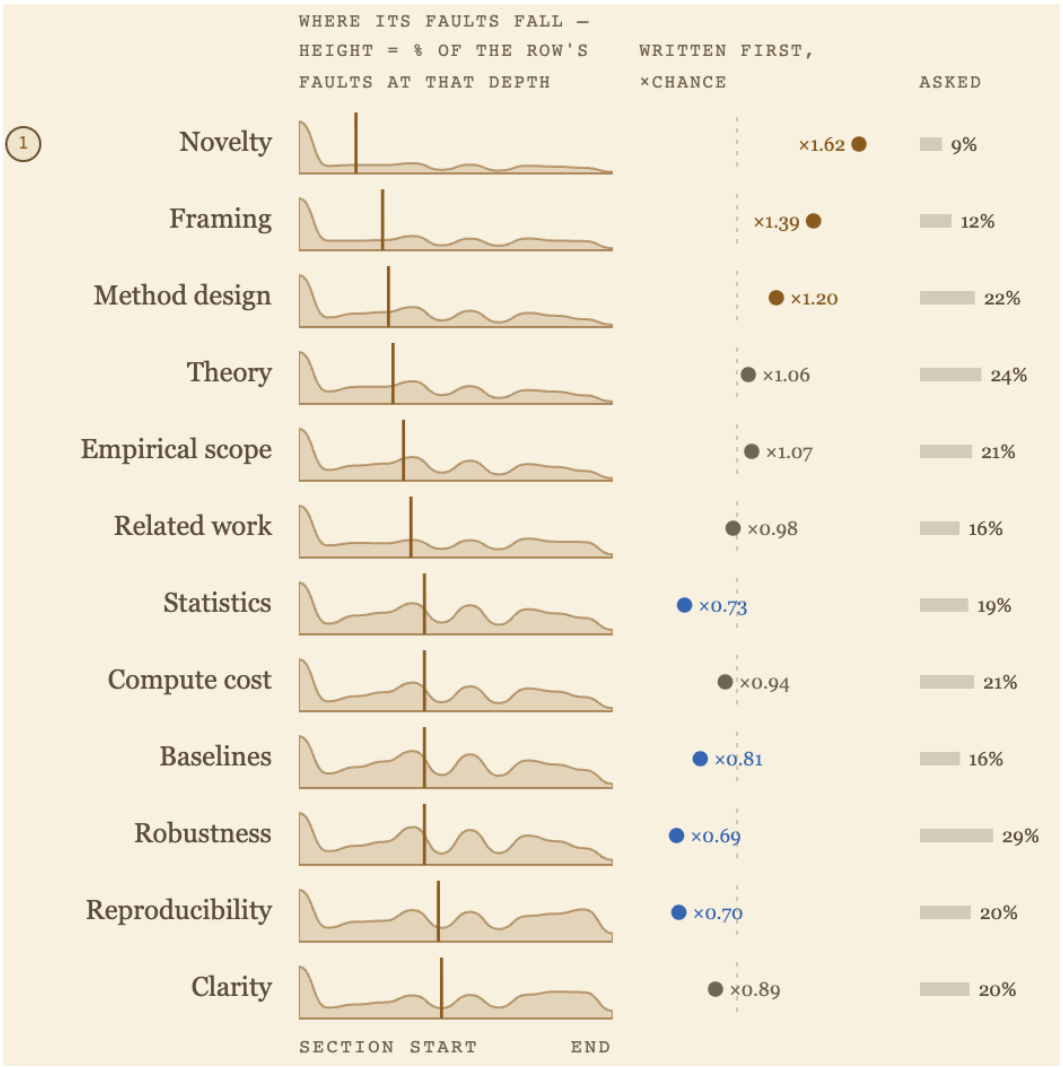


Figure 15: Positions within the weaknesses field, with lead-off lift and the separate share raised in questions. Relative density shapes are scaled within each object and cannot be compared as absolute counts.

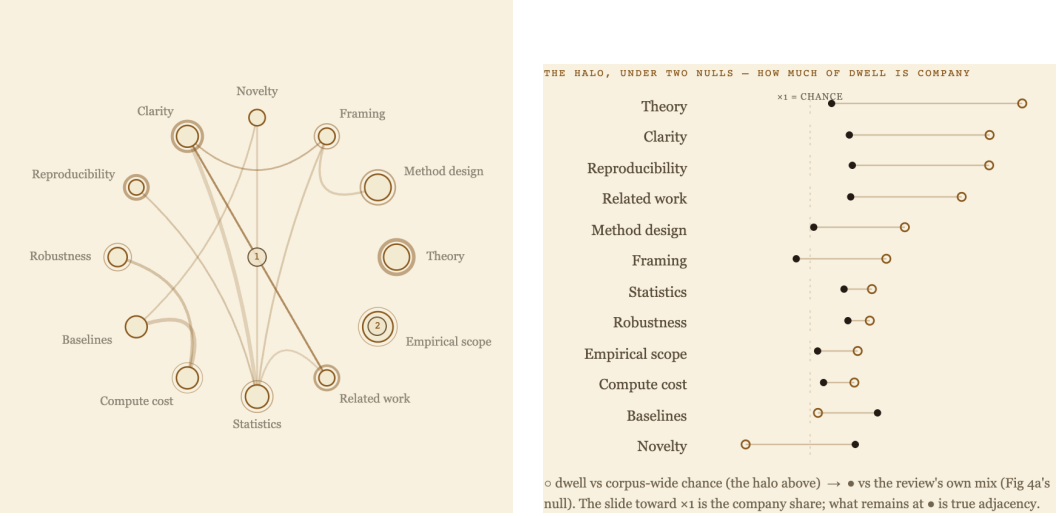


Figure 16: Left: Object adjacency against a corpus-wide baseline. **Right:** The stricter within-review comparison, which removes much of the apparent sequence structure.

4.4 A review's rating is readable from a short, weighted tally

Sequence need not be strongly organized for an opening to convey most of a document's evaluative information. We therefore ask how much of a review's own rating can be recovered from successively longer prefixes, and whether that information comes from valence counts, the objects discussed, or fuller wording. Comparing those feature sets separates a largely negative-versus-positive tally from any additional contribution of topic, and restricting the prefix analysis to long reviews separates saturation from simply reaching the end of the document.

A prefix is the opening portion of a review, represented by its first several extracted units. For each prefix length, the prediction model uses counts of the objects and valences encountered so far and does not read the later units. Logistic prediction maps those counts to a probability of the high-rating category. AUC measures how well that probability ranks high-rated reviews above lower-rated ones across possible thresholds; chance ranking corresponds to one half and perfect ranking to unity. Cross-validation evaluates the ranking on held-out reviews, so a larger feature set must improve prediction beyond the reviews used to fit it. The exercise asks how the written evaluations encode their author's own summary rating. That rating is not an independent assessment of the paper's quality.

Figure 17 compares how much of a reviewer's rating can be recovered from review prefixes and different representations of their content. It has four panels. Upper left plots prefix length in units against held-out AUC, with chance at 0.5 and the full-review result as reference. Upper right uses unit index horizontally and the AUC gain from adding that unit vertically, restricted to long reviews. Lower left compares feature sets by gains above chance, with stacked valence and object contributions and a separate full-text bar. Lower right has objects in rows and the magnitude of a negative unit's fitted weight relative to the average negative weight on the horizontal axis; these weight ratios are not AUC values. A logistic model predicts whether a 2026 review's own rating is at least 6 from the first k units, using object and valence counts. Three-fold cross-validation ranges from one to fifteen units. Because the median review contains only five units, a second analysis uses the 2,649 reviews with at least ten units, so that the exhaustion of short reviews is not mistaken for early informational saturation.

One unit gives AUC 0.69; the full coarse profile reaches approximately 0.779. Valence counts account for $+0.260$ of the $+0.279$ improvement over chance; object information adds $+0.019$, and -0.006 in the long-review subset. A weighted tally improves on the unweighted tally by only 0.009 AUC. Its per-unit weights nevertheless differ substantially: a novelty negative carries about 2.4 times the average negative weight, and a robustness negative about 0.4. The ordering correlates with the within-paper association in Section 6 at Spearman $\rho = 0.90$. The two analyses are related observational designs on the same reviewing process and should not be counted as independent replications.

The full-text baseline uses TF–IDF, a word-based representation that gives more weight to terms frequent in a review but less common across the collection. It supplies wording that the object–valence counts discard. The comparison therefore asks whether that discarded language helps recover the same rating. It does not treat the predictive features and the earlier interpretation of the review as independent measurements.

TF–IDF over full review text reaches AUC 0.822, 0.043 above the coarse profile. At the level of this representation, valence carries most of the coarse model's rating correspondence, while wording adds further information. The full-text model provides one predictive baseline; its performance need not exhaust the information available in the text. In long reviews, gains become negligible after approximately the ninth unit. The interpretation concerns redundancy in the written document and agreement with its own summary rating; it does not locate the moment at which the reviewer decided.

Held-out prediction does not resolve the upstream information constraint. The unit representation is produced from public review-form text whose numeric fields may have been visible to the memo model. The same-review rating task therefore measures correspondence with a summary that may already have informed extraction, and its AUC cannot show that score-blinded reading recovers an independently withheld judgment.

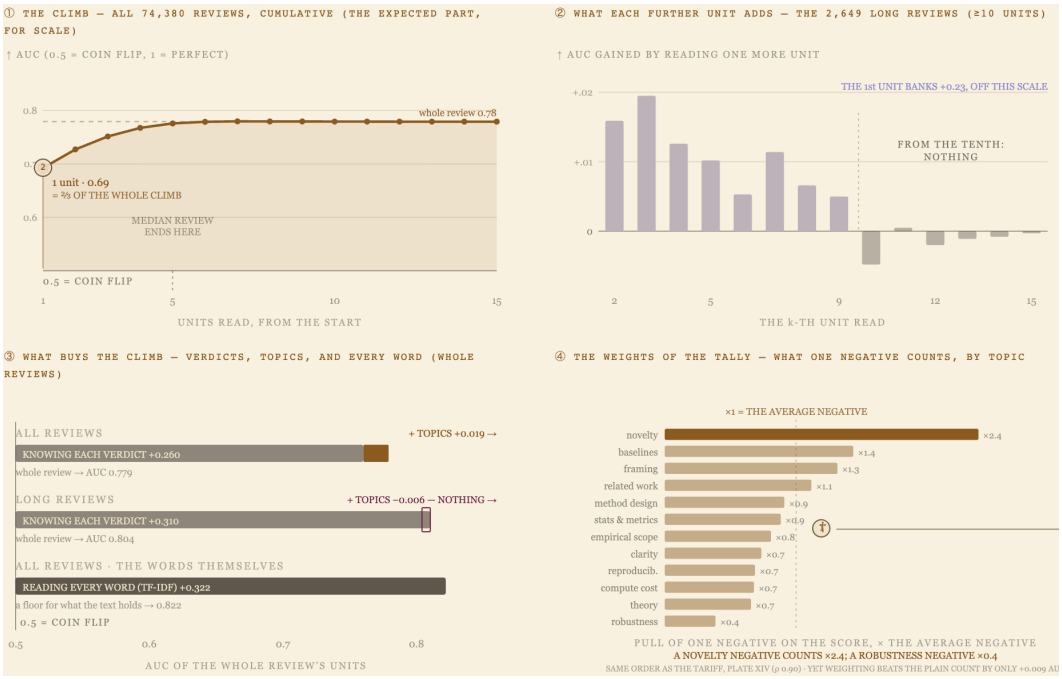


Figure 17: Rating prediction from successive units, the incremental contribution of units in long reviews, comparison of valence, topic, and full-text features, and descriptive per-object weights. The target is the same reviewer's rating, evaluated by three-fold cross-validation.

5 What criticism claims and what it requests

5.1 Grounds, stated rules, and remedies

An object label such as novelty says what a criticism targets, but not which deficiency is alleged or what would satisfy the reviewer. To examine those distinctions, we group the text in the unit's observation, reasoning, and improvement fields into grounds, rules, and remedies. The groupings let us describe the content of a criticism, compare the work requested across objects, and ask whether a given observation predicts a particular rule and remedy. The last question matters because a diagram that links three fields can look like a tightly constrained argument even when most combinations remain possible.

Three terms carry the analysis. A "ground" is the alleged state of the work recorded as an observation, a "rule" is the principle by which that state is evaluated, and a "remedy" is the proposed response recorded in the improvement field. Not all three need appear in the reviewer's own words. Describing a contribution as an assembly is an observation, requiring an additional contribution is a rule, and asking for a further experiment would be a remedy. The example only illustrates the roles of the fields. Whether these elements travel together, and whether an objection determines a unique repair, is what the analysis tests.

The forum track contains 698,507 negative units across the twelve objects. We examine their observation, reasoning, and improvement fields. The three fields are not an independently observed syllogism, because the model may have inferred the rule that connects the other two. Normative-sentence mining picks out the smaller subset in which a rule is explicitly stated, estimated at 26–40% of negative units depending on the object. The mining pattern embodies an assumption about how norms are phrased, so these shares omit norms expressed in other ways.

The finer grouping used here is local to this question. The shared object and standard categories organize the whole corpus, whereas these clusters summarize variation inside an observation, a rule, or a requested remedy. K-means partitions the embedded descriptions around representative centers, and reading examples then guides the merging and naming of groups. The labels thus reflect the

resolution chosen for this material, and a minor variant that receives no cluster of its own can still occur in individual evaluations.

Figure 18 arranges novelty criticism in three text columns: observations, reasoning rules, and requested remedies. Group labels and percentages describe the represented content, and the links between columns trace connections between fields, with widths summarizing estimated within-object frequencies. Color distinguishes kinds of content or requested work and is not a severity scale. Because a unit may carry several rules and the fields were sampled separately, the links do not add up as a conserved flow through all three columns. Samples are capped at 40,000 units per object where necessary. The evidence rule appears in all twelve objects, novelty's rule in six other objects, and clarity's in three, so what a criticism is about and the principle used to evaluate it are separate things.

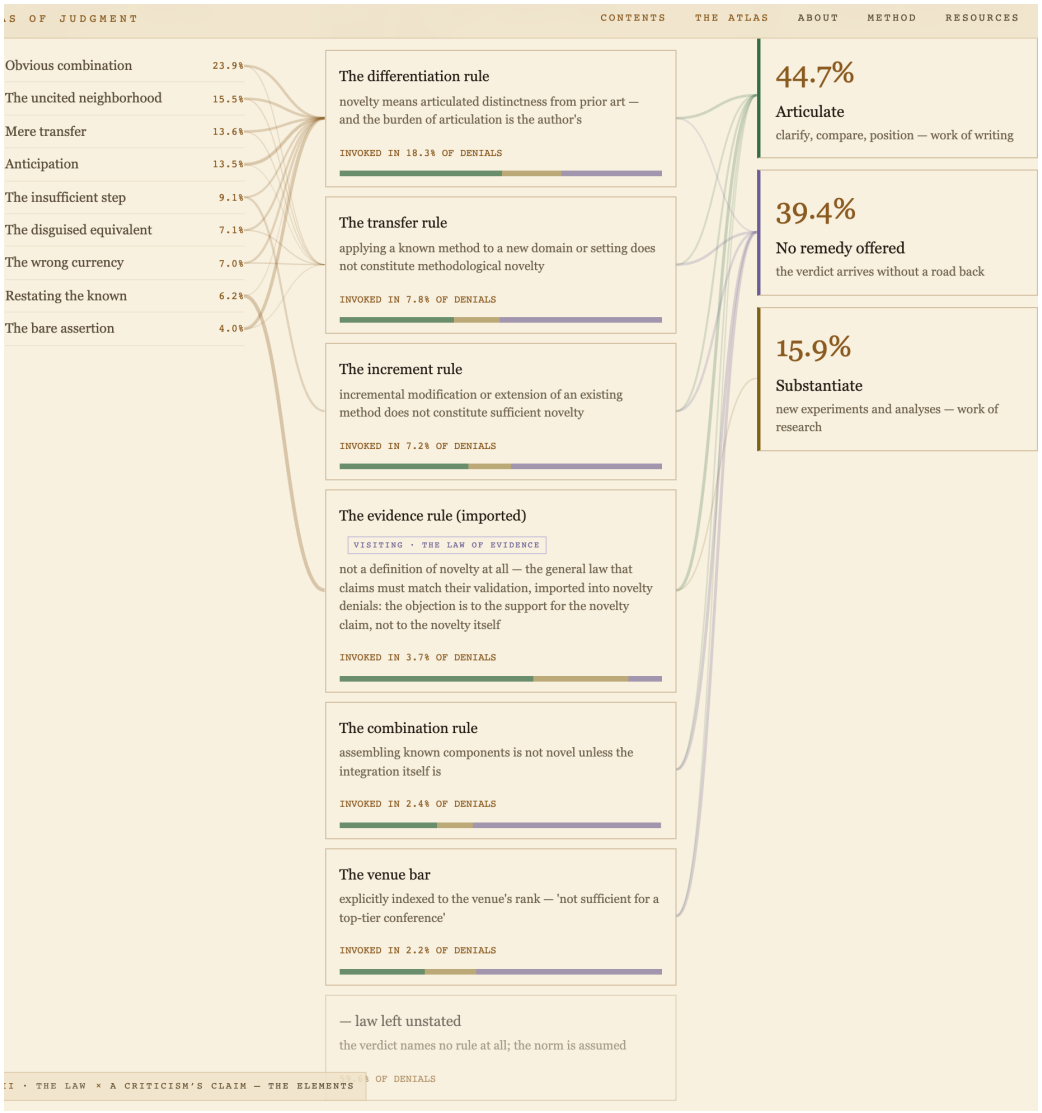


Figure 18: Ground–rule–remedy decomposition for novelty. Flow widths are estimated shares within the object; multiple rules and separate sampling mean the ribbons are not a conserved flow. Other objects use the same instrument.

Figure 19 is a set of six clause cards rather than a plot with axes. Each card gives a novelty-rule family, representative restated wording, its interpretation, and a source-review reference. The percentage on a card is the family's estimated share of negative novelty evaluations, not the prevalence of the displayed sentence, and card size and placement carry no frequency information. The novelty rulebook is built from 12,789 normative sentences. Its operative clauses require articulating differen-

tiation, exceeding an increment, providing more than an assembly, going beyond transfer, meeting a venue-indexed standard, and supplying evidence for the claimed contribution. Because the clauses unpack a novelty denial into distinct requirements, two judgments that a work is not novel need not allege the same failure. The representative wording is the model's tidied version of the review even when it is linked to an original sentence. A three-reader audit further separates false priority from ordinary similarity, and an unlocatable contribution from an explicitly incremental one, but these are refinements of the reading rather than new frequency estimates for the whole corpus.

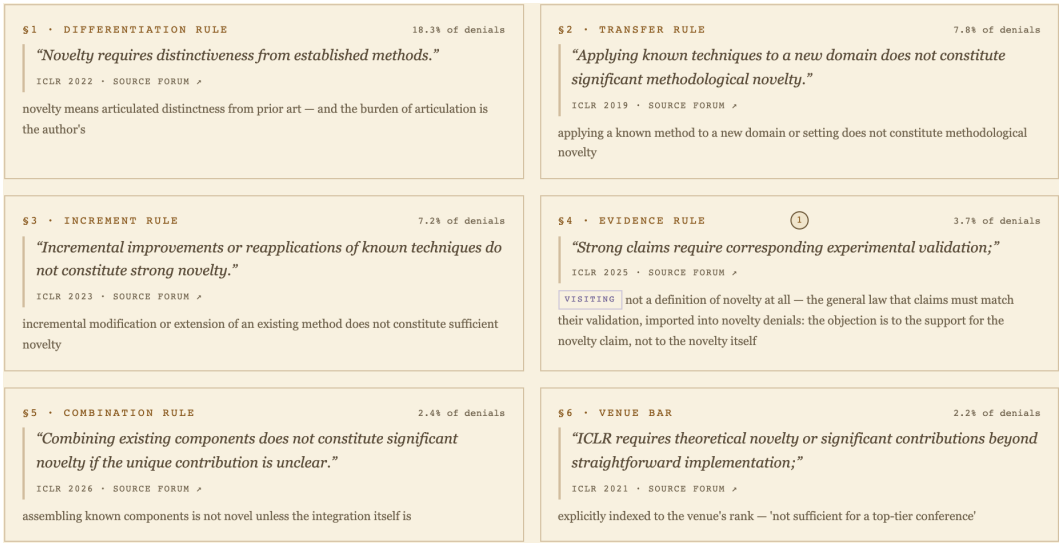


Figure 19: The novelty clause families and representative formulations. Exemplars are interpretations linked to source reviews; they should not be treated as exact raw quotations without checking the originals.

The remedy comparison asks whether a criticism points toward work an author could undertake, and of what kind. When the improvement field is empty, the unit is recorded as having no suggested fix rather than filled in with the analyst's own proposal. A terse unit may omit a remedy simply through brevity, so a length-stratified version groups units of similar length before comparing novelty with the other objects. That controls one descriptive difference between units; it does not explain why reviewers do or do not offer a response.

Figure 20 compares the kinds of remedy recorded for criticism of each object, with criticized objects in rows. Each horizontal bar spans all sampled negative units on that object, and its colored segments give the shares requesting articulation, substantive work, redesign, disclosure, or no remedy; the no-remedy percentage is also printed at the right. A row is therefore normalized to criticism of that object, including units with an empty improvement field. Clarity criticism mostly requests writing (90%), baselines criticism experiments (80%), and reproducibility criticism disclosure (76%). Novelty leaves approximately 39% of its denials without a fix, whereas the other objects supply one in 82–92% of cases. The gap survives stratification by unit length: within length quartiles, novelty offers a fix in 60–61% of units against 76–94% for other objects. Novelty units are in fact slightly longer than the pooled remainder (221.5 versus 203.0 characters), so brevity alone does not account for the difference. A novelty denial thus arrives less often with a recorded path for the author, even among units of comparable length. A zero redesign share at this cluster resolution does not prove that no review ever requests a new method; theory retains a small redesign component in the broader object-specific instrument.

The coupling analysis compares the observation, rule, and remedy categories within the same negative novelty units. Under strong coupling, knowing the observation would greatly narrow the likely rule or remedy. Cramer's V summarizes the overall strength of association between two categorical groupings, and lift compares one observed pairing with its expected frequency under independence. Rejecting independence and measuring the size of an association are different matters, since a large dataset can distinguish a small association from exact independence.

Figure 21 shows how observations, warrants, and requested remedies are associated within negative novelty units, one column for each. Node heights and ribbon widths show relative frequency. Rib-

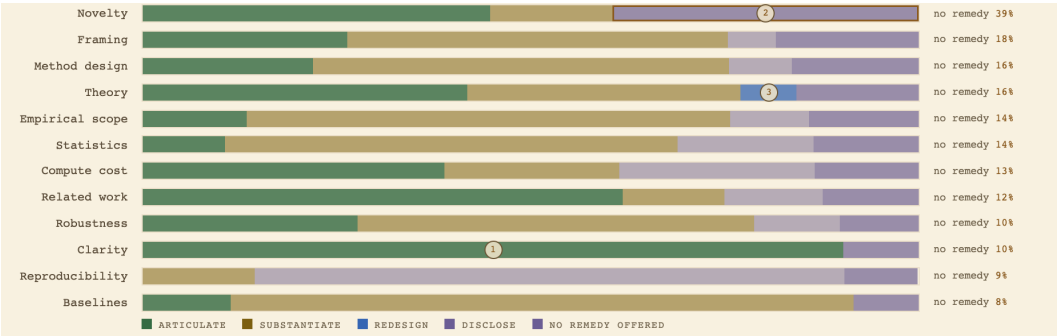


Figure 20: Remedy families for negative units by object in the forum-level track. Samples and clustering limit the precision of individual shares, particularly rare redesign requests. The 2026-only no-fix calculation is reported separately below.

bon color shows lift over the product of the relevant marginal frequencies, warmer for excess and cooler for deficit, and the printed multipliers are dimensionless lifts. The empty-request node is an observed category, and the vertical ordering within a column is categorical rather than a numerical scale. In 30,447 negative novelty units, the observation, reasoning, and fix fields are embedded separately, clustered into 10, 10, and 8 raw groups (seed 46), and then interpreted at a coarser grain. Cramer's V between successive stages is approximately 0.12–0.14. The largest named observation–rule lift is about 1.9, while the link from a combination observation to the combination rule is only 1.22. Combination reasoning requests nothing in 49% of its uses, versus 25% for rigor-doubt reasoning. The sample is large enough to reject exact independence, but the association itself remains small. Within the extracted account, knowing the observation category therefore leaves considerable uncertainty about which rule was applied and which remedy requested. Generic memo phrasing can attenuate such associations, so the result describes what survives this extraction rather than all the reasoning a reviewer might have performed.

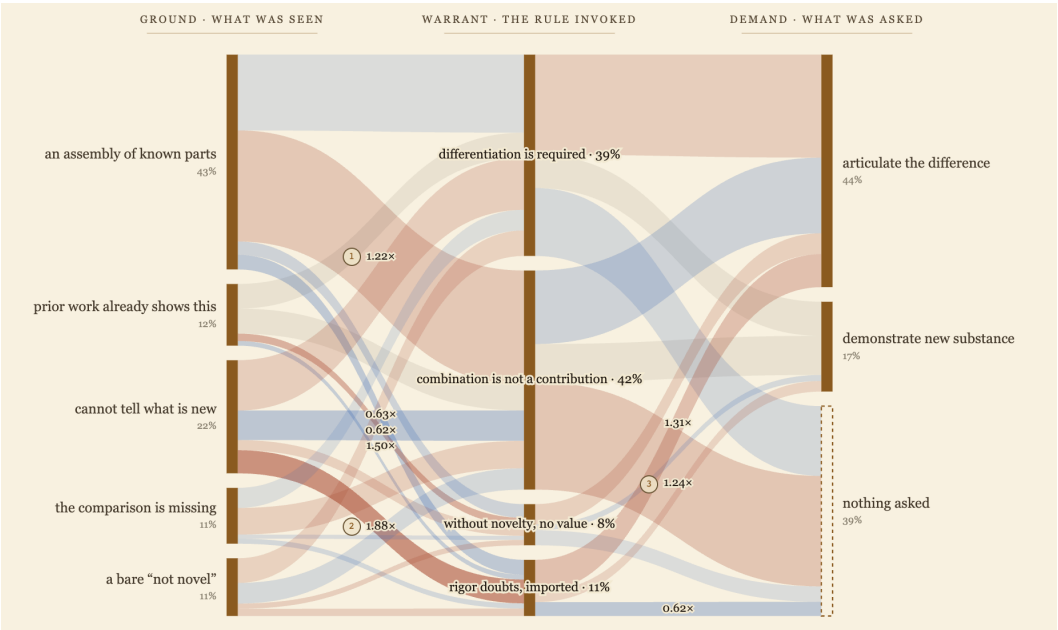


Figure 21: Coupling within negative novelty units. Ribbon width is frequency and color is lift against the product of marginal frequencies. "No request" is an explicit outcome, so coupling includes whether any remedy is requested.

5.2 Combination objections and the absence of a named precedent

The combination clause offers a useful case for asking what an apparent rejection rule actually permits. Calling an idea a combination might reject all integration, or it might only require some additional contribution. We isolate combination-rule sentences, cluster their formulations, and read exemplars to separate stated exceptions from unqualified boundaries. A second grouping, of the corresponding observations, records whether the components or precedents are named. Rule sentences and observations answer different questions and have different denominators.

Where the preceding decomposition pooled many kinds of novelty criticism, this analysis follows one rule into the specific conditions it expresses. An exception is a stated way in which combining existing elements could still count as a contribution; a boundary statement offers no such route in the passage analyzed. The distinction concerns how the rule is formulated, not whether the criticized paper actually satisfies it. What the observation names is a separate question: the components being assembled, a particular predecessor, or neither. Keeping the rule apart from its referent prevents an apparently general standard from being mistaken for a documented comparison with a specific prior work.

Figure 22 branches from a combination objection into six stated exception families and two boundary formulations. Branch weights and printed percentages are shares of the 4,316 combination-rule statements; brown tracks specify a qualifying route and gray tracks give no exception. The horizontal layout follows the logic of the statement, not time or any measured amount of novelty. Among the 4,316 combination-rule sentences, drawn from 4,262 units, six exception families account for 76%: added value, new insight, differentiation, a new mechanism, an isolated contribution, and innovation beyond integration. The remaining 24% are boundary statements with no stated route to qualifying as novel. On the dominant reading, then, combination is conditionally insufficient: integration can qualify if it supplies an additional contribution. Explicit conditional language appears in 43.6% of sentences, so the semantic families are broader than a keyword count of "unless" or "without." Local k-means uses eight clusters and seed 46, and near-duplicate generic referent clusters are merged after exemplar review.

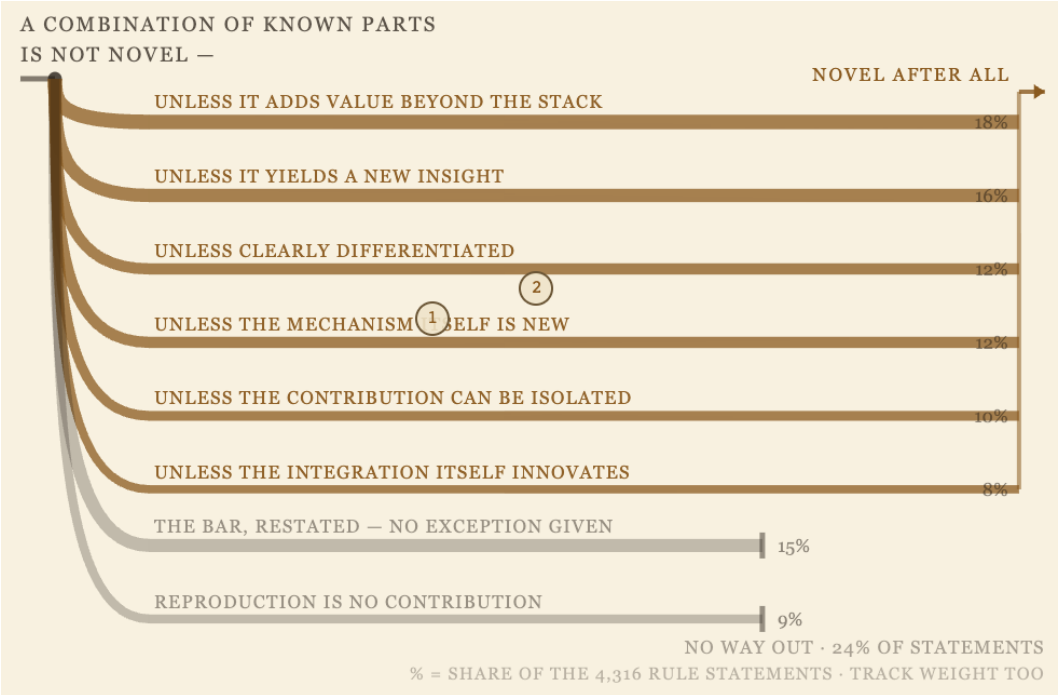


Figure 22: Exception and boundary readings of the combination rule. The figure opens one rule from the preceding decomposition; percentages refer to normative statements, not all reviews.

Figure 23 displays five observation categories as cards, each with schematic component symbols, a description, representative wording, and the category's percentage. The percentages divide the

combination charges by what they identify, such as named parts or a named precedent, so the denominator is observations rather than the rule statements of the preceding figure. Symbols and card areas are illustrative. Of the corresponding observations, 46% do not name the assembled parts, 18% name them, 16% identify one specific precedent, 12% concede a new problem but dispute the tools, and 8% identify no new insight. The represented observation of a combination objection thus often leaves the concrete comparison unspecified. A 30-case audit of generic observations against the source reviews largely agrees on specificity where the charging sentence can be located, with one ambiguous case. The audit qualifies the corpus-level proportion without fully validating it.

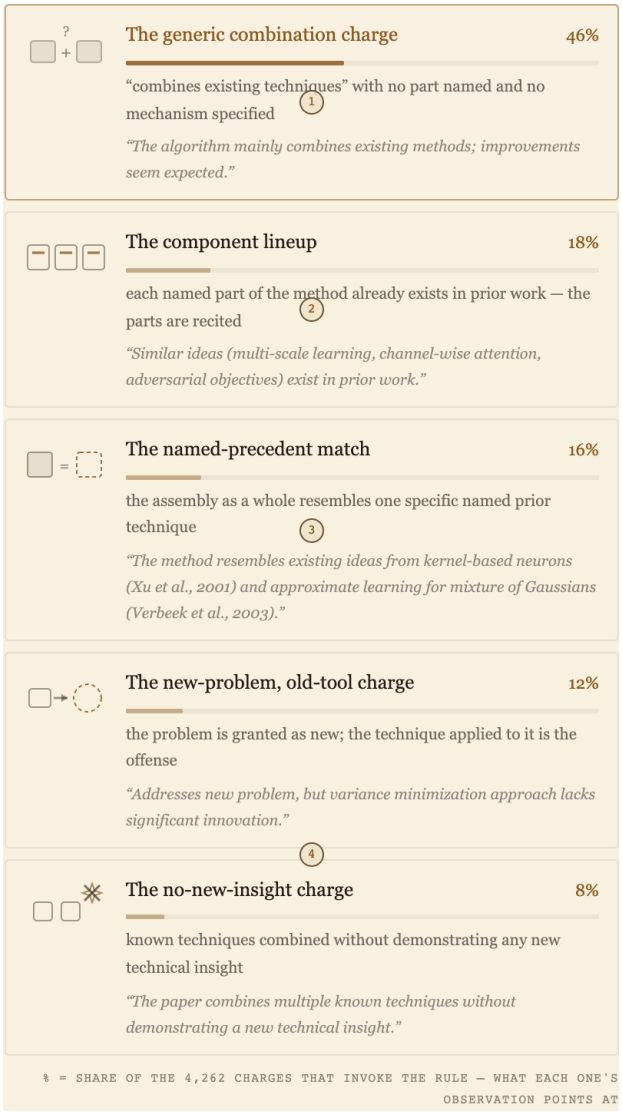


Figure 23: Referents of combination objections. The denominator is charges represented by the units' observations, whereas the preceding figure counts rule statements.

A broader question follows for novelty criticism in general: how often is the comparison anchored to an identifiable prior work? We search both locatable source-text windows and the distilled units, since the memo could have dropped a citation that the review contained. Named-reference shares are computed separately for the two representations. Counting recurring surname–year keys across forums then asks whether the identifiable comparisons converge on a common canon.

The source window matters because the structured unit is a compressed interpretation. Searching only the unit would confuse an unnamed criticism with a citation lost in compression. Locating nearby source sentences restores part of that context, and the separate count on the unit shows how

much the choice of representation moves the estimate. Neither search amounts to a full bibliographic reconstruction of the reviewer's comparison.

Figure 24 shows how often novelty criticism names a specific prior work in the surrounding raw review text. The left side is a percentage grid in which colored tiles partition locatable raw-text novelty windows into a named specific work, generic prior work, or no extractable referent, one tile for approximately one percentage point. The small line plot on the right has year horizontally and the percentage naming a specific work vertically. Tile color encodes referent type, and the line gives that type's yearly frequency rather than a count of citations. The citation analysis begins with 30,824 negative novelty units. For 24,233 of them (79%) a raw-text window can be located: every novelty-containing sentence plus two neighboring sentences on each side, in the source review with the best token overlap. A citation regex extracts author–year patterns and arXiv identifiers. It finds a named work in 12% of windows, a generic reference to prior work in 25%, and no referent in 63%. The named share falls from approximately 20% in 2018 to 11% in 2026. Applied to the distilled unit instead, the regex gives 5/29/66% and a decline from 8.4% to 3.5%. The level thus depends on the window searched, but the large no-referent share and the declining naming rate appear under both readings.

The most frequent surname–year key occurs in only 21 of 23,251 affected forums, even though works sharing a key were deliberately over-merged. The instrument therefore finds no dominant named canon. It misses prose references such as a well-known model's paper, and the generous raw window can absorb citations that belong to an adjacent criticism. An absent extractable citation does not show that a reviewer lacks relevant expertise or that the novelty objection is wrong.

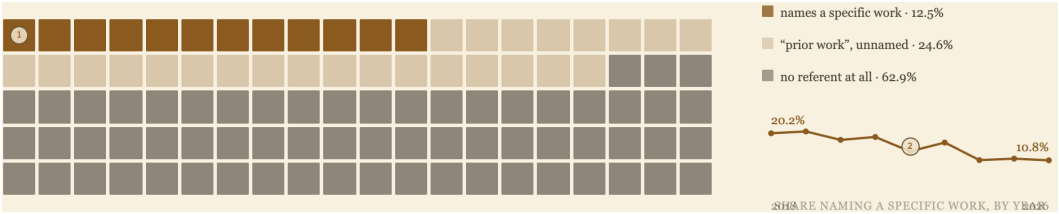


Figure 24: Citation referents in locatable raw-text novelty windows, with historical naming rates. Extraction misses nonstandard prose references; the raw-window and distilled-unit comparisons assess this instrument sensitivity.

5.3 Portable criticism and the taxonomy of suggested work

A criticism can omit a named precedent and still be specific to the paper. To keep the absence of a citation distinct from the recurrence of wording, we find each negative observation's nearest neighbor on a different paper and summarize the resulting similarities. Splitting threshold exceedances by object and rating shows where the extracted language is most portable. Setting those object-level shares beside the within-paper rating coefficients then asks whether the recurring formulations are also the objections most closely associated with low scores.

"Portable" here means that an extracted observation resembles one made about another paper. Requiring the neighbor to come from a different paper ensures that similarity is not just repeated discussion of the same submission. Cosine similarity compares the directions of the text embeddings, the vector representation introduced in Section 3, and so measures resemblance within that representation rather than whether two papers really share a defect. The threshold-based "formula" label is a descriptive convention for picking out highly similar formulations to compare across objects. It does not establish a literally shared sentence or a common author.

Figure 25 shows how closely extracted criticisms resemble criticisms of other papers. The horizontal axis is cosine similarity to the nearest extracted observation on a different paper. Curve height is the number of negative observations in each similarity bin of width 0.01, not a probability that copying occurred. Vertical guides mark the chosen similarity thresholds and the median, and the shaded upper tails supply the exceedance percentages printed beside the guides. For each of the 293,671 negative 2026 units, similarity of formulation is measured against the nearest embedded observation from another paper. The median nearest-neighbor cosine is 0.81; 19.9% of units exceed 0.85 and 3.8% exceed the conventional formula threshold of 0.90. The formula share rises from 3.6% at

ratings 0–2 to 6.8% at rating 10. Because the comparison is between the model's restatements, whose uniform voice can inflate absolute levels, it does not identify copying by reviewers.

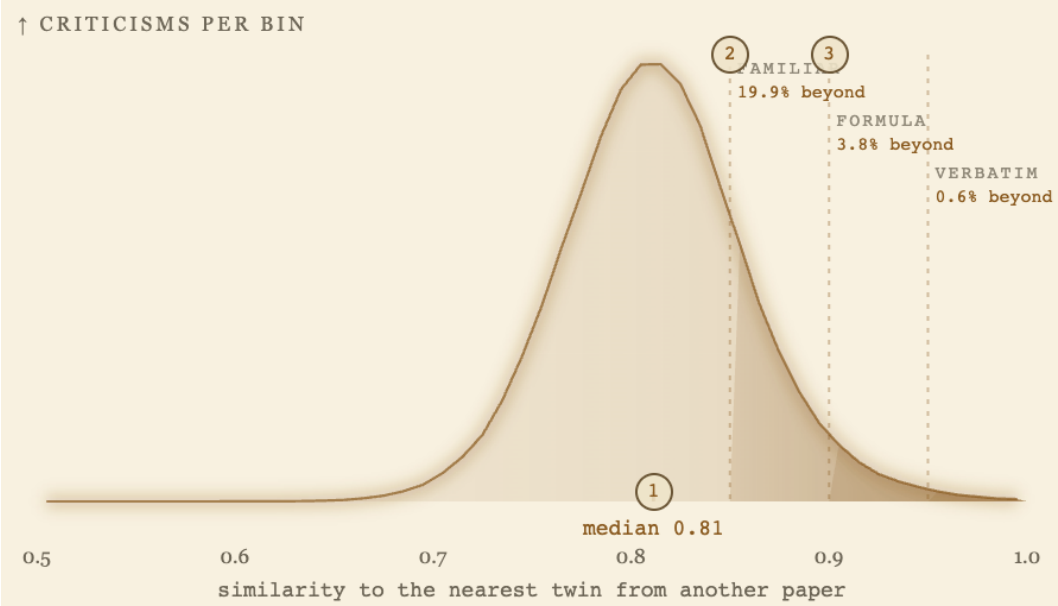


Figure 25: Distribution of nearest criticism similarity across different papers. Thresholds mark conventions for descriptive comparison, not validated categories of reviewer behavior.

Figure 26 sets the recurrence of criticism wording beside its rating association, object by object, as two ranked lists joined by curves. The left list gives the percentage of negative observations whose cross-paper nearest-neighbor similarity exceeds 0.90, and the right list gives the adjusted rating coefficient in points. Each curve connects the same object across the lists. Vertical position ranks the values separately on each side, so a crossing shows a change in ordering, not a time trajectory or a causal link. Novelty and framing have formula shares of 1.3% and 1.2%, whereas reproducibility and clarity have 8.7% and 8.0%. The broad inversion holds at thresholds 0.85, 0.90, and 0.95, although the fine rankings change. An unnamed novelty objection can therefore still be specific in its wording. Portable standards may also be justified responses to recurring defects, so formulaicity alone does not measure review quality.

Figure 27 makes the comparison concrete with paired text columns, each pair holding observations from different papers and grouped by similarity level. Highlighting marks the wording that changes between the two restatements, and the remaining text shows the shared formulation; source identifiers allow inspection of the originals. Rows are selected examples rather than frequency bins, and their text is the instrument's restatement.

The improvement field gives a complementary view of what evaluation asks of an author. Rather than counting only whether a fix exists, we cluster the suggested-fix texts, name the resulting types from their exemplars, and sum their shares within broader research, writing, and disclosure families. The aim is to separate requests for new evidence from requests to explain or disclose existing work. A sweep over clustering resolutions checks whether the family-level conclusion depends on the chosen number of clusters.

This taxonomy differs from the earlier object-specific remedy comparison in both population and purpose. It covers improvements attached to units of any valence in the review-level track, whereas the earlier comparison concerned negative units in the forum track. Research work means producing or analyzing further evidence, writing means changing how the work is explained or presented, and disclosure means making relevant information available. The families summarize the work requested in the extracted text; they do not show that the work was done or that doing it would resolve the objection.

Figure 28 lists fix types in rows, grouped and colored as research, writing, disclosure, and redesign. Horizontal bar length and the printed percentage give each type's share of the sampled improvement

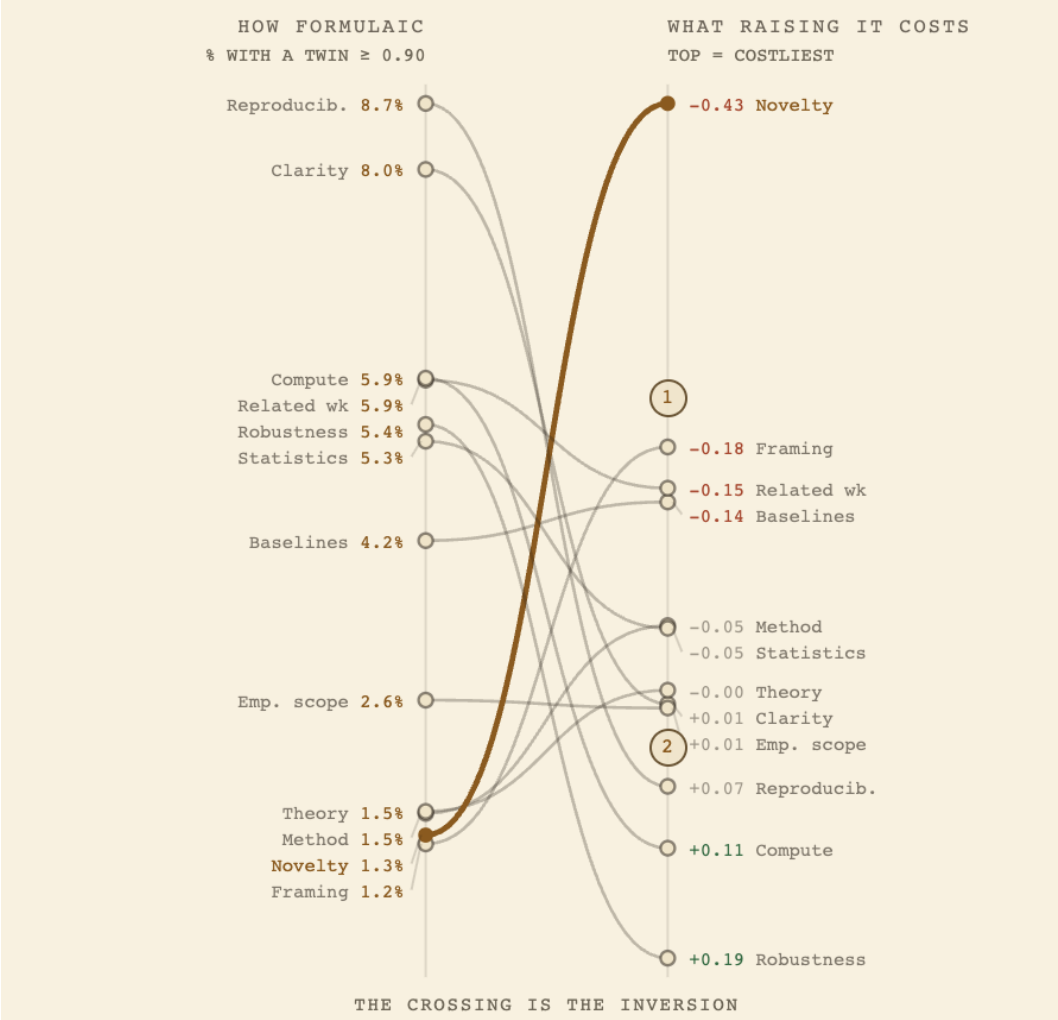


Figure 26: Formula share by object and its within-paper rating association. The rating coefficients are the same estimates analyzed in Section 6; this comparison is not an independent estimation.

texts, and the colored strip at the top aggregates the broader families. The denominator is fixes attached to units of any valence, not all extracted units or negative units alone. Of the 410,586 review-level units, 307,388 contain a suggested improvement. A sample of 80,000 fixes is embedded, clustered into 22 groups (seed 7), and merged into sixteen named types. Extending the evidence is the largest single type at 22%. At the family level, which is more stable across resolutions, 69% of fixes request research work, 26% writing, and 5% disclosure. The dominant requested response is thus further research, with writing and disclosure as smaller complementary demands. The family assignments are interpretive: error bars count as research because they must be computed, and surface fixes count as writing. These 22 clusters resolve no standalone redesign type. That is a statement about this clustering, not about the literal absence of redesign requests from every source review, and it sits beside the small theory redesign share found under the object-specific clustering above.

To set negative judgment beside stated routes to improvement, we return to the full 2026 track. For each object, each valence count is divided by all units on that object, and negative units lacking an improvement are divided by all negative units on it. The two ratios separate how often an object is criticized from how often its criticism arrives without a suggested response, and neither can be read off the distribution of fixes alone.

Figure 29 compares evaluative valence with the absence of a suggested fix, with objects in rows in both panels. On the left, colored segments give the within-object shares of negative, hedged, and



Figure 27: Illustrative near-neighbor pairs at three similarity levels. Highlighted differences reveal what varies in the model-restated observations; links in the source Atlas allow comparison with raw reviews.

positive units, arranged around a common visual boundary. On the right, horizontal length is the percentage of that object's negative units with no suggested fix. The panels therefore have different denominators, all units on the object and only its negative units. Related work and novelty have the largest negative shares, 88% and 82%. In this 2026 review-level track, novelty negatives lack a suggested fix in 37% of cases, versus 8–18% for other objects. The 37% does not revise the nine-year figure of 39%, since the two come from different tracks and windows. Shares of reviewer-explicit support lie within 17–22% across objects here. Novelty is thus distinctive in two respects: it is frequently judged negatively, and its criticism is comparatively unlikely to carry a proposed repair. A suggested repair is a property of the written interpretation, and its absence does not establish that the problem is objectively unfixable.

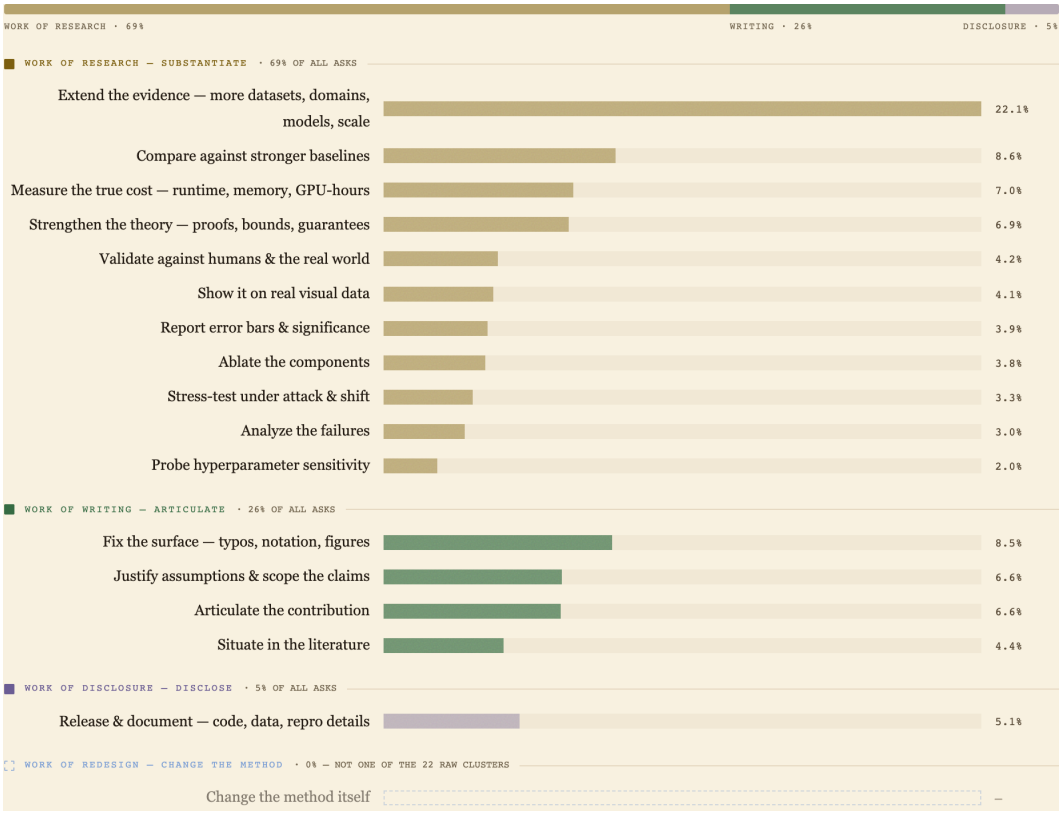


Figure 28: Sixteen suggested-fix types and their research, writing, and disclosure families in a 2026 sample. Fine type shares depend on clustering resolution; family totals are the more stable summary.

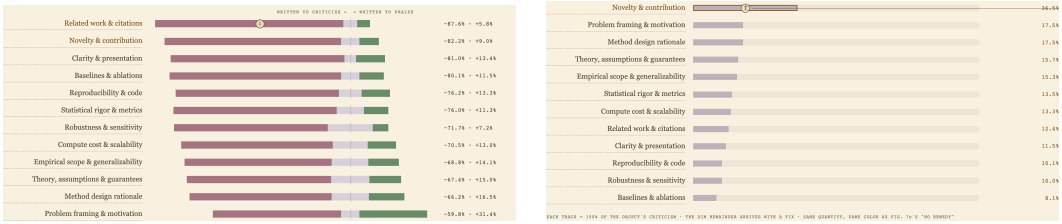


Figure 29: Left: Valence by object in the 2026 review-level track. **Right:** The proportion of negative units without a suggested fix in that same track.

5.4 Rules co-occur differently within and across reviewers

The decomposition also allows criticisms to be compared within a document. We ask which stated rules occur together more or less often than their marginal frequencies predict. Such co-occurrence can arise because the paper attracts several related complaints or because of how one reviewer assembles a review. We therefore measure rule-pair lift first within reviews and then across different reviewers of the same paper, holding the paper fixed while changing the writer. The comparison narrows the interpretation of co-occurrence without pinning down a single behavioral cause.

A rule pair is present when both named principles are identified in the relevant reviews. Attraction means excess co-occurrence relative to the chosen reference and repulsion means a deficit; neither term refers to personal affinity or disagreement between reviewers. The same-paper comparison is informative because distinct reviewers face the same submission. A relation that persists across them need not depend on one writer's way of composing a document, whereas a relation largely confined to a single review may reflect its composition or the extraction process. Several causes remain possible under either reading, and the comparison does not separate them experimentally.

Figure 30 shows which stated rules co-occur more or less often than their separate frequencies predict. The same 61 stated-rule categories appear on both axes, grouped by the object they concern. Each cross-object cell is a co-occurrence lift, the observed joint frequency divided by the product of the marginal frequencies. Brown denotes excess and violet a deficit, with stronger color for a larger departure from unity. Pale gray cells have too few joint observations, within-object blocks are omitted, and outlined cells mark the extreme pairs examined next rather than confidence intervals. Of 191,946 initial reviews in the forum track, 120,540 contain an identified stated rule. Cross-object pairs among the 61 categories are compared when they have at least 60 joint filings, and within-object pairs are excluded. Of the eligible pairs, 82% have a lift above one, with a median of 1.19. Strong attractions connect formal precision, theory rigor, statistical reporting, and presentation, and the largest displayed pair reaches approximately 3.6. Novelty-related clauses show some of the strongest negative associations, down to approximately 0.66.

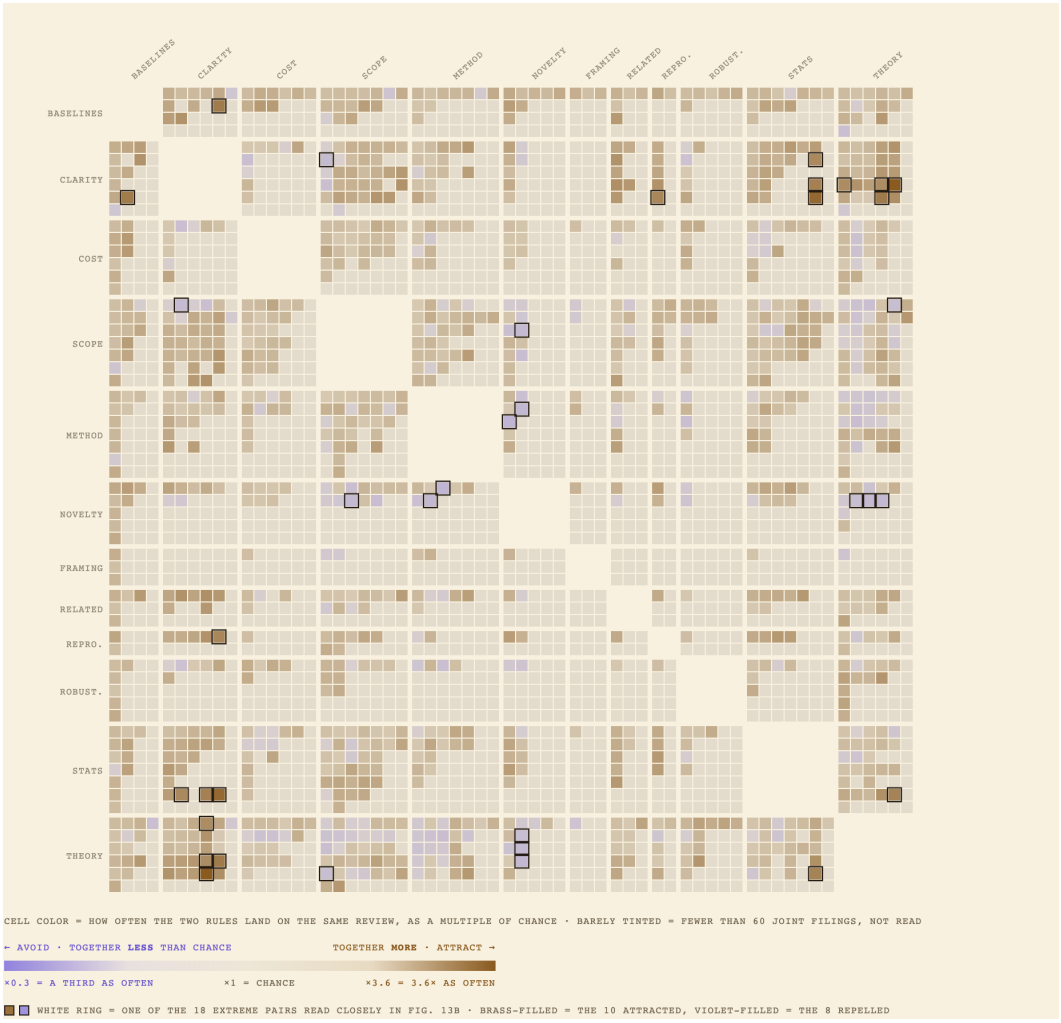


Figure 30: Cross-object stated-rule co-occurrence. Only sufficiently frequent pairs are interpreted; a positive background can arise because longer or more thorough reviews contain more of many rules.

Figure 31 compares selected rule-pair associations within a single review with the same pairs across different reviewers of the same paper. Rows name rule pairs, and the horizontal axes are logarithmic co-occurrence lift with unity marking independence. The left panel shows attracted pairs and the right repelled pairs. Filled dots are within-review estimates, open circles are the cross-reviewer estimates, and the connecting segment shows the change between the two. Whiskers give paper-bootstrap uncertainty for the within-review estimate; the connector itself is not an interval. Under paper-clustered resampling, the displayed attraction intervals stay above 1.25 and the repulsion intervals below 0.91. When the two rules are instead taken from different reviewers of the same paper,

the median lift among displayed attracted pairs falls from 1.87 to 1.44, and among repelled pairs it moves from 0.75 to 0.97. Re-basing the cross-reviewer comparison by year changes the medians by only 0.01. The pattern is consistent with shared paper characteristics contributing to the attractions, while the suppression of additional novelty clauses is mainly a within-review phenomenon. It does not single out a human habit, because extraction and normative-sentence sampling can also induce within-review dependence. Because units are sampled separately within each object, we report ratios rather than absolute co-filing prevalence, and even the ratios are preserved only approximately under the sampling assumptions.

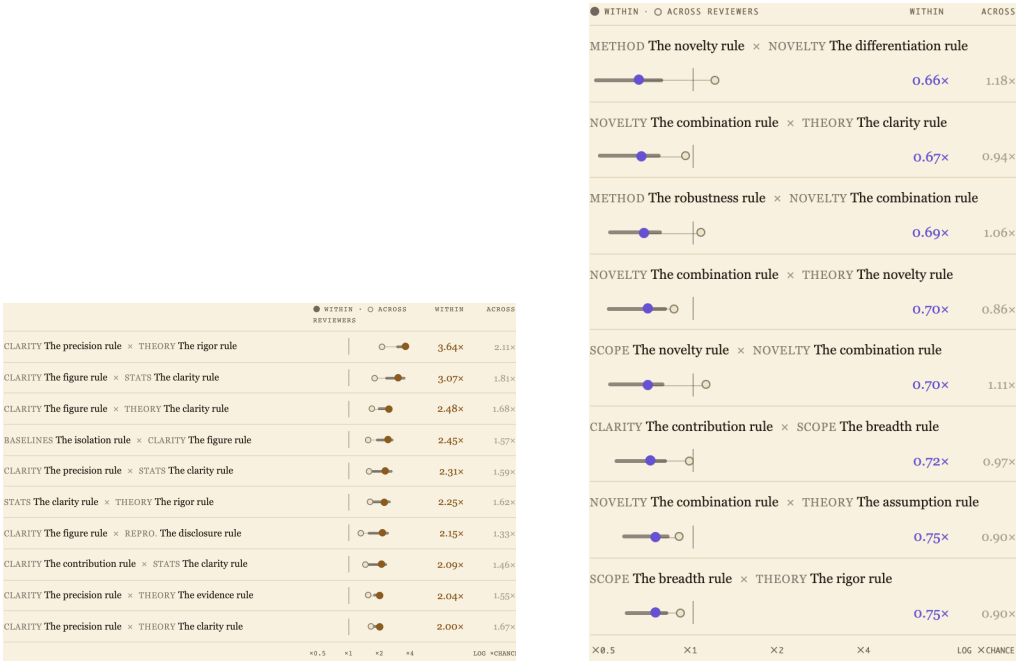


Figure 31: Left: The most attracted rule pairs, with within-review and cross-reviewer comparisons. **Right:** The most repelled pairs, whose suppression largely disappears across reviewers of the same paper.

6 Associations with ratings and decisions

6.1 Within-paper rating differences by object

Having described which criticisms occur, we ask whether different criticisms have different associations with the score attached to the review. A plain tally of negative units treats all objects as interchangeable, and object-specific coefficients test that simplification. We also set the coefficients beside fix prevalence to ask whether the criticisms associated with low scores are the ones that come with an apparent route to repair.

The observational unit is a review with its attached rating. Its object indicators record whether each kind of criticism is present, and its negative-unit count records how much criticism the review contains. Including both separates what a reviewer criticizes from how much the reviewer criticizes. Subtracting the paper mean gives the rating a local reference, so a negative value means that this reviewer rated the paper below its panel average, not that the paper received a low absolute score. The comparison is thus among co-reviewers of the same submission.

Comparing across papers would confound the criticism with the paper that attracts it. The 2026 rating analysis therefore centers the rating and every predictor within paper by subtracting the paper-specific mean, then regresses the centered rating on the centered object-criticism indicators and two unit-count controls, total negative units and total units. Bootstrap resampling of whole papers supplies 95% intervals. The resulting coefficient, called a rating deficit here, is a conditional association; no objection is randomly assigned.

The bootstrap repeatedly draws complete papers with all their reviews and refits the model, so that dependence among a paper's reviews is preserved when uncertainty is estimated. Suggested-fix prevalence is a separate summary of whether units record an improvement proposal. Placing it beside the rating deficit asks whether criticism associated with lower ratings also tends to name something the author could do. The analysis does not assess whether the proposal was feasible, carried out, or successful.

Figure 32 shows the within-paper rating association of each criticism together with its suggested-fix prevalence. On the left, criticized objects occupy rows and the horizontal axis is the fitted within-paper rating coefficient in rating points. Filled dots are adjusted estimates with paper-bootstrap 95% intervals, open circles are unadjusted comparisons joined to them by dotted connectors, and the vertical zero line marks no adjusted difference. Red and green mark intervals entirely below and above zero, and neutral dots mark intervals that cross it. On the right, each bubble is an object, placed horizontally by the share of its negative units with a suggested fix and vertically by the same adjusted coefficient. Larger bubbles indicate more criticisms, with a square-root radius rule and a minimum marker size. Novelty has a coefficient of -0.43 rating points, approximately 2.4 times the next-largest negative coefficient. Framing is -0.18, and related work and baselines are about -0.15. Robustness and compute have positive coefficients after adjustment. The category of a criticism thus distinguishes rating positions within a panel beyond what a count of negative evaluations captures. The signs may arise because a favorable reviewer finds relatively minor faults, or because a skeptical reviewer challenges what the paper contributes; they do not measure the effect of asking an author to fix a particular problem.

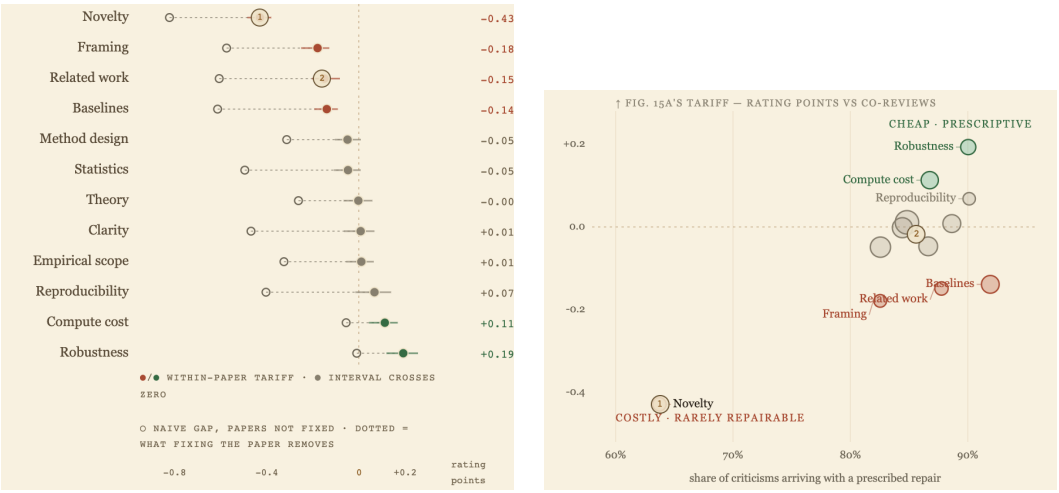


Figure 32: Left: Within-paper object coefficients, with unadjusted comparisons and paper-bootstrap intervals. **Right:** The same coefficients plotted against suggested-fix prevalence. These are two presentations of overlapping measurements, not independent evidence of a causal repairability mechanism.

Reviewers who inspect the same object can still disagree. We therefore form same-paper reviewer pairs on each object, split them by whether their dominant reasoning standard matches, and compare the proportion taking opposite stances. Holding the object fixed makes this a check on differences associated with the invoked standard rather than with the choice of subject.

The counting unit is now the pair rather than the individual unit. A dominant standard summarizes the standard used in a reviewer's extracted evaluations of that object, and an opposite stance records disagreement in evaluation rather than a mere difference in wording. Reviewers who produce no unit on the object cannot enter the comparison, so the answer concerns disagreement among reviewers who share a represented subject of attention.

Figure 33 compares disagreement over an object when co-reviewers invoke the same standard and when they invoke different ones. A pooled row is followed by object-specific rows. From the central reference, the brown segment extends left by the percentage of same-standard reviewer pairs with opposite stances, and the red segment extends right by the additional percentage points for different-standard pairs; these are two quantities rather than one signed score. Right-hand annotations give that increase and the number of eligible pairs. The jagged line shape is decorative and

encodes neither a distribution nor uncertainty. Pairs with the same dominant standard take opposite stances in 20.5% of cases, pairs with different standards in 30.1%, a difference of approximately 9.7 points. The same-standard groups for reproducibility and related work are small, with fewer than 1,500 pairs each. Shared attention to an object therefore does not settle disagreement, and the represented standard adds a descriptive distinction. Conditioning on paper and object reduces some confounding, but standard and stance are extracted from the same units and may be two expressions of one evaluation. The contrast cannot show that changing a reviewer's standard would change their stance.

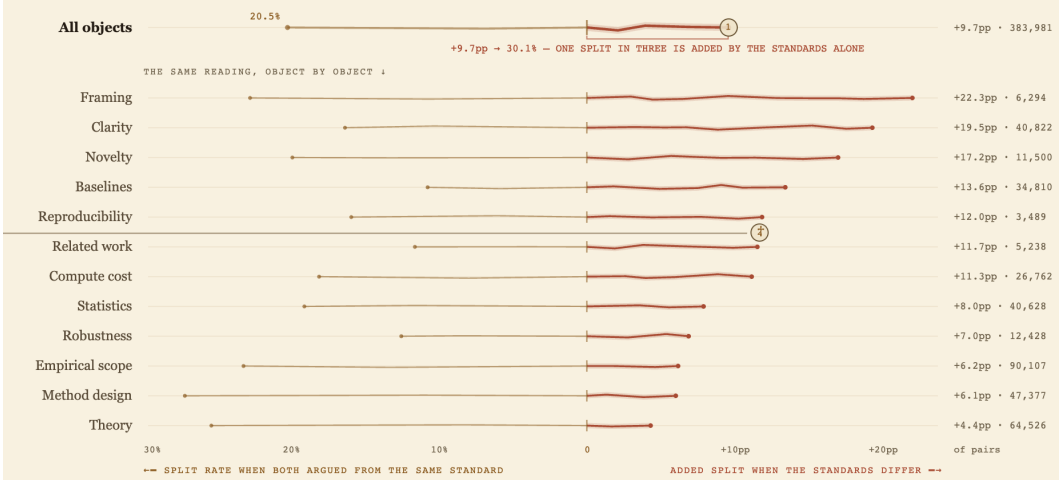


Figure 33: Opposite stances among reviewers inspecting the same object of the same paper, comparing shared and different dominant standards. Differences are descriptive and can include joint extraction effects.

6.2 The same stated rule has different rating associations across objects

An object-specific coefficient leaves the rule behind the objection unspecified, and a rule such as requiring evidence may carry different implications when aimed at a paper's motivation than at its clarity. We therefore compare the same stated rule across objects, and different rules within novelty, using each invoking review's score relative to its co-reviewers. Together the two comparisons ask how much of the association the rule's name alone summarizes.

A stated rule here is an explicitly formulated requirement identified in the normative-sentence analysis, not a standard inferred for a unit. An evidence rule demands support for a claim, while an increment or combination rule asks whether a contribution goes far enough beyond an extension or an assembly. Rules describe evaluation at a different level from the object labels. The cross-object comparison holds a rule family fixed while changing what it evaluates, and the within-novelty comparison holds the object fixed while changing the requirement. Expressing the rating difference in units of that year's standard deviation relates its size to the year's observed rating dispersion.

The analysis pools years, measuring each invoking reviewer's score relative to co-reviewers in units of that year's rating standard deviation. Within-year standardization expresses each deficit relative to that year's rating dispersion; it does not establish that the rating forms have the same meaning. Sampling is inherited from the normative-sentence clusters. Because the evidence rule appears in all twelve objects, its association can be compared across objects with the coarse rule category held fixed.

Figure 34 compares the rating associations of stated rules across objects and across novelty clauses. Objects occupy the rows of the left panel and novelty clauses the rows of the right. Both horizontal axes measure the invoking review's within-paper rating deficit in units of that year's rating standard deviation. Filled dots are estimates adjusted for criticism volume, open circles are raw estimates, and whiskers extend one standard error on either side of the adjusted estimate. Zero means no deficit, and the dashed line in the left panel is the adjusted average across objects. On the right, color separates novelty's own clauses from its imported evidence rule. After adjustment for criticism volume, the evidence rule is associated with a -0.30σ deficit in framing and approximately none in clarity; the

framing–clarity contrast is -0.29σ with a paper-bootstrap 95% interval of $[-0.38, -0.20]$. Within novelty, the increment and combination rules carry deficits of approximately -0.44 to -0.46σ, against -0.19σ for the imported evidence rule. The name of the rule alone therefore does not summarize its rating association. Whether the rule was a cause, a justification, or a consequence of the reviewer's judgment remains open.



Figure 34: Left: The evidence rule's standardized within-paper deficit across objects, before and after volume adjustment. Whiskers represent standard errors. **Right:** The same comparison within novelty, separating its own clauses from the imported evidence rule.

6.3 Which criticism accompanies rejection?

A reviewer's own rating and the institutional decision are different outcomes. To ask whether object-specific criticism also separates accepted from rejected submissions, we reduce each paper's reviews to twelve indicators of whether any reviewer raises a negative unit on each object, and compare acceptance proportions with and without each criticism. Unlike the preceding comparison of reviewers of the same paper, this is a paper-level comparison across submissions.

The acceptance gap is the acceptance proportion among papers with the criticism minus the proportion among papers without it. Each paper contributes once to each object comparison however many reviewers raise the objection, which avoids counting a heavily reviewed paper repeatedly but discards the intensity of its criticism and the agreement behind it. Because different papers supply the two groups, the gap can reflect differences among the submissions as well as differences in their reviews.

Figure 35 compares acceptance rates for papers with and without criticism of each object. Rows are criticized objects and the horizontal axis is acceptance percentage. Filled markers locate papers with at least one negative unit on the object and open markers those without, and the connecting segment shows the gap. Printed differences are in percentage points, with whiskers giving paper-bootstrap 95% uncertainty for the gap. The vertical venue line marks the overall acceptance rate in this join, and gray markers flag gap intervals that span zero. Of 19,474 reviewed 2026 submissions, 13,704 have a decision and are not withdrawn in this join. Novelty has the widest gap: -10.6 pp, with a 95% paper-bootstrap interval of $[-12.1, -9.0]$. Statistical rigor and baselines follow at -9.5 and -9.3 pp, while compute cost and robustness have intervals spanning zero. The subject of an objection thus matters to its association with the final decision, and criticism of different objects is not an interchangeable signal. Unlike the within-paper rating analysis, this comparison holds neither the paper nor the total amount of criticism fixed.

The next analysis predicts each paper's accept/reject decision in order to compare how much decision-related information different representations retain. A count model keeps how often each kind of evaluation appears but discards its wording. A text representation instead converts judgment sentences into numerical vectors and averages them per paper, keeping distinctions that the coarse labels need not express. A score-based reference uses the reviewers' own numerical summaries. Cross-validation fits each model on one portion of the data and evaluates it on held-out portions

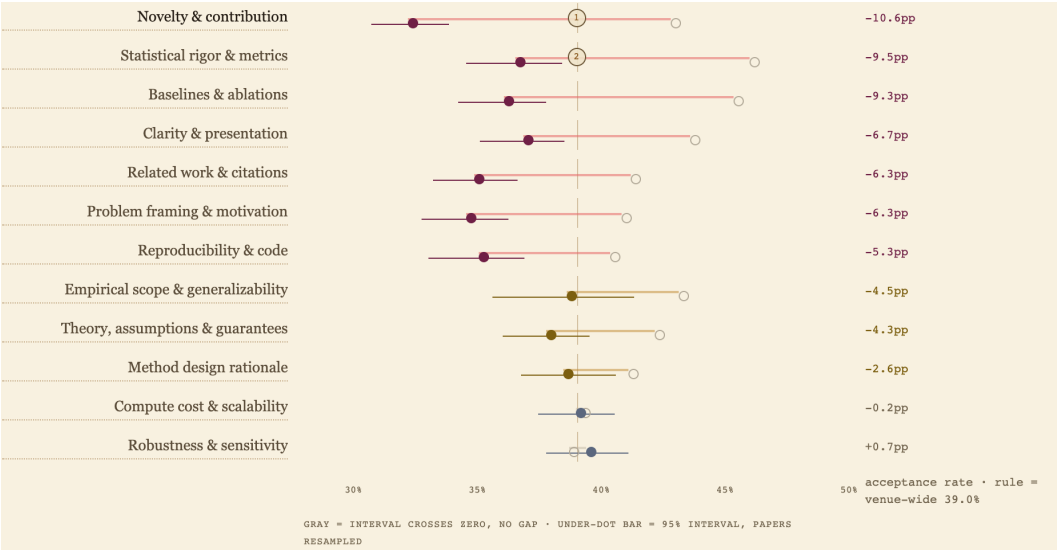


Figure 35: Acceptance with and without at least one negative unit on each object. The denominator is the 13,704 decided, non-withdrawn 2026 papers in this extraction join, not all collected submissions.

rather than on the observations used for fitting. AUC measures how well the predictions rank accepted papers above rejected ones, not the fraction of decisions reproduced at a chosen cutoff. The exercise assesses retained information; it does not propose a replacement decision rule.

Figure 36 compares how well representations of review content and panel scores predict acceptance. The upper part has feature sets in rows, review count, evaluation counts, judgment text, and panel scores, with cross-validated AUC as the horizontal quantity and the values printed beside the bars. AUC is a ranking measure on a 0–1 scale with chance at 0.5. The lower part plots year horizontally and AUC vertically for the count and text models. The hatched score bar is the outcome-proximate reference, whose historical scoring forms differ. The review-count baseline checks whether simply receiving more reviews explains performance. Across 39,484 decided, non-withdrawn, non-desk-rejected papers with eligible units labeled as initial, a logistic model reads twelve negative-object counts plus positive and hedged counts. Stratified five-fold cross-validation with seed 7 gives an AUC of 0.645, against 0.51 for review count alone and 0.652 when review count is added. Mean-pooled embeddings of the unit judgment sentences give 0.740, and reviewer scores give a reference of 0.896. The judgment text therefore retains decision-related distinctions that category counts discard, while recorded scores remain the strongest of these descriptive summaries. The score reference pools changed scoring forms and is not directly comparable to a stable cardinal scale.

In the count model, novelty again has the largest negative weight (-0.36 per unit), followed by framing (-0.22) and baselines (-0.17), while compute and robustness contribute little. Within-year AUC for the count model falls from approximately 0.75 in 2018 to 0.63 in 2025–2026, whereas the text model falls less, from approximately 0.73 to 0.71. No decisions are available for 2019. The temporal contrast concerns how readable the decision is from coarse counts, not any decline in scientific rigor, since review prose, rebuttals, institutional rules, and other unobserved factors may change together. Scores themselves also become less predictive in later years.

The initial label selects a position in the extracted forum account, not a verified historical version of the input. The whole-forum memo was produced with later context available, including the decision reasoning. The performance values therefore describe how readable the outcome is in that representation, not how accurately an outcome-blinded pipeline could predict a future decision.

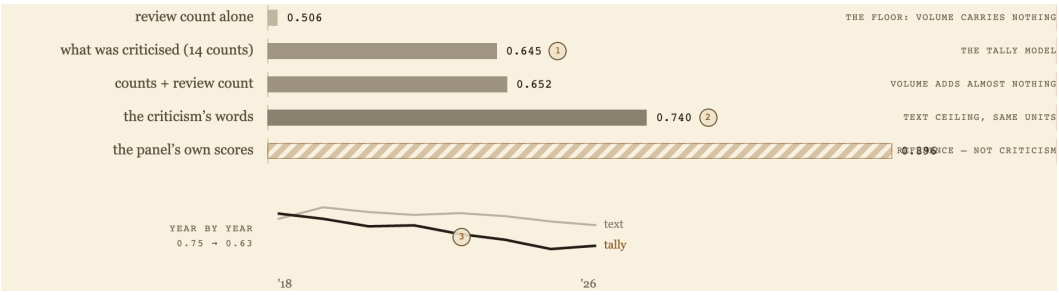


Figure 36: Cross-validated decision readability from counts, text, and scores, with the historical tally–text comparison. Units labeled initial are the input; their source is not a verified pre-rebuttal snapshot. The task does not identify the mechanism producing the decision.

6.4 The contents of the rating scale

What distinguishes the contents of a low-rated review from a high-rated one, beyond the expected difference in negativity? We join ratings to units and summarize, within each rating, the valence and standard shares, the fraction of reviews touching each object, reviewer confidence, and suggested-fix prevalence. We then compare low- and high-rating co-reviewers of the same paper to separate a change in the subjects inspected from a change in stance toward a shared subject.

The displays use different denominators. Valence and standard composition count units within a rating group, so a review with many extracted units weighs more in those shares. Object coverage asks instead whether a review contains any unit about the object and counts the review once. Praise-only comparisons restrict the denominator further, to positive units. Keeping these distinctions explicit prevents a change in the amount of text from being read as a change in the fraction of reviewers who inspect a subject.

Figure 37 compares the valence and reasoning-standard composition of evaluations across overall-rating groups. In the left panel, rows are the six available overall ratings from 10 down to 0, and each stacked bar partitions all extracted units at that rating, from left to right, into negative, uncertain, mixed, conditional, and positive valences. Thin bars beneath show review counts relative to the largest rating group, and the side text gives review count, units per review, and mean confidence. The right panel has standards in rows and, on the horizontal axis, each standard's percentage of units at rating 2 or 8; purple and green dots mark the two groups and are connected to show the difference. The connectors are not intervals. The 2026 scale runs from 0 to 10 in steps of two. Even rating 10 contains 35% negative units, and rating 0 contains 91%. The novelty and fair-comparison standards decline between ratings 2 and 8 (18% to 12%, and 10% to 6%), while merit recognition rises from 7% to 29%.

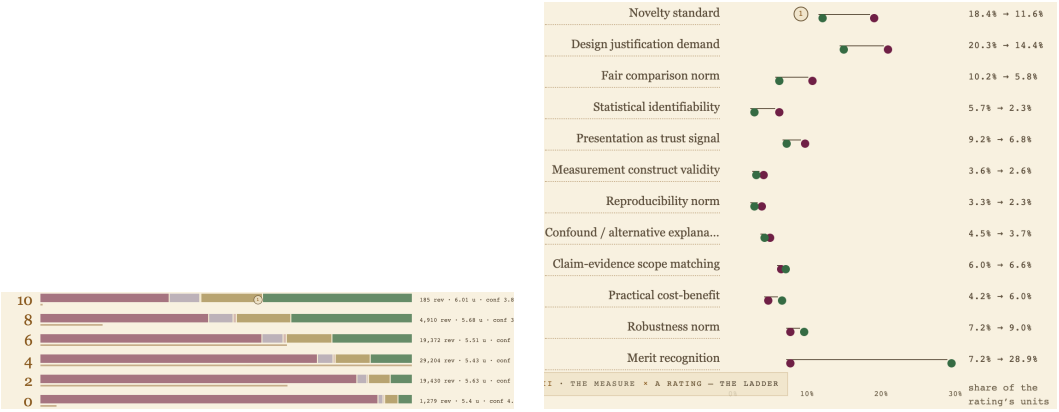


Figure 37: Left: Valence composition by rating. **Right:** Reasoning standards at ratings 2 and 8. Both summarize the same rating–unit join.

Figure 38 shows how coverage of the twelve objects differs across overall-rating groups, with rating groups in rows, objects in columns, and review counts in the row labels. A cell's brown intensity is the percentage of reviews at that rating with at least one extracted unit on that object, each review counted once, under a single color rule for all cells. Coverage is thus a review-level presence rate, unlike the unit shares in the preceding stacked bars. Novelty is inspected in 20% of rating-2 reviews but 9% of rating-8 reviews, and baselines fall from 42% to 29%, while theory, method design, and compute become more common. A high rating therefore does not mean an absence of criticism; the mix of subjects inspected and standards invoked differs across rating groups as well.

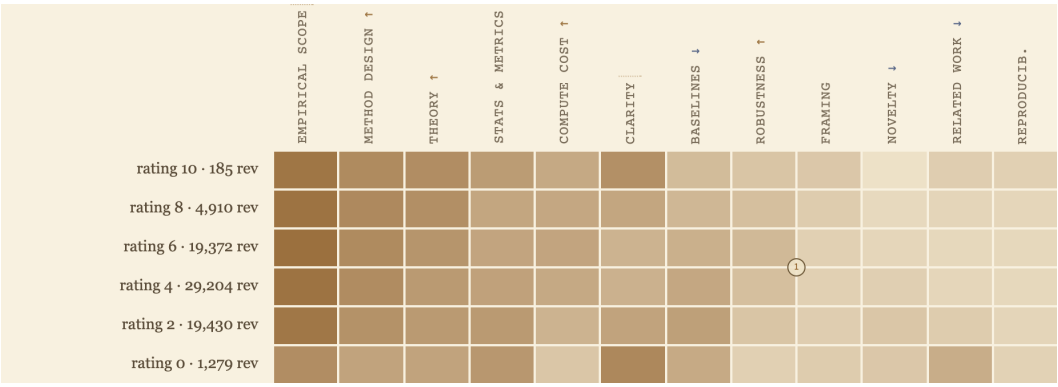


Figure 38: Object coverage by rating. Coverage means at least one extracted unit; a lack of extracted coverage does not prove that the reviewer never considered the object.

Praise at low ratings tilts toward writing and motivation: clarity accounts for 13% of positive units in rating-2 reviews versus 8% at rating 8, while theory rises from 13% to 16%. Topic diversity within a rating decreases toward the favorable end. Reviewer confidence is U-shaped (approximately 4.2 at rating 0, 3.4 at 6, and 3.9 at 10), and fix prevalence peaks in the middle (78% of rating-2–4 units versus 56% at 10). These patterns describe the recorded population and are not necessary properties of excellent or weak papers.

Figure 39 compares what low- and high-scoring co-reviewers inspect and criticize on the same papers. Objects occupy rows, and mirrored percentage extents show rating-2 co-reviewers on the left and rating-8 co-reviewers on the right, within papers that received both ratings. Pale wings give inspection of any valence and solid wings give negative evaluation. The extents measure the proportion of reviewers in each scoring group, rather than the volume of extracted text. Similar pale extents with different solid extents indicate shared inspection but differing stances. Across the 1,789 papers that received both a 2 and an 8, the two sides inspect theory and method at similar rates but differ in stance. Novelty is condemned by 21% of the low-rating reviewers and by 4% of their high-rating co-reviewers. The comparison separates differences in attention from differences in evaluation, subject to reviewer selection and the limits of extraction.

Panel disagreement raises a separate question: does score dispersion distinguish decisions beyond what the panel mean already says? For each multi-review paper we compute the rating range and mean, then compare acceptance between unanimous and divided panels at identical means. Exact conditioning is needed because within a broad mean band the unanimous and divided panels can have systematically different mean scores.

A unanimous panel gives the paper the same score in every review; a divided panel does not. Fixing the mean ensures that neither group has a more favorable central rating. The question is whether dispersion adds a visible decision distinction at that mean; the score range alone does not describe the content or severity of the underlying disagreements.

Figure 40 relates score dispersion to panel mean rating and compares acceptance at the same mean with and without disagreement. The upper plot has panel mean rating horizontally and mean within-paper score standard deviation vertically, in rating points, with one dot per mean-score group. The lower plot fixes an exact mean in each row and places acceptance percentage horizontally; open and filled dots compare unanimous and divided panels at that mean, with sample sizes printed beside the rows. The upper dispersion curve and the lower acceptance comparison are different quantities. Multi-review papers disagree in score in 93% of cases, and approximately half span at least four

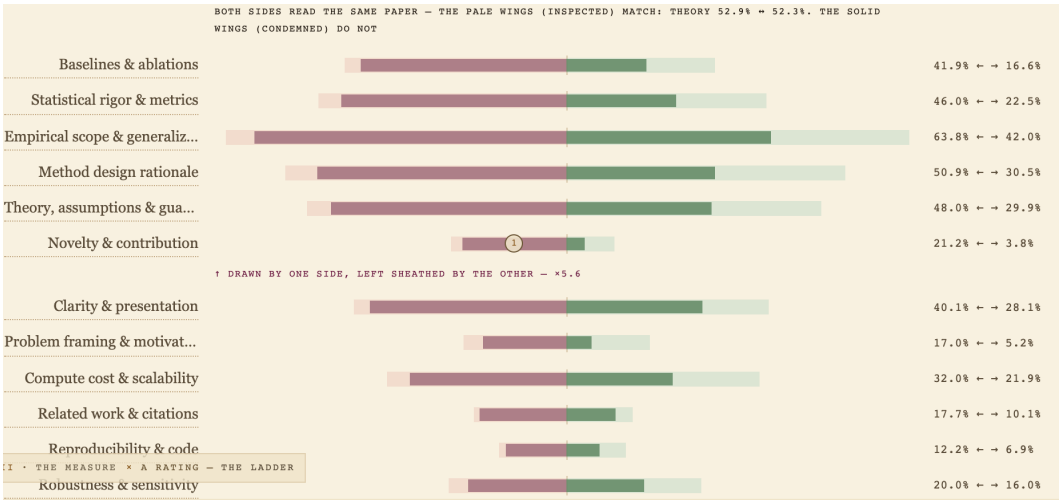


Figure 39: Same-paper rating-2 versus rating-8 comparison. Pale extent represents inspection; the darker extent represents condemnation.

points. Yet at an *exact* mean of 6.0, unanimous and divided panels have acceptance rates of 92.1% and 91.8%, and at a mean of 4.0 approximately 12.1% and 13.1%. At these two fixed means, consensus adds little separation in acceptance despite widespread disagreement in scores. A banded-mean comparison can confound dispersion with mean differences within the band. The exact-mean result is a narrow empirical contrast at these values, not proof that decision-makers ignore disagreement in general.

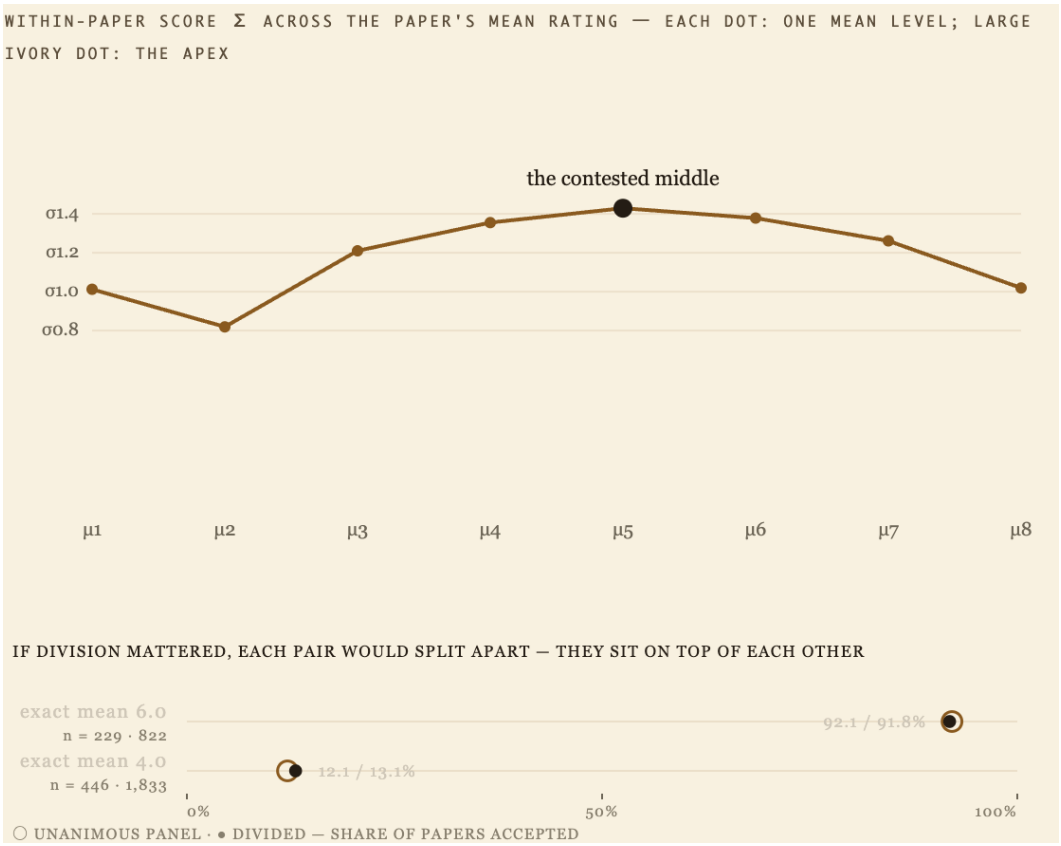


Figure 40: Score dispersion and acceptance. Interpretation requires exact conditioning on panel mean; comparing broad mean bands can create an apparent consensus association.

7 Response, deliberation, and the decision hierarchy

The preceding section related written evaluation to ratings and decisions. This section turns to the public process between an initial review and a decision: whether participants reply, whether a reviewer writes a changed judgment, how judgments combine within a panel, and how the chair represents them in a meta-review. A thread is the branch of replies attached to an official review, and a panel is the set of reviewers assigned to a paper. The two are different units of observation, and an active thread does not by itself imply a revised judgment or a changed panel outcome.

7.1 Reply structure without language-model interpretation

Before a written concession can be interpreted, we need to know whether a discussion took place at all. We reconstruct reply trees from the public metadata, classify each official-review thread by author response and reviewer return, and tabulate the classes by year and decision. The result is an engagement baseline that does not depend on extracting semantic units, and it separates an unanswered review from an author response that received no return.

A reply link gives the parent–child relation between notes, a signature gives the participant's role, and timestamps order the turns. The reconstruction records participation without judging whether a reply is persuasive or even substantive. The names below describe tree shapes only. A reviewer return means that the reviewer posts again, not that the reviewer agrees with the author, and a long branch can contain persistent disagreement.

All 199,031 official review threads can be classified from signatures, reply links, and timestamps. A stump has no reply, a monologue has an author response without a reviewer return, a return has one subsequent reviewer turn, and a volley has two or more. Rare asides involve other roles. Roles in early years depend on recovering historical signature conventions and are less certain.

Figure 41 shows how participation in official-review discussions changes across years. Year runs along the horizontal axis and the percentage of that year's official-review threads along the vertical one. Colored stacked bands distinguish no reply, an author response without reviewer return, and one or multiple reviewer returns, and the full height is 100% of threads each year. The line and crest annotations give the combined return/volley percentage. Band area is a yearly proportion, not an absolute number of messages or a confidence band. Reviewer return rises from approximately 9% in 2018 to 43.8% in 2025, then falls to 17.7% in 2026 as submissions grow by 70% and reviews by 62%. Stumps rise from 9% in 2021 to approximately 26% in 2026. In 2026, return/volley threads sit on papers accepted at approximately 50–51%, monologues on papers accepted at 39%, and stumps on papers accepted at 1.2%. In the last cycle, then, more reviews coexist with proportionally fewer reviewer returns, so volume alone does not describe reciprocal engagement. The correlations with acceptance may reflect author triage, abandonment, discussion rules, and review workload as well as engagement, and they do not show that a reply causes acceptance.

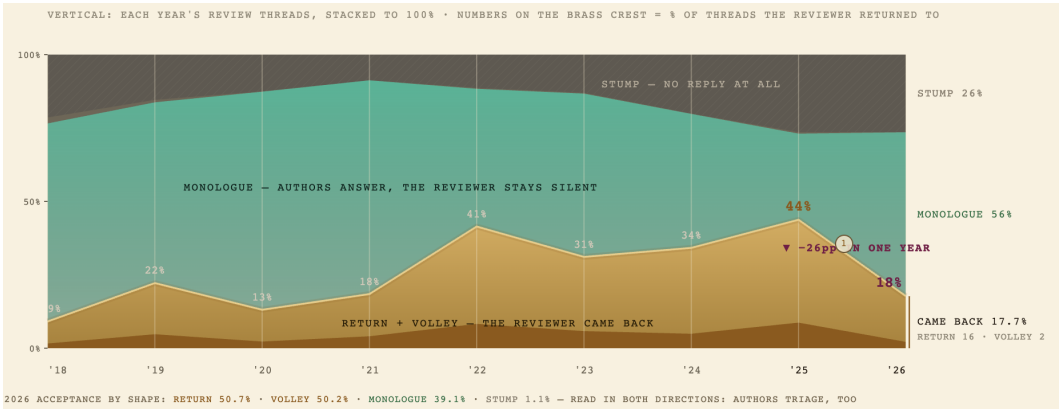


Figure 41: Reply-tree census across all official review threads. Counts do not rely on semantic extraction, but historical role reconstruction and changing discussion rules affect comparability.

Figure 42 gives five selected paper discussions as separate tree panels. Squares are official-review roots, later nodes are replies, edges are parent–reply links, and node color distinguishes participant roles. Horizontal separation arranges branches and vertical depth orders reply generations; neither axis measures calendar time or message frequency. Paper identifiers and decision labels identify the specimens. The five 2026 discussions illustrate extremes: three with the most reviewer returns among threads of six to fourteen notes, plus strongly monologue-dominated and stump-dominated cases with at least three reviews. The selection is purposeful, not representative.

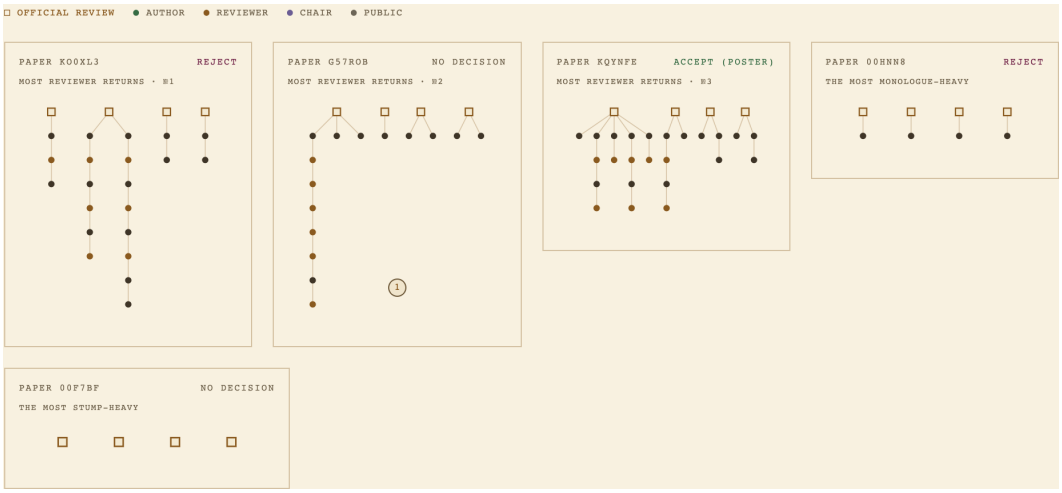


Figure 42: Five deliberately selected reply-tree specimens. These illustrate thread forms; prevalence is estimated by the census above.

7.2 Three denominators for post-response judgment change

Whether an author response changes evaluation has three observable versions: what later judgments say, which exchanges contain a concession, and what becomes of every original objection. Table 5 lists the corresponding counting units in rows, their observed counts in the middle column, and the question each denominator can answer in the last column. The counts refer to later evaluation units, observable author–reviewer exchanges, and original reviewer–object objections respectively; they are not nested stages whose differences measure attrition. Looking only at later judgments would omit the objections that are never discussed again. We therefore treat the three denominators separately: tabulating change labels and recorded triggers among post-response units, relating response features to exchange-level softening, and following initial reviewer–object objections into later discussion. All three describe visible written updating, with no random assignment of replies or tactics.

Written updating is distinct from an unobserved score change. The snapshot contains no complete rating history, and an author response can change a score without producing another unit of written reasoning. The forum protocol records whether a post-response judgment strengthens, weakens, reverses, clarifies, or has no legible fate, and these labels are interpretations of prose.

Change and valence are different properties. Strengthening means that a later judgment reinforces the earlier evaluation, so praise can be strengthened as well as criticism. Weakening reduces an earlier commitment, reversal changes its direction, and clarification makes a position more explicit without necessarily changing it. In the comparisons that follow, concession or softening covers weakening and reversal together. A recorded trigger is the reason for an update as represented in the extracted account, such as a reference to new experiments, not an independently assigned treatment.

Table 5: Different denominators answer different questions about response.

Counting unit	Number	Question
Post-response unit	130,650	What happens in judgments that are written again?
Observable exchange	67,290	Which authorial moves accompany a written update when the reviewer returns?
Original objection	551,595	What share of objections is ever revisited by the same reviewer?

Figure 43 summarizes recorded judgment changes after author responses and the triggers associated with them. The left panel partitions all 130,650 post-response units into held, strengthened, clarified, weakened, reversed, or unreadable fates, with colored band width giving the share of that full denominator. On the right, rows are recorded trigger groups, and each bar is renormalized to the units in that group with a classified strengthened, clarified, weakened, or reversed fate, in that order from left to right. The dashed vertical guide is one minus the unweighted mean softening share across the four trigger rows, which aligns a common non-softening boundary for comparison. Among post-response units, reinforcement is approximately 2.4 times as common as concession. The trigger analysis takes the moved judgments and groups their recorded triggers by keyword. Even when fresh experiments are named, strengthening is 2.9 times as common as softening; rare unspecified and bare-commitment triggers, affecting 163 judgments, are omitted from the display. Topic-specific softening rates lie around 3.9–5.4%. In the visible post-response record, continued discussion more often affirms an existing evaluation than relaxes it, even when new experiments are mentioned. These are judgments selected by having been written after a response, not a causal evaluation of experiments as a rebuttal tactic.

The panel-position comparison asks whether written updating is symmetric between the favorable and unfavorable evaluations of a shared paper. It links a reviewer's identity in the discussion to the rating attached to that review, then contrasts the low- and high-rating sides of panels with a large spread. The recorded rating only locates a reviewer relative to co-reviewers; the comparison recovers no complete numerical trajectory and does not show that a concession closes the score gap.

Figure 44 compares recorded softening and strengthening by a reviewer's position within the rating panel. Reviewer-position groups occupy rows, and the horizontal axis is the percentage of reviewer–paper profiles with at least one recorded change. Red dots mark at least one weakened or reversed post-response unit and brown dots at least one strengthened unit; a profile can satisfy both. The pale

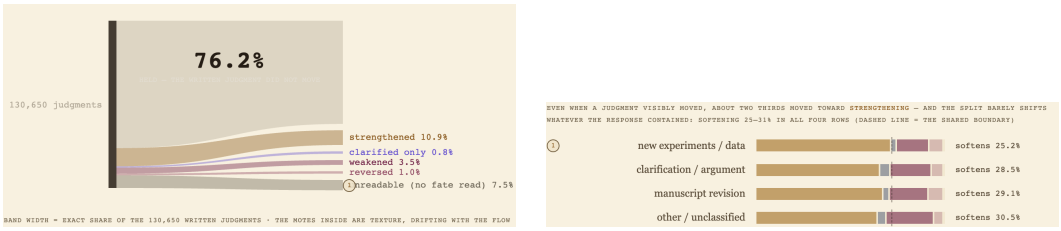


Figure 43: Left: Recorded fate of post-response units, keeping unclassified fates separate. **Right:** Fates by recorded trigger among the relevant revisited judgments. The denominator differs from original objections.

band spans the softening rates of the other groups and is not an uncertainty interval. Low, middle, and high refer to the reviewer's rating within a split panel, and the last rows give all-reviewer and unanimous-panel references. On the 9,329 split 2026 panels with a rating spread of at least four points, the low-rating reviewer weakens or reverses a judgment 2.4 times as often as the high-rating reviewer. Reinforcement still exceeds concession, and unanimous panels show little movement. An aggregate concession rate therefore conceals an asymmetry between opposing rating positions within the panel. Matching reviewer identities to ratings succeeds for approximately 95% of reviews, which adds one more selection constraint.

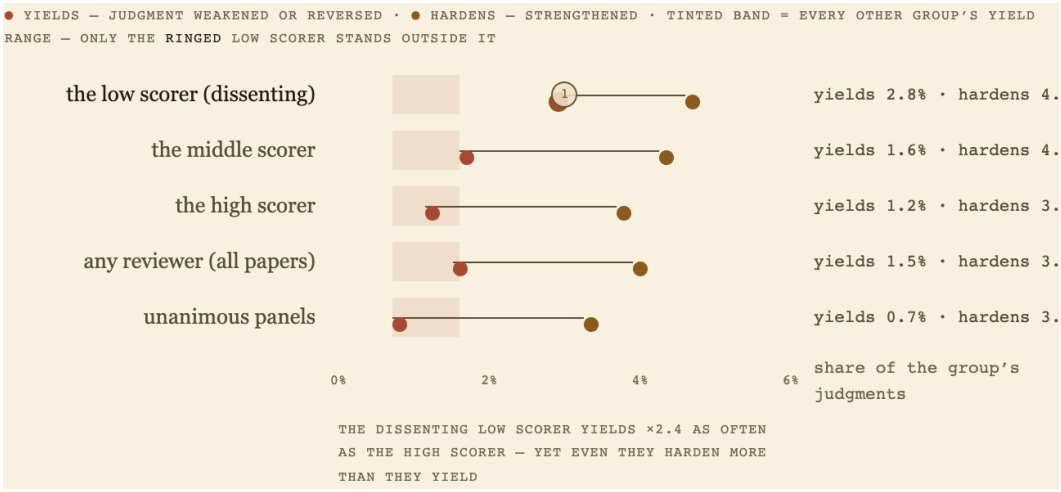


Figure 44: Post-response movement by a reviewer's position in a split panel. The denominator is reviewer–paper profiles, including profiles without a later statement; a profile qualifies for a movement if at least one eligible post-response unit records it. Silent score changes are not observed.

The exchange-level analysis asks which kinds of response accompany concession when a reviewer returns. A longer reply offers more opportunities both to detect a move and to address an objection, so an unadjusted comparison of moves may mostly measure response length. We first summarize softening by reply-length decile, then compare exchanges with and without each detected move within length quartiles. A sentence-clustering check asks whether a less prescriptive representation of replies yields clearer action categories than the deterministic patterns do.

An observable exchange contains an author response and a reviewer return, so that a later written judgment is available for comparison. Its softening outcome records whether any eligible unit weakens or reverses; it does not measure how much the reviewer's score changes. A detected move is a textual marker in the reply, not a judgment that the author has carried out the claimed work. A contest challenges the objection, a delivered-experiment marker refers to reported work, and a promise refers to work proposed for later. The categories can capture different amounts of language, which is why stratifying by length matters. Within a length group, the displayed contrast is the softening rate with the marker minus the rate without it.

Figure 45 relates recorded softening to reply length and to detected author moves. On the left, the horizontal axis gives the median reply length in words for each length decile and the vertical axis the

percentage of observable exchanges with any weakened or reversed judgment; the horizontal line is the overall softening rate. On the right, rows are detected author moves and horizontal position is the softening-rate difference, in percentage points, with versus without that move. Filled dots give the four length-quartile contrasts, open circles the raw contrast, and the bands span the quartile values rather than confidence intervals; zero is no difference. Seven authorial moves are detected with deterministic text patterns, and the outcome is whether any post-response unit weakens or reverses. Across 67,290 observable exchanges, approximately 6% soften. Softening rises from 4.6% to 9.6% across reply-length deciles, which is why move associations are also computed within length quartiles. Contest retains a positive association, approximately $+1.6$ to $+2.8$ pp in the three adequately populated strata and $+1.4$ pp in the shortest, which contains only 265 contests. Contest is detected in about 5% of exchanges, and detector breadth differs by move.

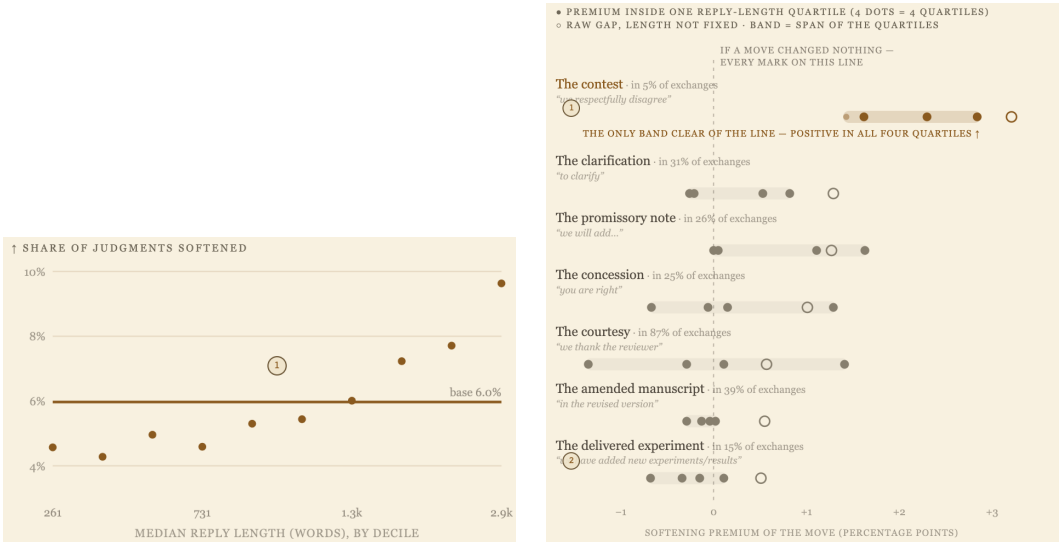


Figure 45: Left: Softening by reply-length decile. **Right:** Move-associated softening differences within reply-length quartiles, compared with raw differences. Chosen tactics and length are both non-random.

No other move keeps a consistently positive association after this coarse control for effort. Delivered experiments lie within ± 0.7 pp and are negative in three of four quartiles, and an object-specific analysis finds differences below half a point across the eight most common objects. Concession and thanks vary in sign, and promises keep a small non-negative residue. An embedding clustering of 722,406 reply sentences, drawn from 26,428 exchanges with a 30-sentence cap, separates topics more clearly than tactics and gives a similarly weak result. The raw ranking of tactics therefore partly reflects how much authors write, and after adjustment contest is the clearest remaining positive correlate of visible softening. None of this shows that running experiments is ineffective. Authors choose their move, stronger cases may invite contest, and the outcome omits silent score updates and exchanges in which no reviewer returns.

The exchange-level result conditions on a reviewer speaking again, so it cannot say how much of the original critical record ever receives a visible response from that reviewer. We now keep the initial objections in the denominator whether or not a later exchange appears. The link is the combination of paper, reviewer, and object, so a later evaluation of the same object counts as revisitation at that resolution without establishing that every detail of the original criticism was addressed.

Figure 46 shows how often initial objections are revisited and what change is recorded when they are. At the top, a full-width bar partitions all initial reviewer–object objections, including those never revisited. Below, rows identify objects and magnify the revisited fates on a common 0–14% horizontal range relative to all original objections on that object. Gray, purple, and green segments distinguish revisitation without movement, hardening, and softening, and the remaining objections lie outside the revisited slice. Right-hand labels give the revisitation percentage and the original objection count. Initial criticism is grouped by forum, reviewer, and object, and each group is checked for a post-response judgment by the same reviewer on the same object and for the change it records. The fate shares use all initial objections as the denominator rather than observable exchanges alone.

Of 551,595 reviewer–object objections, 92.1% have no recorded revisitation by that reviewer; approximately 6.5% are revisited without movement, 1.0% harden, and 0.4% soften. Revisitation by object runs from approximately 4–12%, highest for related work at 11.7%. Among revisits, hardening exceeds softening for every object, by approximately 1.5 times for novelty, 3.2 for clarity, and 2.5 overall. Visible revision thus reaches only a small fraction of the original critical record, and an analysis restricted to returning reviewers describes a far more selected process. Matching is at the reviewer–object level, not a validated one-to-one match of every semantically distinct complaint.

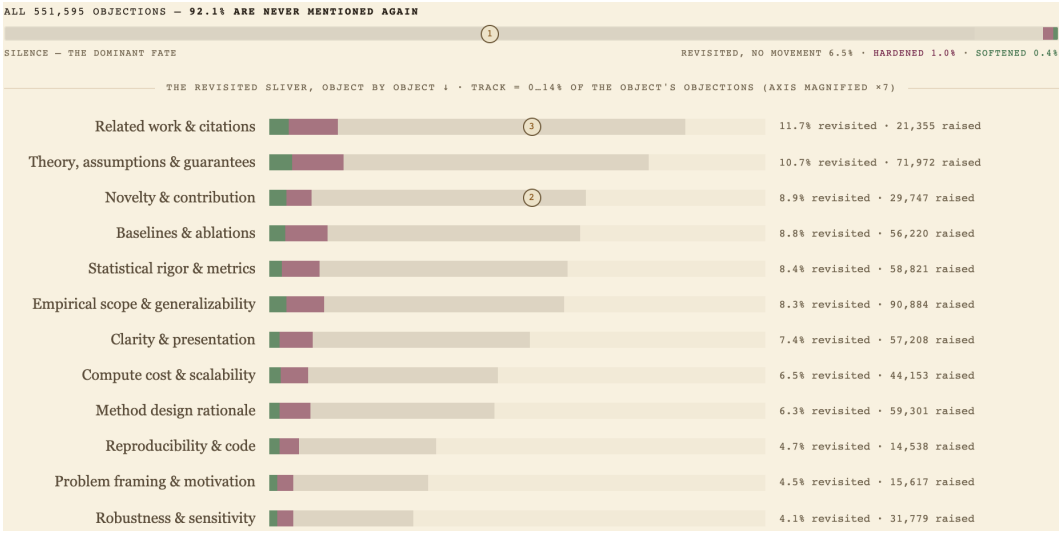


Figure 46: Fate of originally raised reviewer–object objections. This broader denominator exposes the large unobserved-after-response share. Silence is not automatically an unchanged private belief.

A further comparison asks whether an objection posed as a question has a different recorded fate. We locate objections in the 2026 form, classify them as placed in the questions field only, the weaknesses field only, or both, and compare revisit, hardening, and softening proportions across the groups. Question form is thus operationalized by field placement, not by an inferred degree of uncertainty or politeness.

Field placement is an observable way to separate a request for an answer from a criticism filed as a weakness, but it classifies by location: a question outside the questions field and an assertion inside it are labeled by where they sit. The comparison therefore tests the form-based distinction in the anchored subset, not the broader proposition that expressing uncertainty makes a reviewer more willing to update.

Figure 47 compares the later fates of objections placed in the questions field, the weaknesses field, or both. Its three rows are revisit, softening/reversal, and hardening. Colored dots distinguish the three placement groups, and horizontal position is the percentage of that group's original objections with the given fate. Each row has its own horizontal maximum, printed at the right, so distances are comparable within a row but not across rows on a common scale; the group sample sizes appear in the legend. In a separate 2026 line-anchored subset, 109,062 reviewer–object objections are located in form fields. Excluding 5,000 outside the weaknesses and questions fields leaves 104,062: 13,568 asked, 85,343 asserted, and 5,151 in both. Asked and asserted objections have nearly identical recorded fates, revisited in 5.1% versus 5.2% of cases, softened in 0.2% versus 0.3%, and hardened in approximately 0.6% of both. Objections in both fields are revisited in 7.4% and harden in 1.1%. Recorded fates differ little by field-defined question form; this does not establish that politeness or genuine uncertainty has no epistemic effect.

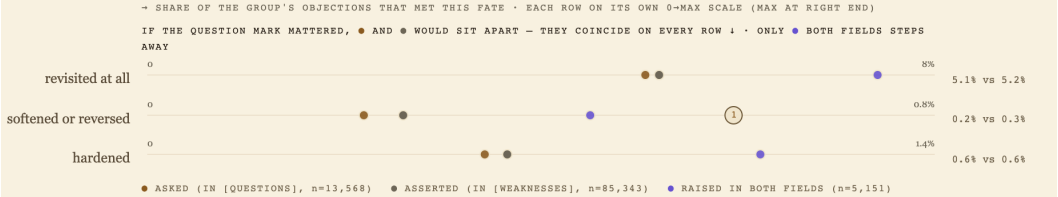


Figure 47: Post-response fates by questions-field and weaknesses-field placement. This comparison is confined to anchored 2026 objections, rather than the full nine-year denominator.

7.3 What a panel covers and where it disagrees

A panel can disagree because its members inspect different things or because they evaluate the same thing differently. For each paper–object combination touched by at least two reviewers, we record whether any co-reviewers take opposite stances and compute that proportion by object. Two further panel summaries concern aggregation: whether the decision departs from a simple panel-mean rule, and whether adding reviewers expands object coverage or duplicates it.

The stance comparison requires shared inspection, because an object missing from one review is not an opposite evaluation. The event counted is that the paper's reviewers include opposing stances on the object, so a panel can contain such an event without every reviewer disagreeing with every other.

Figure 48 lists objects in rows, with horizontal bar length giving the percentage of eligible papers whose co-reviewers take opposite stances on that object. Eligibility requires at least two reviewers to inspect the object, so each row has its own paper denominator, and the printed value is the disagreement rate rather than the object's prevalence among all submissions. Opposite stances are most common for problem framing (38.6%) and least common for robustness (16.3%). Sharing a subject of inspection therefore leaves substantial room for disagreement, especially about the paper's framing.

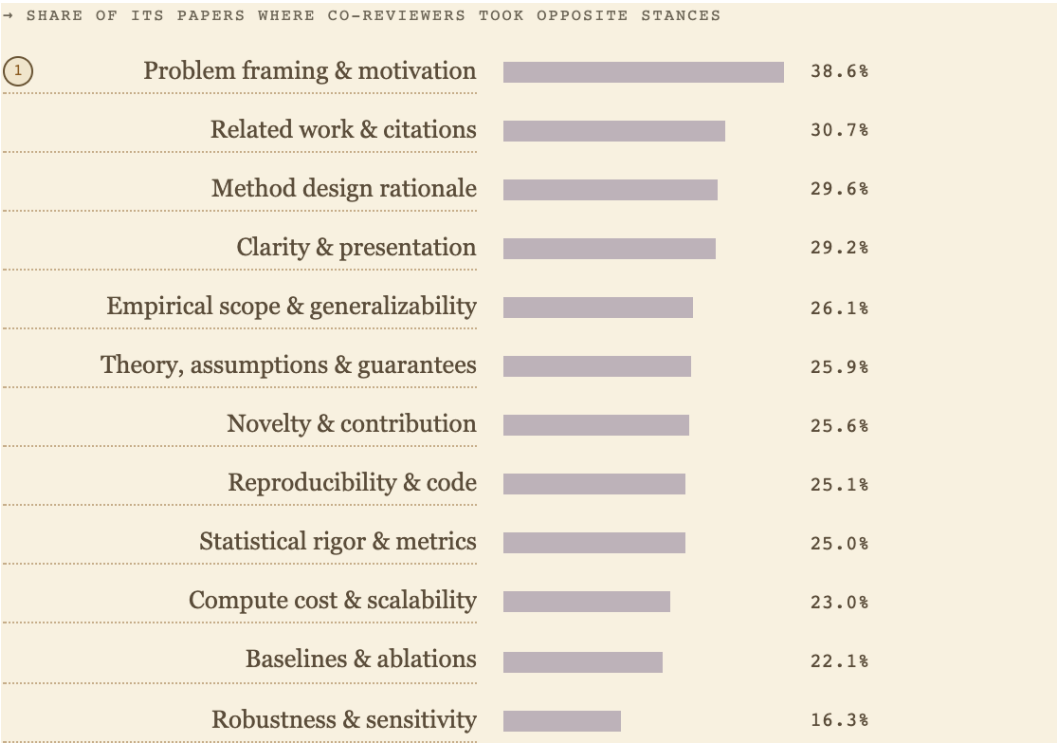


Figure 48: Object-specific stance disagreement among papers with at least two reviewers touching the object. The selected denominator differs by object.

Counting departures from panel scores requires a reference rule. The inferred rule accepts papers above a mean-rating cut chosen to reproduce the observed acceptance rate. We then count actual acceptances below the cut and actual rejections above it, excluding papers within a stated margin of the cut so that tiny boundary differences play little role. Comparing the two directions asks whether departures from this simple summary are symmetric. It does not assume that chairs used the summary as a rule.

Figure 49 compares actual decisions with the mean-rating cut rule and shows the direction of departures from it. The left panel has distance from the inferred cut, in score-point bands, horizontally and the percentage of decisions agreeing with the rule vertically. The right panel counts departures beyond the stated margin, with rejected-above papers in the left wing and accepted-below papers in the right. Wing length encodes paper count, printed at the tips, and the ratio of the two measures directional asymmetry rather than a probability for any individual paper. The comparison uses 13,700 decided 2026 papers with at least three rated reviews. An empirical mean-rating cut of 5.0 reproduces the acceptance rate. At a half-point margin around that cut, 919 papers are accepted below it and 338 rejected above it, a ratio of 2.7. Margins of 0.25, 0.75, and 1.0 give ratios of 2.4, 3.9, and 3.9, so the direction of the asymmetry persists across these margins. Relative to this reference rule, departures in favor of papers below the cut are more frequent than departures against papers above it. "Overrule" is shorthand for departure from an inferred rule, not a documented chair action or a claim that the mean ought to determine the outcome.

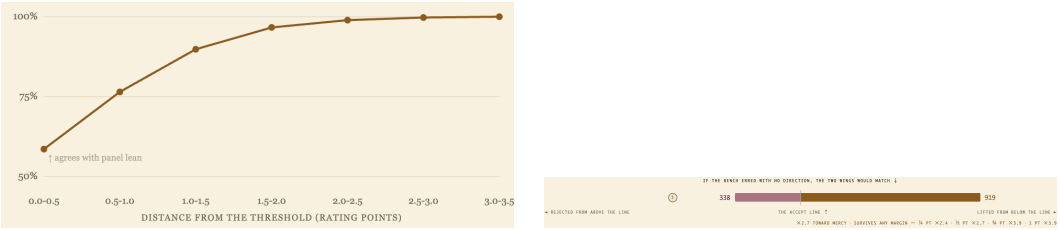


Figure 49: Left: Agreement with the inferred panel-mean rule as distance from its cut increases. **Right:** Counts of accepted-below versus rejected-above cases at the stated margin.

For the coverage comparison, each review is represented as the set of objects with at least one extracted unit. Counting how many panel members touch each object sorts objects into omitted, individually covered, and jointly covered, and the union and intersection of the sets summarize the panel's total reach and its common ground. The question is what the written record gains from multiple reviewers, without assuming that every object is relevant to every submission.

The union holds every object mentioned by at least one panel member and the intersection only those mentioned by all. An object covered by a single reviewer extends the panel's reach without adding to its shared coverage, while repeated coverage can supply contrasting evaluations of a common subject. The counts describe both possibilities without declaring either duplication or single-reviewer attention inherently beneficial.

Figure 50 shows how widely each object is shared as a subject of inspection within a panel. Rows are objects, and each stacked horizontal bar partitions eligible panels by how many reviewers inspect that object: none, one, two, or three or more. Segment width is the percentage of panels, so every row totals 100%. Darker brown means more reviewers share coverage, and the unseen percentage at the right is the zero-reviewer segment. These are panel frequencies, not unit counts. The analysis uses 18,074 rated 2026 panels of three or four reviewers, and coverage requires a single extracted unit of any valence. The mean panel touches 7.95 of the twelve objects; all its reviewers share only 0.83, and 4.05 are touched by no one. Over half of panels have at least three reviewers on empirical scope, while approximately two-thirds have none on reproducibility, and no reviewer mentions novelty in 56% of panels. The written panel record thus combines broad collective coverage with very little coverage shared by every reviewer. These are representations of written coverage: a paper may not make every object applicable, and an object absent from extraction is not evidence that no one considered it.



Figure 50: Panel coverage by the number of reviewers who inspect each object. Duplication and omissions are both visible, without treating all twelve objects as equally necessary for every paper.

7.4 Panel trajectories and selected complete deliberations

Per-unit updating can be rare while a panel still contains substantial exchange. To describe a discussion as a whole, we group matched reviewer trajectories by paper and summarize whether later units record strengthening, softening, reversal, or no movement. We compare these panel summaries with decisions and across years, and we contrast same-paper object-set overlap with random reviewer-profile pairs drawn within year. Selected complete trajectories then show the sequence and coexistence of events that the aggregate categories compress.

A trajectory is the ordered record of a reviewer's extracted initial and later judgments within a paper's discussion. Aggregating trajectories to the panel shifts the question from how often a unit moves to whether any movement is visible in the deliberation. The ending categories summarize that record; they do not establish a definitive collective belief. Silence means no later recorded contribution, whereas discussion without a detected change can contain substantive replies, and keeping the two apart is why thread engagement and judgment movement need not track each other.

Figure 51 has ending indicators in rows: softening, entrenchment, reversal, a large score split, and unanimity. Printed percentages and counts give the fraction and number of eligible panels satisfying each indicator, and bar lengths are scaled to the largest displayed share. Symbols identify the ending type. The rows overlap rather than partitioning the panels, so they cannot be summed into a total ending distribution. Among 17,848 rated 2026 panels with at least two matched reviewers, 85.7% show no recorded judgment movement. Where movement occurs, strengthened units outnumber weakened and reversed units, 3,543 to 1,336. At the panel level, then, discussion activity and visible revision of judgment are distinct, since discussion can occur without a recorded change in evaluation. Silence, procedural discussion, softening, entrenchment, and contested endings are interpretive categories that need not be mutually exclusive under every summary, and they should not be added without checking the classification rule.



Figure 51: Visible endings in the 2026 panel record. No recorded change includes both silence and discussion without a detected update.

Figure 52 shows the written evaluations and post-response changes in one selected paper's panel discussion, one horizontal row per reviewer. Circles are extracted evaluations in their written order, colored by valence; horizontal spacing is ordinal rather than elapsed time. The dashed vertical line separates initial units from those written after the author response. Upward and downward glyphs mark strengthening and weakening, and other change symbols appear where present. Curves between initial units connect the same object across reviewers, with darker curves for opposite stances; they are not one-to-one links from an original objection to its later reply. Fourteen deliberately selected cases show complete reviewer trajectories spanning reversals, entrenchment, large rating splits, and unanimity. They illustrate the course of the selected deliberations and how the instrument represents them, rather than estimating event frequencies. The example shown preserves the unit order, response boundary, and stance connections used to read a case, and the population analysis above supplies prevalence.

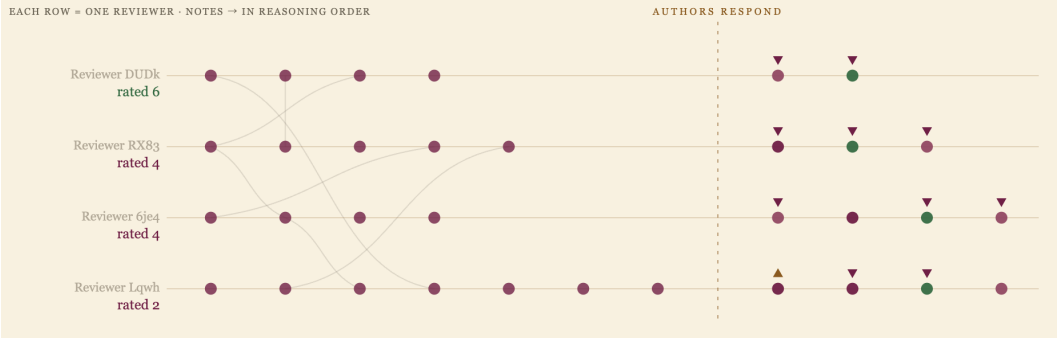


Figure 52: One of fourteen selected deliberation specimens. Rows represent reviewers, symbols represent extracted units, and the response boundary separates initial from later written evaluation. The source archive contains the complete selected case set.

Shared attention supplies a further reference for reading a panel. Pairing unrelated reviewer profiles from the same year shows how much object overlap arises from the year's general reviewing vocabulary and topic frequencies, and comparing co-reviewers with that baseline asks whether evaluating the same paper adds shared attention. Jaccard overlap treats an object as present or absent, so it measures neither how deeply either reviewer examines it nor whether they agree. A final comparison with revisitation asks whether an initially disputed object is later taken up, instead of presuming that discussion concentrates on disagreements.

Figure 53 presents acceptance rates by panel ending beside a comparison of co-reviewer object overlap with random pairs. The left panel has ending categories in rows, acceptance percentage horizontally, and larger bubbles for the more common categories among 2026 panels; the dashed venue line is overall acceptance. The right panel plots Jaccard overlap of inspected-object sets horizontally and the share of reviewer pairs in each overlap bin vertically. Separate curves compare same-paper co-reviewers with random cross-paper profiles from the same year, vertical dashed marks locate their means, and color distinguishes the groups rather than uncertainty. In 2026, panels ending in silence are accepted at approximately 27% and panels ending in entrenchment at 54%. Because strengthening can reinforce praise, the latter does not mean that hardening causes acceptance. For shared attention, initial-review object sets are compared by Jaccard overlap, the size of the intersection divided by the size of the union. The mean for co-reviewers of the same paper is approximately 0.27, against 0.21 for profiles paired at random within year (seed 7). Dissonant objects are revisited slightly less often than agreed ones. The overlap indicates modest shared attention above the random-pair reference together with substantial differences in what co-reviewers discuss; it is not a validated measure of reviewer complementarity.

Figure 54 shows how the recorded endings of panel discussions change across years, with year horizontally and the percentage of that year's panels vertically. Stacked colored bands represent the recorded ending categories, the unfilled region above them represents silence, and the full height is 100% of panels each year. The procedural band is discussion without recorded judgment movement. The graph therefore counts panel endings, not posts, returning reviewers, or changing units. Across 49,441 panels in the forum track, silence falls from 51% in 2018 to 20% in 2022 and returns to 52% in 2026. Recorded movement reaches approximately 29% of panels in 2022 and falls to about 13%

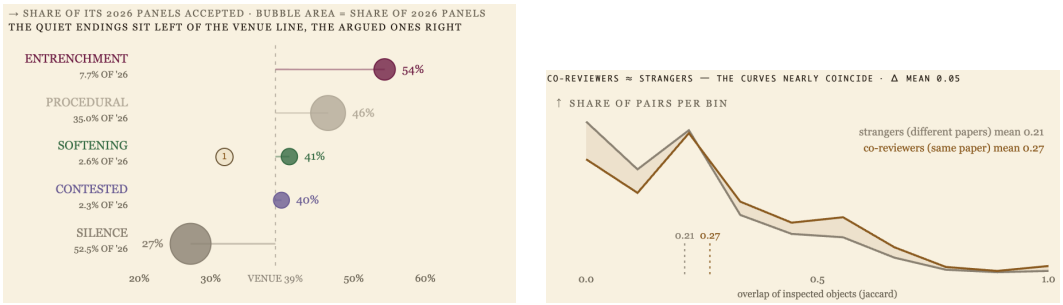


Figure 53: Left: Acceptance associated with panel-ending categories. **Right:** Co-reviewer object overlap versus randomly paired reviewer profiles within year.

in 2026. Visible deliberation in the panel record thus expands and then contracts rather than rising steadily with venue growth. Format, institutional rules, volume, paper composition, and model reading all vary together, so the trajectory does not isolate a workload effect.

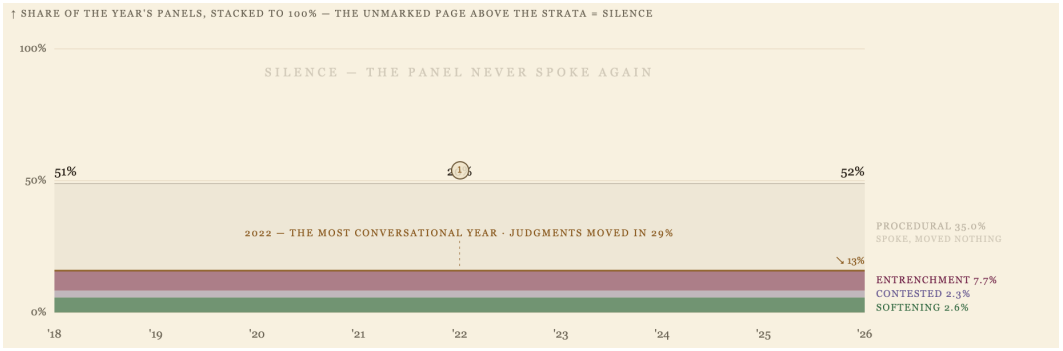


Figure 54: Discussion endings by year. These panel-level rates differ from thread-return rates and from the per-unit updating denominator.

7.5 Area-chair evaluation and textual proximity to reviewers

An area chair synthesizes a panel and justifies a decision, and that account need not reproduce the distribution of inspection across individual reviews. We compare object and valence shares in meta-review units with those in official-review units. To examine decisions near the margin, we separately compare object-specific criticism between accepted and rejected borderline papers within narrow panel-mean bins. The first comparison describes attention by role; the second is an outcome association conditional on a similar panel score.

A meta-review is the chair's account of the evaluation and decision, read here as a document with its own extracted units. Attention shares divide units about an object by all eligible units for that role and do not count how many chairs independently inspected the object. In the borderline comparison, grouping papers by nearby panel means supplies a common score context before criticism and decisions are compared. The aim is to ask what distinctions remain near a similar numerical recommendation, bearing in mind that a chair's prose may be written to explain a decision already reached.

Figure 55 is a slope comparison with official reviews on the left and meta-reviews on the right. Each line follows one object, and its endpoint height is that object's percentage of all extracted units for the corresponding role. Brown marks increases and purple decreases, and right-hand multipliers compare the two role shares. The horizontal dimension identifies roles rather than time, and the lines do not pair individual reviewers with chairs. The forum track includes 56,736 meta-review units, approximately 90% of them from 2024–2026, against 952,856 official-reviewer units. Statistics and metrics rise from roughly 10% of reviewer attention to 26% in meta-reviews, related work from 5% to 16%, and novelty approximately doubles. Empirical scope, theory, and method design fall to less than half their reviewer shares. Meta-review units are approximately 30% positive versus 19% for

reviewer units, and across the nine years they are 13.7 points less negative and 10.5 points more positive than official-review units. Genre plausibly contributes, since a meta-review summarizes a decision rather than repeating every inspection. The distributions are consistent with selective synthesis at the decision stage rather than a proportional inventory of everything reviewers inspected, and the differences mix institutional role and writing genre with differences in what is evaluated.

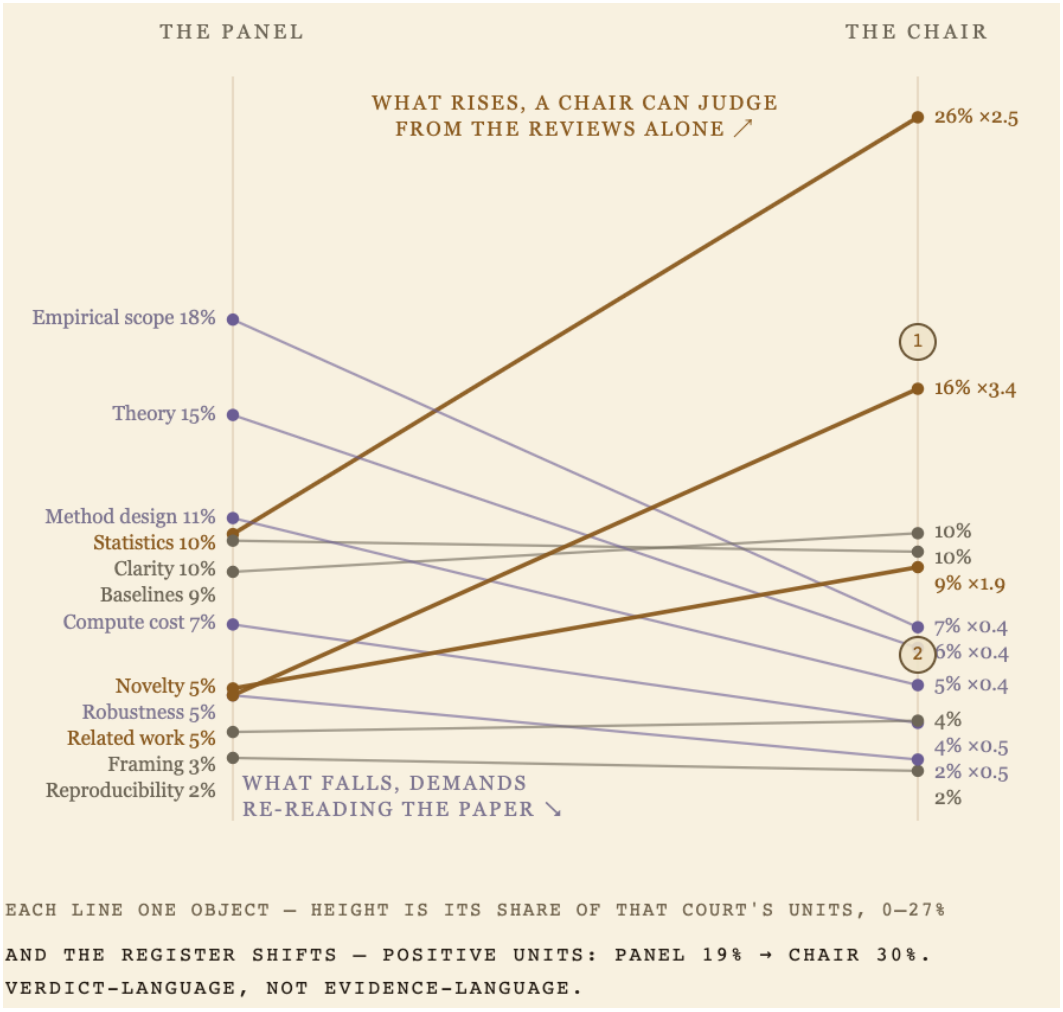


Figure 55: Attention in official reviews and meta-reviews. Uneven temporal availability and genre differences qualify comparisons between roles.

Figure 56 compares the objects criticized or praised in meta-reviews of accepted and rejected papers at similar panel means. The left panel has objects in rows and, horizontally, the matched difference in criticism prevalence, accepted minus rejected, in percentage points. Dots are estimates, whiskers are paper-bootstrap 95% intervals, and the dashed zero line is equal prevalence. The right panel has objects in rows and, horizontally, the percentage of each decision group's meta-review units that praise the object; green and neutral dots compare accepted and rejected papers at the stated borderline means, and the connectors show differences rather than intervals. For 6,547 decided borderline 2026 panels with mean ratings of at least 3.5 and below 5.0, comparisons are made within 0.25-point mean-rating bins. Within each bin, the proportion of accepted papers with an object-specific criticism is compared with the corresponding proportion of rejected papers, and the differences are then aggregated. Only clarity has a paper-bootstrap interval excluding zero: after matching, clarity criticism is 4.4 percentage points less prevalent among accepted than among rejected papers. This is a difference in criticism prevalence conditional on the decision, not a difference in acceptance probability conditional on criticism, and the other object-specific differences are small. On accepted borderline papers approximately 75% of meta-review units are praise, versus 8% on rejected ones.

Near the numerical margin, then, decision justifications differ sharply in how much praise they contain even though most object-specific criticism differences are small after matching. Because the chair writes the justification for the decision, the contrast is not independent evidence of the paper's quality.

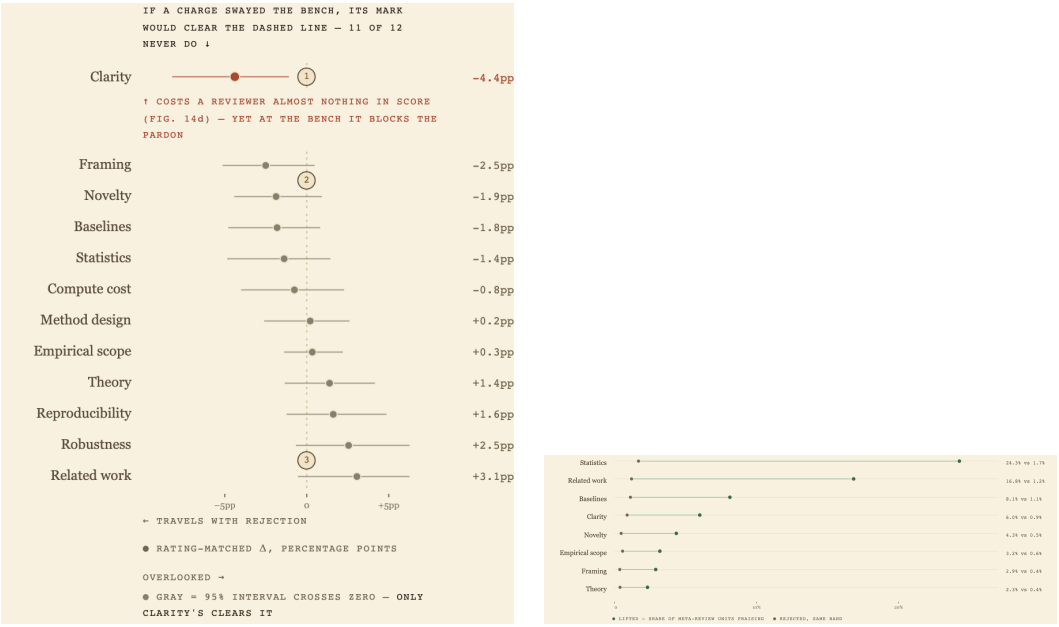


Figure 56: Left: Differences in criticism prevalence between accepted and rejected papers at comparable borderline panel means. **Right:** Objects praised in meta-reviews for accepted and rejected borderline papers.

A different question concerns textual proximity: which parts of a meta-review most resemble arguments the reviewers already wrote, and whose wording is nearest? We embed the reasoning units, match each meta-review unit to its nearest official-review unit on the same paper, and summarize threshold exceedances by valence. We then count how the matches are distributed across reviewers and rating sides. Matching within a paper keeps the comparison relevant to that panel, but it leaves shared subject matter as an alternative explanation to direct borrowing.

An embedding represents a unit's wording as a numerical vector, and cosine similarity compares the directions of two such vectors. For each meta-review unit, the nearest match is the reviewer unit with the largest similarity among that paper's candidates. A sufficiently close match is called an echo, and a higher cutoff is reserved for near-copies. Both names describe similarity in the extracted representations. A very close match can concern an observation that the paper itself makes obvious, and a summary written in different terms can still rest on the reviewers' arguments.

Figure 57 shows how closely meta-review evaluations resemble reviewer evaluations of the same paper. All meta-review units, negative units, and positive units form separate rows, each totaling 100% of its own units. Horizontal colored lengths classify the nearest same-paper reviewer-unit cosine similarity as below 0.75, from 0.75 to below 0.90, or at least 0.90, labeled own wording, echo, and near-copy in the source. The stacked widths are proportions of units, not amounts of text, and the threshold names do not establish provenance. Textual proximity is measured in 8,439 papers with at least three meta-review units and two official reviews. Each of 31,598 meta-review reasoning units is matched to its nearest reviewer unit on the same paper. At cosine thresholds of 0.75 and 0.90, approximately 73% fall below the echo threshold and 0.5% are near-copies. Negative meta-review units are echoes more often than positive ones (28.8% versus 22.4%), and the gap persists at cutoffs of 0.70, 0.75, and 0.80 (+7.0, +6.6, and +3.3 pp). Under this convention most meta-review units do not closely match any reviewer unit, and the matches that occur are more common for criticism than for praise. The thresholds are similarity conventions, not labels proving authorship or influence.

Once matches are assigned, concentration asks whether they spread across the panel or point repeatedly to one reviewer. The directional comparison then uses the matched reviewers' ratings to describe which side of a split panel the echoes favor. Weighting by echoes gives more weight to

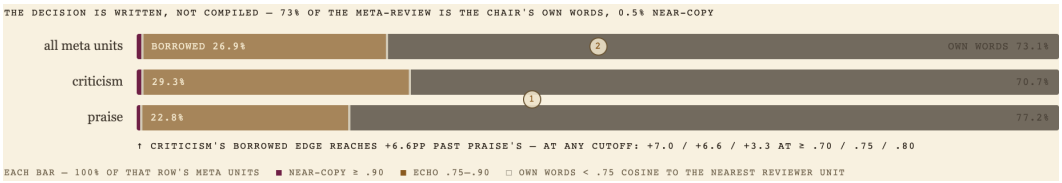


Figure 57: Meta-review units by nearest reviewer-unit similarity and valence. "Own wording" means below a threshold, not that the chair used no information from reviewers.

a reviewer whose units are matched more often, without estimating how much authority the chair granted that reviewer. Requiring complete ratings and excluding unanimous panels makes the direction interpretable but restricts the cases to which the comparison applies.

Figure 58 shows how meta-review echoes concentrate on individual reviewers and how they align with ratings. The upper left plots the largest single reviewer's share of a paper's echoes horizontally and the number of eligible papers per bin vertically. The upper right gives the percentages of panels whose most-echoed reviewer is uniquely lowest, middle, uniquely highest, or tied at an extreme in rating. The bottom rows separate rejected and accepted papers; their horizontal axis is the percentage of papers whose echo-weighted rating lies on the harsher side of the panel mean, with 50% as the even-direction reference and horizontal 95% intervals. The three panels have different denominators and axes. Where at least three units echo a reviewer, approximately 60% of echoes concentrate on one voice. In fully rated, non-unanimous panels, the echo-weighted side favors the harsher side 60/40 for rejected papers but only about 51/49 for accepted papers. The analysis uses 3,334 eligible panels, and the directional decided subsets are 2,220 rejections and 752 accepts. The most-echoed reviewer is uniquely lowest-rated in 16% of panels and uniquely highest in 13%, and ties on the coarse scale leave approximately half of panels uninformative for that comparison. The matches thus portray an uneven textual connection to the panel, with rejection-side echoes leaning toward harsher ratings. A chair and a reviewer may describe the same paper independently, so nearest-neighbor similarity cannot establish that borrowing occurred.

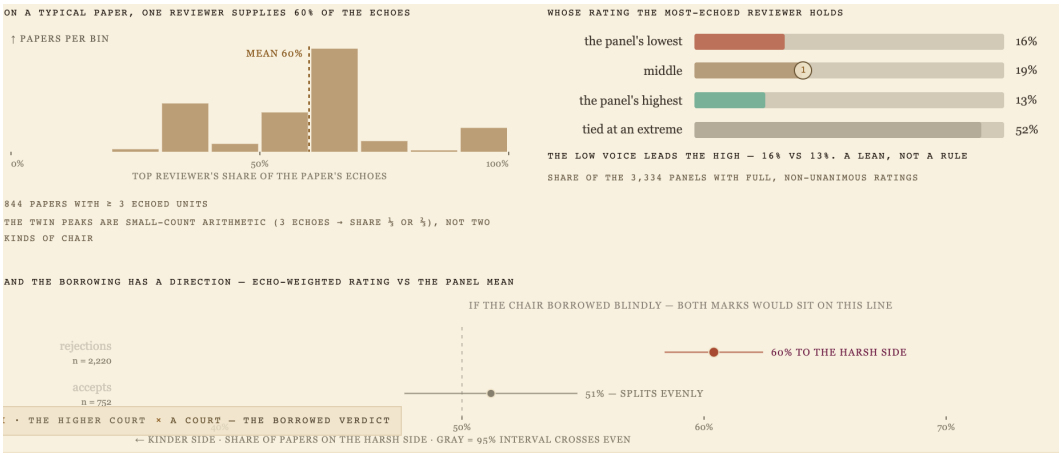


Figure 58: Concentration and direction of meta-review echoes. Thresholds, eligibility, and score ties matter; the stronger directional statement concerns the rejection-side 60/40 comparison.

8 Ratings as measurements

The preceding sections use ratings as outcomes, as comparison variables, and as summaries of a panel. What kind of measurement do those numbers provide? This section examines four sources of variation in turn: when a review is filed, which score categories the form offers, how far reviews of the same paper disagree, and what a simulated replacement panel would decide. None of these checks treats a numerical score as a direct observation of scientific quality. Read together, they show how much can be inferred once a written evaluation has been reduced to a point on a scale.

8.1 Filing time and changes in the rating instrument

Before treating ratings as repeatable measurements, we ask whether the time at which a review was filed is associated with what the review says. Each 2025 and 2026 official review is placed in a bin by the days between its first posting and that year's filing peak. Within each bin we compute the mean rating and median word count for both years, and the mean extracted-unit count for 2026 only, expressing rating differences relative to the same year's peak-day mean. The comparison tests whether filing-time groups differ, without equating the two years' rating units.

The filing peak supplies a within-year time reference, so the analysis compares earlier and later reviews inside the same cycle rather than calendar dates across cycles. Rating, word count, and unit count capture different properties: the numerical evaluation, the length of the document, and the amount of evaluation the extraction represents. Displaying them together shows whether reviews filed at different times differ in both assessment and elaboration. The design does not follow the same reviewer across filing dates, and the peak is an observed concentration of postings rather than an assigned treatment.

Figure 59 compares review ratings and length across filing-time groups in two review cycles. Both panels place filing-time groups on the horizontal axis, measured relative to each year's observed posting peak: progressively earlier groups, the peak day labeled deadline, and later reviews. Brown and gray curves identify 2026 and 2025. In the left panel the vertical axis is mean rating minus that year's peak-day mean, in rating points; its zero line is the peak-day reference, and whiskers are bootstrap 95% intervals around each bin mean. In the right panel the vertical axis is median words per review. The categories are ordered time bands rather than equally spaced days, and the two panels have different vertical units. In both 2025 and 2026, reviews filed at least two weeks before the filing peak carry lower mean scores and run longer than peak-day reviews. Within each year's own scale, the mean difference is approximately -0.34 in 2025 and -0.48 in 2026, with about fifty additional words. In 2026, the mean extracted-unit counts are 5.68 versus 5.40. Reviews filed at different times thus differ in both assessment and elaboration, so filing-time groups are not descriptively interchangeable. The remaining analyses do not control for filing time. Timing may stand in for reviewer behavior, paper difficulty, or selection; the data do not show that filing early itself lowers a score.

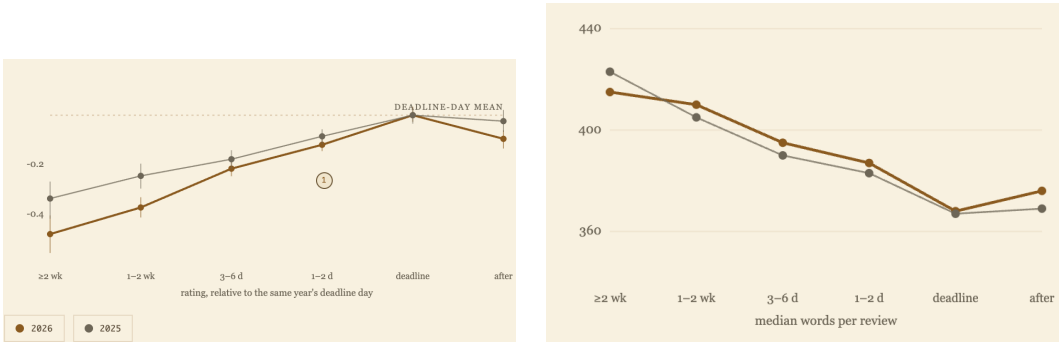


Figure 59: Left: Mean score relative to the same year's peak-day mean, by filing time. **Right:** Median review length by filing time. Separate-year normalization avoids equating rating units across changed forms.

Four scoring forms occur in the public record (Table 6), so historical rating means and raw point gaps do not share one ruler. Confidence fields also change; 2020 substitutes an experience assessment. Soundness, presentation, and contribution sub-scores are available only from 2024 onward; the construct check uses the 2026 form.

In Table 6, each row is a group of years that used a common rating form. The next column lists the values actually offered, and the final column records their verbal anchors or other qualifications. A row describes an instrument, not how often each score was used, and the same numeral in two rows need not express the same recommendation. The endpoints of the scale are only part of the problem. Removing intermediate values changes which judgments can be expressed, and changing the descriptions attached to values changes the reference available to a reviewer. Within-year standardization rescales a number relative to the year's observed mean and spread; it cannot recover labels that were never offered or show that equally standardized scores express the same recommendation.

Analyses that pool years therefore need a different interpretation from comparisons made entirely inside a single rating form.

Table 6: Observed rating forms. Standardizing within year does not recover a shared semantic threshold.

Years	Observed scores	Qualification
2018–19, 2021	1–10	Labeled ten-level rating; 5/6 straddle an acceptance threshold.
2020	1, 3, 6, 8	Reject / weak reject / weak accept / accept; no ordinary confidence field.
2022–25	1, 3, 5, 6, 8, 10	Recommendation then rating; labels through 2024, bare values in 2025.
2026	0, 2, 4, 6, 8, 10	Bare even values; no shared threshold anchor with the earlier form.

8.2 Within-paper variability and an observational redraw

How different might another review of the same paper be? We approach the question in three ways: partitioning rating variance within each year, tabulating one reviewer's score conditional on the other reviewers' mean, and simulating an alternate panel under an explicit variance model. The first describes aggregate variation, the second exposes the empirical conditional distribution, and the third translates model assumptions into a decision-change frequency.

The variance comparison separates how much ratings vary between papers from how much reviewers of the same paper differ. In the expression below, Y_{ij} is review j's rating of paper i. The overall mean μ is common to the year, the paper term u_i is shared by its reviews, and the residual ϵ_{ij} is the departure of an individual rating from that paper's modeled value. A large between-paper share means that knowing which paper was reviewed accounts for more of the observed variation; a small share leaves more variation among ratings within papers.

Figure 60 shows the estimated share of rating variation attributable to differences between papers in each year. The horizontal axis is year and the vertical axis is the estimated intraclass correlation, a dimensionless fraction of total rating variance attributed to differences between papers. Each dot is one year's estimate; the dashed horizontal reference at 0.5 marks equal between-paper and within-paper shares. The connected dots describe yearly estimates, not a confidence band, and the changed rating forms in Table 6 qualify their comparison. We express this intraclass correlation (ICC) through the random-effects decomposition

$$Y_{ij} = \mu + u_i + \epsilon_{ij}, \qquad \mathrm{ICC} = \frac{\mathrm{Var}(u_i)}{\mathrm{Var}(u_i) + \mathrm{Var}(\epsilon_{ij})}. \tag{1}$$

The residual combines disagreement, differences in expertise, interactions, and other unmodeled factors. It cannot be interpreted as a measure of reviewer incompetence, nor does the decomposition separate objective paper quality from human error. Estimated ICC is 0.34–0.45 in 2018–2025 and 0.18 in 2026. Under this decomposition, most rating variation lies within papers rather than between them. Changes in scale resolution and use can move the estimate, so the lower 2026 value should not be read as a comparably measured collapse in reliability. The source flags the four-level 2020 form for the same reason.

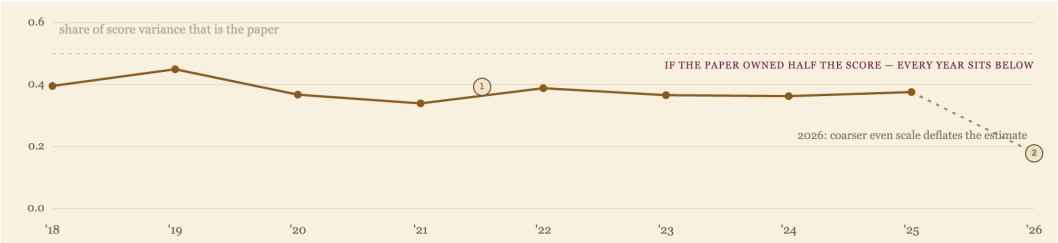


Figure 60: Estimated within-year ICC. The 2026 point is qualified by its changed scoring form. The denominator includes all observed rating variation, not only differences attributable to reviewer identity.

The variance decomposition compresses each year into one number, so we next return to the observed scores themselves. Conditioning on the co-reviewers' mean deliberately leaves the focal rating out of its own reference; otherwise that rating would help define the group against which it is compared. Each displayed distribution shows the scores actually observed among reviews whose co-reviewers had similar means.

Table 7 fixes the rounded mean of the other reviewers in each row and distributes the focal reviewer's score across the six rating columns. Entries are percentages of the rating observations in that row and sum to approximately 100% after rounding. The final n column counts the rating observations contributing to the row; several can come from one paper. The mass spread across rating columns shows how much uncertainty remains after conditioning.

Table 7: Next-reviewer rating probabilities (percent), conditional on the rounded mean of the other reviewers in 2026 papers with at least three ratings. Bins have at least 400 observations. Counts are rating observations; several can come from one paper.

Other mean	0	2	4	6	8	10	n
1	17.2	51.1	22.0	8.3	1.2	0.3	1014
2	5.4	46.9	31.3	14.0	2.4	0.1	4475
3	2.0	33.8	39.8	20.2	4.1	0.1	18928
4	1.1	25.4	41.5	26.2	5.7	0.2	17895
5	0.8	20.3	40.7	30.1	7.8	0.3	23241
6	0.6	16.3	38.3	32.8	11.4	0.5	6895
7	0.5	12.9	33.1	37.6	15.1	0.8	2997

Figure 61 illustrates the spread of scores within a panel and relates that spread to acceptance in the observed population. The upper part places one illustrative panel's individual scores on the horizontal 0–10 rating scale, marks its mean, and spans its maximum-minus-minimum range. The lower part groups the observed panel population by that range on the horizontal axis. Tan bar height is proportional to the square root of the number of panels in each group, and the green portion occupies the group's accepted fraction of that displayed height; printed percentages give acceptance rates. Because the heights are square-root transformed, they cannot be read as raw accepted-paper counts, and the upper example illustrates one panel rather than a second population estimate. Grouping panels by range gives the following contrast: spread-eight panels accept at 47% (168 panels), versus 30% for unanimous panels (1,319). These are unconditional rates. The groups differ in mean rating as well as in spread, so the contrast does not show that disagreement benefits a paper.

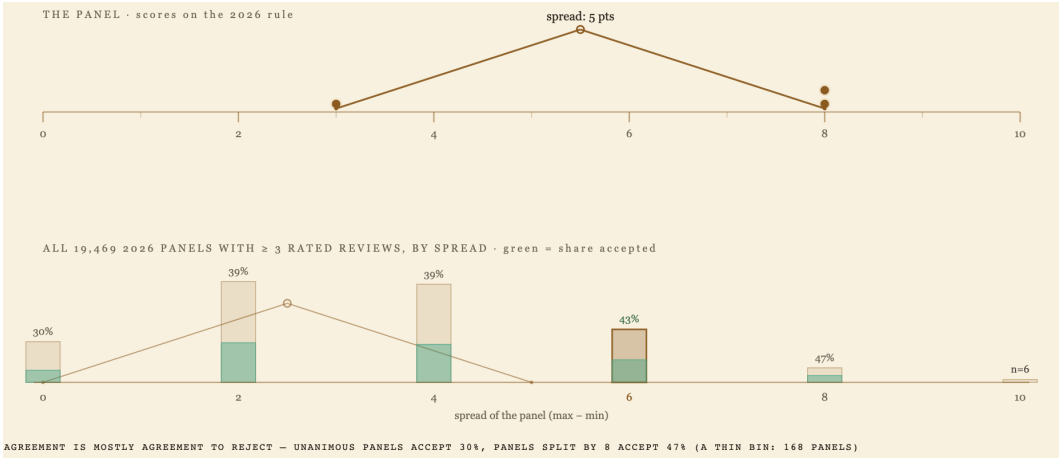


Figure 61: Panel spread and its associated acceptance rate. The exact-mean control in Section 6 is needed to separate dispersion from mean composition.

A separate corpus-wide check asks whether a year's ratings concentrate on a few values or spread across all the available ones. Normalized rating entropy summarizes the yearly score frequencies as $-\sum_r p_r \log p_r / \log K$, where p_r is the observed share of reviews using rating r and K is the number of rating levels observed that year. Because the measure pools ratings across papers, it

describes score use rather than within-paper agreement. It is approximately 0.74–0.82 in most years and 0.92 under the four-level 2020 form. The normalization adjusts for the number of levels, but not for changes in how those levels were used or understood.

The redraw turns the repeatability question into an explicit model exercise. The observed panel mean is an uncertain estimate based on a few reviews. Rather than treating an extreme observed mean as an error-free property of the paper, shrinkage moves it toward the venue mean by an amount set by the variance model. A replacement panel of the observed size is then generated around that estimated value, and its mean is mapped to a decision using the acceptance pattern observed in the data. Because that pattern allows different decisions at similar means, the simulation includes variation at the decision stage as well as among reviewers.

Figure 62 shows the model-estimated probability that a paper's decision changes under a newly drawn reviewing panel. The horizontal axis is the observed panel mean rating and the vertical axis is the simulated percentage of redraws that change the decision. Brown and gray curves identify the 2026 and 2025 models; the dashed 50% line is a reference probability. Pale bars along the bottom summarize where decided papers lie on the mean-score axis, on a separate display scale. Curve height gives a conditional flip probability rather than a number of papers, and each yearly curve keeps its own rating scale. Across 4,000 redraws, the probability of changing a decision peaks near 54% around an observed mean of 5.75. Under these assumptions, 37% of 14,175 decided 2026 papers change outcome, including 58% of the 5,359 accepted papers. The corresponding accepted-paper figure for 2025 is 45%.

These results use two denominators. The all-paper rate counts changes in either direction, whereas the accepted-paper rate conditions on acceptance in the observed process and asks how often that status is lost in a redraw. Near a decision boundary, a modest change in either the panel mean or the decision draw can flip the outcome more readily than at a clear extreme. The reported rates describe this conditional simulation; they do not measure how often the conference's actual decisions were mistaken.

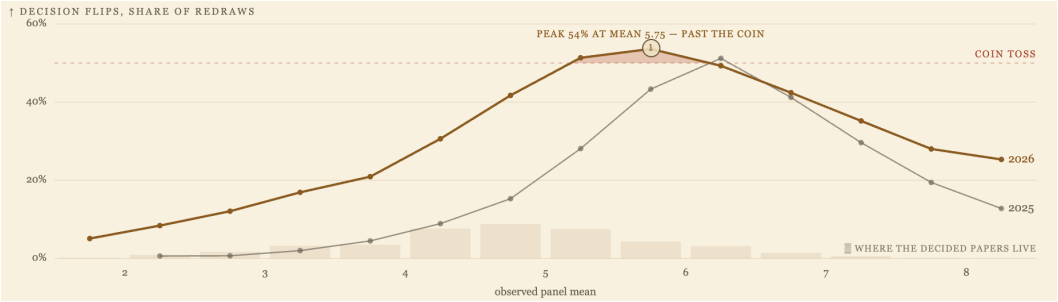


Figure 62: Simulated probability of a changed decision under a new panel and an empirical decision rule. The redraw includes decision-layer variation as well as panel variation; it is conditional on the shrinkage and noise model.

Figure 63 shows one cell for each originally accepted 2026 paper, ordered by observed panel mean from the strongest accepts toward the margin and wrapped onto successive rows. Filled brown cells keep their accepted status in this draw; outlined reddish cells lose it. Row position only reflects where the ordered list wraps and carries no second measured variable. The display follows the survival of the original accepted set; papers newly accepted from the originally rejected set are not shown. The accepted-paper flip rate falls within the range of the actual repeated-committee NeurIPS experiments [2, 8], but the resemblance does not validate the simulation. The experiments convened a second committee; this analysis estimates one from a single observed process, and the model may not reproduce review assignment, correlated reviewer errors, paper-specific uncertainty, or changed decision rules. The survival display is one realization of that same model.

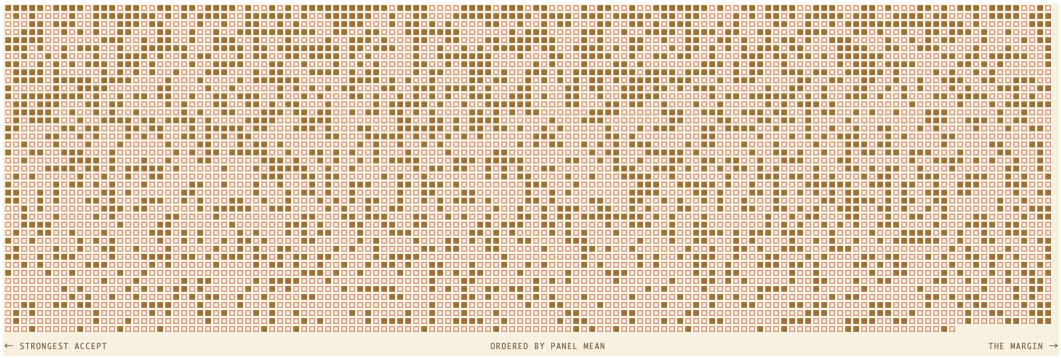


Figure 63: Survival of the originally accepted 2026 papers in one draw under the same simulation. Filled cells retain acceptance and outlined cells lose it; newly accepted papers are not shown. This is an illustrative realization.

9 Change across years, fields, and the LLM era

The preceding analyses separate the objects of evaluation, the standards applied to them, and what later happens to a criticism. We now ask how far those descriptions depend on when and where a review was written. The historical comparisons distinguish changes in how often an evaluation occurs from changes in how it is formulated, and the field comparison asks whether a venue-wide average represents every research area. The final analyses return to raw review language to examine changes associated with the LLM era. These analyses use different units and address different aspects of the changing record, so their results need to be interpreted separately.

9.1 Attention changes more than the conditional grammar

The earlier summaries might describe a recent configuration rather than a persistent feature of the venue. To check, we tabulate object and standard counts separately for each year of the forum track and divide by that year's unit total. Expressed as shares, the counts separate a change in the allocation of written attention from growth in the number of submissions. Accompanying time series track valence, post-response change labels, and unit production, and full distributions show whether a shift in an average conceals a change in shape.

Here, attention means the share of extracted evaluations assigned to a category. It differs from the share of papers with that feature, from the share of reviewers who care about it, and from the weight the feature carries in a decision. A review containing more extracted evaluations contributes more to the unit-level distribution. Keeping the forum extraction track and category definitions fixed makes the yearly distributions comparable under one instrument, although the submitted papers and the review process are not held fixed. Attention also differs from the conditional grammar examined earlier: a standard can become more common simply because the objects that tend to invoke it become more common, even if the standards attached to each object change little.

Figure 64 tracks the objects and reasoning standards represented in written evaluations across years. Both panels use year horizontally and the percentage of that year's forum-track units vertically. Lines represent the twelve objects on the left and the twelve standards on the right; the same category centroids, and hence the same category definitions, are applied throughout the forum-level track. Brown highlights selected rising shares, violet selected falling shares, and pale gray the other categories. Endpoint annotations give percentage-point differences between the first and last years, not relative growth or yearly unit counts. From 2018 to 2026, clarity's attention share falls $4.9\,\mathrm{pp}$ and theory's $4.1\,\mathrm{pp}$, while compute cost rises $3.2\,\mathrm{pp}$ and statistical rigor $3.0\,\mathrm{pp}$; baselines and reproducibility also increase. In relative terms, clarity declines by approximately 37% and compute cost grows by 67%. Reviewer-clustered bootstrap checks, 1,000 draws with 2,721 versus 73,427 reviewer clusters at the endpoints, place the four named changes at 8–16 standard errors. Read alongside the earlier within-object comparison, these shifts show attention being reallocated while the organization of evaluative standards stays comparatively stable.

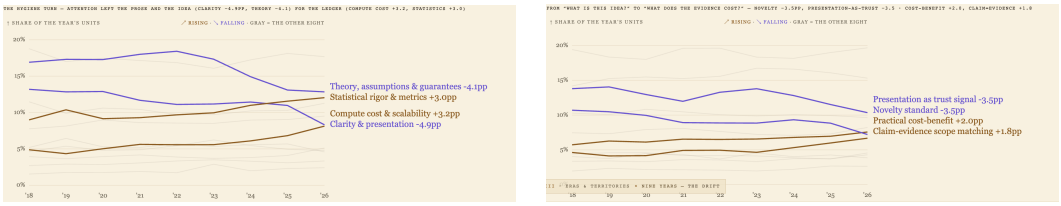


Figure 64: Left: Historical object shares. **Right:** Historical reasoning-standard shares. These compare one extraction protocol through time, rather than pool the two tracks.

Novelty and presentation-as-trust standards each decline about 3.5 pp, while cost-benefit and claim–evidence scope matching increase. The forum-level negative share stays near 77%, which differs from the separate review-level track's 71.8%. Figure 65 summarizes yearly changes in recorded updating, post-response evaluation, and extracted-unit production. It has three small time series, each with year horizontally and its own vertical range. The left panel is the percentage of post-author-response units labeled weakened or reversed; the middle panel is the percentage of all units placed after the author response; the right panel is total units divided by the number of forum–reviewer-key profiles. The first two are percentages and the third is a mean count. The vertical ranges do not share a zero baseline, so slopes and heights cannot be compared across panels as if they were one quantity. The saved series in the left panel declines from 7.7% in 2018 to 3.4% in 2026. Its denominator includes every unit labeled post-author-response, including those without an observed change, so in this representation later written judgments soften less over time. Category-assignment similarity remains near 0.81 across years, which helps diagnose gross changes in category fit but does not establish that extraction error stayed constant. Submission composition, review forms, and reviewer behavior remain entangled.



Figure 65: Historical shares of softened post-response judgments and post-response units, alongside units per reviewer profile. Changing institutional context and review forms qualify the temporal comparison.

Figure 66 shows the full yearly distributions behind the rating, review-length, unit-count, and argument-form summaries. Rows are years and the four columns are rating, review word count, units per reviewer profile, and argument form. Within each small panel, horizontal position identifies the rating value, count bin, or form, and bar height gives its percentage within that year's relevant population. Each small panel sums to 100%; ratings and words describe reviews, units describe profiles, and forms describe classified units. Vertical scales are shared down a column but differ between columns, and rating values that a year's form did not offer are left absent. Review length shifts after its 2022 peak, and on-balance weighing gradually replaces some precedent-citing. Because ratings are plotted on each year's observed levels, raw positions are not directly compared across changed forms. These displays reorganize the same record that the earlier summaries compress.

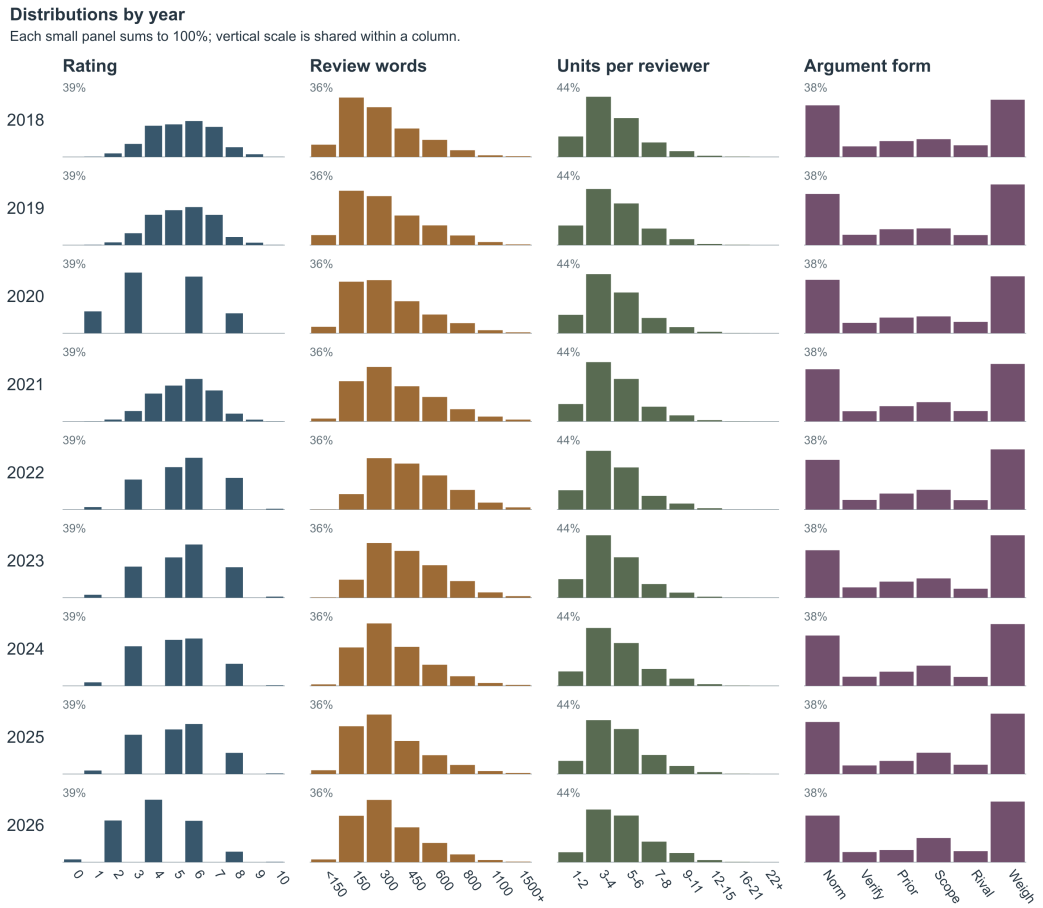


Figure 66: Distributional context for all nine years: ratings, review length, unit counts, and argument forms. This static view reorganizes the source distributions so each year is visible simultaneously. Vertical scale is shared within each column; rating positions retain each year's available levels, and the changing forms must qualify comparisons.

9.2 Coarse novelty rules remain comparatively stable

A change in how frequently novelty is criticized need not mean that the criterion of novelty itself changed. To examine the criterion's wording, we fit clause centers to 2026 normative novelty sentences and measure each earlier sentence's distance to its nearest center. A half-of-2026 fit evaluated on the held-out half supplies a within-year reference distance. We also assign sentences to those clauses and compare their yearly shares, separating a shift in which clause is used from a shift in fit to the clause vocabulary.

Normative sentences state what would count as novelty, rather than only asserting that a particular paper lacks it. The novelty "code" is a compact set of clause categories derived from these sentences; it records how reviewers phrase the requirement and cannot decide whether a paper is new. Sentence embeddings place similar formulations near one another, and each clause center summarizes one group of formulations. The larger a sentence's distance to its nearest center, the less well this reference vocabulary represents it. Even fresh sentences from the reference year need not coincide with a center, which is why the held-out comparison matters: earlier years should be judged against that ordinary within-year mismatch rather than against a distance of zero.

Figure 67 compares the wording and clause composition of novelty standards across years against the 2026 reference. It has a multi-part distance display on the left and yearly clause composition on the right. The left upper plot uses year horizontally and mean embedding distance to the nearest

2026 clause center vertically, with the held-out calibration as a dashed line. The middle plot converts distance minus calibration into sentence-bootstrap standard-error units, with a shaded ± 2 band; values beyond the displayed range are clipped. The lower bars place two different quantities side by side: selected objects' attention changes relative to their early-year shares, and the early years' novelty-distance excess over calibration. Those bars therefore do not share a common substantive measurement. On the right, each year is a row, six clause colors partition its normative sentences into percentages summing to 100%, and the printed n is that year's sentence count. A code fitted to 2026 normative novelty sentences is applied to all 13,138 such sentences across years. Held-out half-of-2026 calibration has mean nearest-center embedding distance 0.152; 2018 is approximately 0.159, and every year is within 0.008 of calibration. Early deviations around 0.006 can be two to three sentence-bootstrap standard errors, although clustering sentences within reviews makes those standard errors optimistic. At this representation the differences are small. Some are statistically detectable, so the finding concerns effect size rather than detectability.



Figure 67: Left: Distance to a 2026-fitted novelty code, with split-half calibration. **Right:** Yearly mix of the code's clauses. Changes in use of a clause need not imply changes in its meaning.

All six clauses occur in every year. Differentiation declines from approximately 50% to 38%, while combination absorbs much of that shift; transfer is about 5–7%. The transfer cluster contains both transfer and differentiation formulations, so its share is less secure. A reverse fit on only 417 early sentences provides weak corroboration. Sentence embeddings can miss semantic changes that preserve register, and coarse clusters can conceal changes within a clause. Under this instrument, the defensible reading is that the novelty criterion stayed comparatively stable while attention shifted. A claim that the definition of novelty never changed would need a finer instrument.

9.3 Field-specific criticism profiles

Venue-wide frequencies can obscure differences in the kinds of evaluation applicable to different research areas. We split 2026 submissions by their recorded primary area and, within each area, count the proportion of papers with at least one negative unit on each object. Subtracting the venue-wide proportions gives an area's criticism profile. Comparing those vectors and projecting them with multidimensional scaling provides a descriptive map of similarities among areas; it does not adjust for differences in the papers submitted to them.

This comparison changes the denominator from evaluations to papers. Repeated criticisms of an object on the same paper count as the same presence indicator, so the profile describes the reach of each criticism across submissions rather than the amount of text devoted to it. A positive deviation

means that more papers in that area receive the criticism than in the venue as a whole. The map compresses the resulting multivariate profiles into a display in which nearby areas have similar combinations of deviations. Placement on the map helps in seeing those combinations; a map direction has no substantive meaning of its own, and position does not rank the areas.

Figure 68 maps similarities and differences in criticism profiles across research areas. Each primary research area is drawn as an island. The two map coordinates come from multidimensional scaling of the object-criticism profiles and have no geographic or chronological units; nearby islands have similar profiles. Island size reflects paper volume, and the radial coastline pattern encodes object-specific departures from the venue's criticism rates in the directions marked by the compass. Labels add paper counts and acceptance rates. The coastline encodes a multivariate profile; it is neither a geographic boundary nor an uncertainty region. All 19,474 reviewed 2026 papers carry a primary-area field, and each area's profile is computed from them as described above. Learning theory has approximately $+25$ pp theory criticism and -31 pp baseline criticism relative to the venue. Datasets and benchmarks receives particularly little theory criticism. A venue-wide profile therefore combines distinct local evaluative environments, and no single criticism pattern is equally typical of every research area. The profiles are conditional frequencies. They measure neither field merit nor reviewer fairness, and they do not fix the fate of a submission, since areas differ in what is submitted and in which evaluations apply.

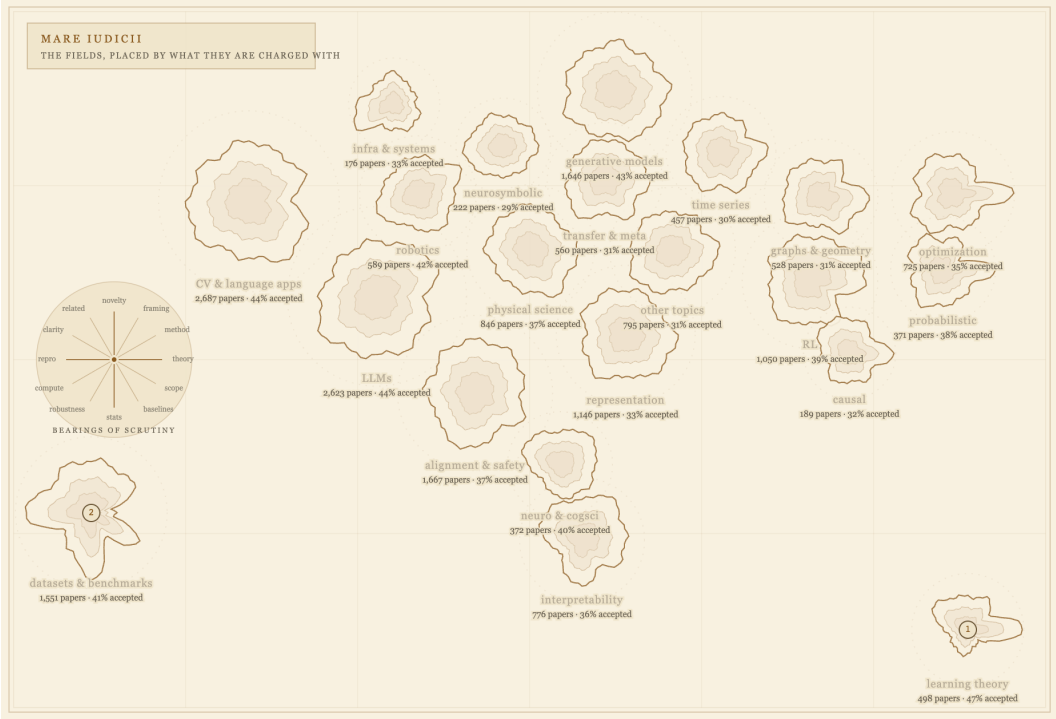


Figure 68: All twenty-one primary areas positioned by criticism profiles. Geographic shape and distance encode the defined summary, not a latent hierarchy of research quality.

9.4 A frozen vocabulary tracer and its changing sensitivity

The arrival of widely available language models raises a separate measurement problem: whether review language changes in a way detectable by a fixed, transparent tracer. To avoid identifying the phenomenon through the extraction models' own paraphrases, the vocabulary analysis uses raw review text. It counts documents containing a frozen style-word marker and compares yearly frequencies with a pre-LLM trend. The placebo year and ordinary-vocabulary control ask whether the apparent departure could instead be an ordinary continuation of earlier lexical change.

A review is marker-positive when its free text contains a word from the fixed list; repeated uses within one review do not make it several positive documents. The reference trend estimates how

frequent the marker would have been had its earlier linear trajectory continued. That projection is a lexical counterfactual; no observed population of reviews written without assistance stands behind it. The placebo applies the same comparison to a designated earlier year, and the negative-control words check whether an apparent surge also appears in ordinary review vocabulary. Together these comparisons let the tracer be interpreted without assuming that a positive document was AI-written.

The LLM-era analysis starts from raw review free text in all 199,031 official reviews. It removes enumerated form fields and freezes thirteen style words before computation: *commendable, meticulous, meticulously, intricate, pivotal, versatile, delve, delves, delving, underscores, underscoring, showcases, showcasing*. The marker list draws on Liang et al. [27] and Kobak et al. [21]. For each year, document frequency is compared with extrapolation from its 2018–2022 linear trend. The design includes a 2023 placebo and a ten-word ordinary-review-vocabulary negative control.

Figure 69 tracks the prevalence of the frozen style-word marker against its pre-period trend. The horizontal axis is review year and the vertical axis is the percentage of raw official reviews containing at least one word from the frozen marker list. The solid line is observed document frequency; the dashed line extrapolates its 2018–2022 trend, and the shaded gap highlights excess after the rise. The 2023 point supplies the placebo-year comparison. The ordinary-vocabulary negative control is reported in the text rather than plotted as another line in this panel. The marker frequency is 1.2% in 2023 against 1.5% expected, then 7.8% in 2024 against approximately 1.6%. Excess is about $+6.1$ pp in 2024, $+4.4$ pp in 2025, and $+1.1$ pp in 2026; the raw 2026 share is 3.1%. The negative control does not spike and moves mildly below trend. Length stratification does not explain the marker's later decline. Together, the trend and controls identify a cycle-specific lexical departure followed by a fading signal from this fixed vocabulary.

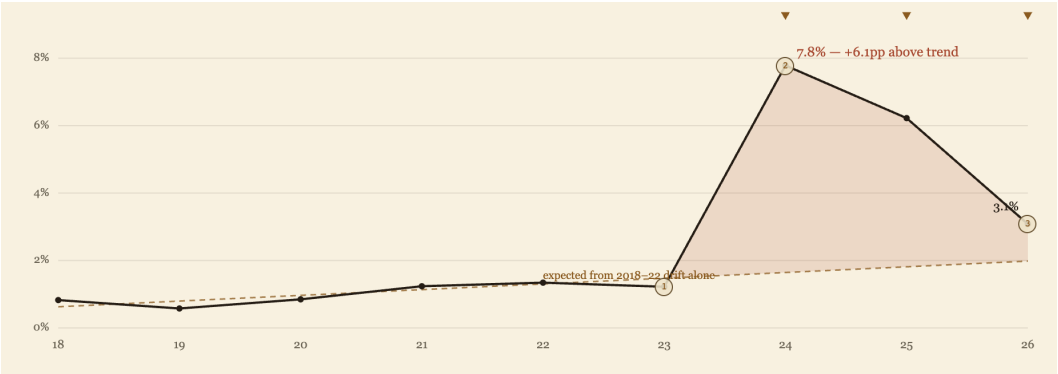


Figure 69: Review-level frequency of the frozen style-word marker and its pre-period extrapolation, including the placebo-year point. The negative control is reported in the text. Marker prevalence is not an estimate of the fraction of AI-written reviews.

Figure 70 shows which individual marker words depart from their expected frequencies after the pre-period. It has individual marker words in rows and 2023–2026 in columns. Each printed cell is observed review-level document frequency divided by that word's extrapolated expected frequency for the year. Redder cells exceed a ratio of one, bluer cells fall below it, and neutral cells are near one. The unit is a frequency ratio, not percentage points or number of word occurrences, so a small expected frequency can produce a very large cell value. In 2024, *meticulously* is 14.4 times its projected baseline, *meticulous* 12.6, *delves* 11.8, and *pivotal* 11.5. By 2026, *delves* is at 0.35 and *delve* 0.72 of baseline, while *underscores*, *meticulous*, and *commendable* remain elevated. The approximately 253-fold value for *underscoring* starts from a near-zero baseline and should not dominate the interpretation. Topic words such as model names undergo much larger adoption shifts but are deliberately excluded from the style tracer.

The marker's sensitivity can change as models, editing practices, disclosure rules, and ordinary human vocabulary change. It cannot identify individual AI use, and without assumptions about false positives and missed assistance it is neither a calibrated prevalence estimator nor a formal lower bound. The external sentence-modification estimate for ICLR 2024 [27] and later review-involvement estimates [33] target different quantities with different instruments, so comparing their percentages directly with marker-positive documents would mix denominators. In particular, a fading fixed lexicon does not imply declining assistance.



Figure 70: Each marker word against its own extrapolated baseline. Extreme ratios at very low expected frequency require particular caution.

9.5 Wording converges without a resolved structural narrowing

The vocabulary marker tracks one feature of review language. To test whether reviews become more alike, we compare raw-text similarity among co-reviews of the same paper with similarity among random cross-paper pairs, holding the vectorizer vocabulary and review form constant across the comparison years. Splitting the text into summaries and evaluative fields checks whether any convergence is concentrated in describing a common paper. A separate analysis compares sets of criticized objects against a null that preserves set size and yearly object prevalence, because longer lists overlap more even without substantive convergence.

For the wording comparison, TF–IDF represents each review by word weights that reflect its vocabulary while reducing the influence of words common across the corpus. Cosine similarity compares the direction of two such vectors; a higher value indicates greater similarity under this lexical representation. Co-reviews share a paper's subject matter, so they are expected to resemble one another even without a change in reviewing practice. Cross-paper pairs supply a comparison for language shared more generally across the venue. Separating summaries from evaluative fields further distinguishes convergence in describing the submission from convergence in how it is assessed.

Figure 71 compares historical changes in review-text similarity with changes in the overlap of criticized objects. Both panels use year horizontally. On the left, the vertical axis is median TF–IDF cosine similarity between raw review texts; the dark line is same-paper pairs and the gray line random cross-paper pairs. Highlighted late-year points identify the constant-form comparison. On the right, the vertical axis is mean Jaccard overlap of sets of negatively evaluated objects: dark is observed co-reviews, brown dashed is the set-size and annual-mixture null, and gray dashed is the random reference. All of these similarity measures are dimensionless, but cosine and Jaccard are defined differently and cannot be read against each other. Within the constant-form 2024–2026 window, a shared TF–IDF vocabulary gives median same-paper review cosine similarities 0.277, 0.299, and 0.317, an approximately 14% increase. Random cross-paper pairs stay near 0.078 to 0.073. Summary sections rise about 20%, while weaknesses-and-questions prose rises about 6% and levels off after 2025. The measured convergence is therefore concentrated more in accounts of the shared paper than in prose evaluating its weaknesses. Describing the same paper is an important shared cause of summary similarity. The fixed-vocabulary, constant-form comparison removes two simple artifacts, but does not identify AI as the cause of the residual trend.

For the object-set comparison, each review is represented by the distinct objects on which it contains negative units. Co-review pairs are scored by Jaccard overlap, the intersection size divided by the union size. The same calculation on simulated sets of the observed sizes supplies the reference distribution. Object entropies and similarities of embedded reasoning provide additional checks on concentration at the corpus, individual-profile, and text-representation levels; they measure different aspects of narrowing and are not substitutes for one another.

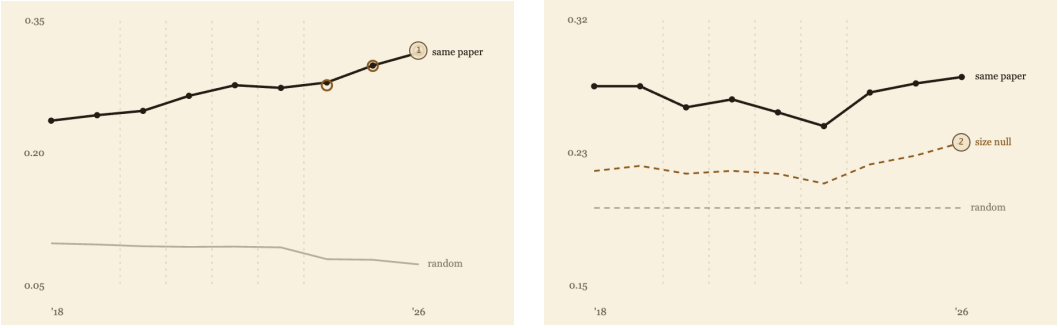


Figure 71: Left: Within-paper wording similarity and random-pair comparison. **Right:** Overlap in criticized object sets, with a null retaining each set size and the year's category mix.

The set-size control is needed because adding criticized objects increases the opportunities for two reviews to intersect. Drawing from the year's category mix also allows for objects that are common throughout that year. An observed overlap above this reference therefore concerns shared objects beyond these two sources of overlap. Entropy addresses a different question: whether attention is concentrated on a small part of the category set or distributed more evenly. Its value depends on how many observations are available in a profile, motivating the count-matched comparison for individual reviewers. Reasoning embeddings add a text-level comparison, whose interpretation still depends on the model-generated restatement being embedded.

Object-set Jaccard overlap also rises, 0.252 to 0.284, but mean set size grows from 3.4 to 3.8. A null preserving those sizes and drawing contents from each year's mix absorbs the rise; observed-minus-null excess is $+0.037$ to $+0.054$ across years, with no trend and the maximum in 2018. The initial content-convergence interpretation was therefore withdrawn.

Figure 72 compares historical changes in evaluation diversity and the similarity of co-reviewers' reasoning. It gives three yearly plots with different vertical quantities. Left is normalized entropy of the corpus object and standard mixes, on a 0–1 scale. Middle is mean normalized object entropy within reviewer profiles, compared with the count-matched null. Right is the mean reasoning-embedding similarity of co-reviewers minus that of cross-paper pairs, in cosine-similarity units. Lines identify these quantities or their references; none of the panels shows counts, and heights are not comparable across panels. Three related summaries point, within limits, to an absence of narrowing. Corpus object entropy rises from 0.916 to 0.940. Mean within-reviewer entropy rises from 0.742 to 0.766, but an n-matched null absorbs most of the increase; only about 0.008 remains, without trend. Because the null draws from the same year's mix, that rise is not independent evidence of broadening. The co-review minus cross-paper similarity gap in distilled reasoning remains approximately 0.012, and overall embedding dispersion is 0.231 to 0.229. Twelve coarse categories and a common model voice can miss within-category or stylistic convergence.

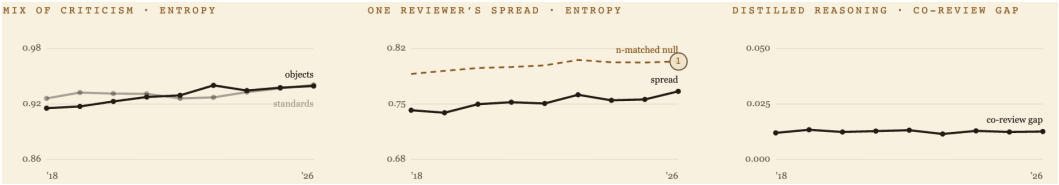


Figure 72: Corpus mix, within-reviewer mix with an n-matched null, and reasoning-embedding similarity. These summaries do not establish unchanged substance at every semantic resolution.

Type–token ratio is the number of distinct words divided by the number of words examined. It describes variety within a review, whereas the earlier cosine comparison describes resemblance between reviews. A shared repertoire can therefore become more varied within each document while also becoming more common across documents. Neither a rising type–token ratio nor rising co-review similarity, taken alone, settles whether the range of substantive judgments has narrowed.

Per-review type–token ratio over the first 200 words reaches 0.638 in 2026, versus approximately 0.578–0.598 earlier. Thus greater similarity between reviews can coexist with greater lexical variety within reviews. The lower lexical diversity reported for fully generated reviews in benchmark work [26] concerns a different population and text window. It does not predict that historical human-review records must show the same change. Studies of fully generated AI reviewers' overlapping criticisms [1, 20] motivate a structural test, but cannot establish that prose-level assistance causes our wording-only pattern. A neighboring review-text diagnostic study locates a 2022–2023 shift [6], whereas our frozen-vocabulary placebo places its own detectable arrival one review cycle later. The two studies measure different things, so their dates do not contradict each other as estimates of one parameter.

9.6 Associates of marker-positive reviews

Finally, we ask what distinguishes the group selected by the vocabulary marker. We compare each marked review's rating with unmarked co-reviews of the same paper, and compare textual and interaction features with adjustment by length decile. The longest-review contrast checks a simple length explanation of the score difference. Permuting marker labels within paper supplies a baseline for score-quality strata that could otherwise generate an apparent gradient through conditioning and mean reversion. These comparisons characterize a marker-defined group rather than an assigned AI-use condition.

Comparing reviews of the same paper holds the submission constant, but still compares different reviewers and their judgments. Grouping reviews into length deciles compares text features among reviews of broadly similar length, since longer documents offer more opportunities to contain both a marker and a reference or criticism. The permutation check addresses a further issue in the score comparison: a review contrasted with unusually low- or high-scoring peers can appear unusually favorable or unfavorable partly because those peer scores define the comparison. Shuffling the marker labels within each paper retains that comparison structure while removing the observed association between the label and a particular review. The excess over this reference is an adjusted association; it does not estimate an effect of language-model use.

Figure 73 compares marked and unmarked co-reviews in rating, citation, reply, and confidence outcomes. It is a grid of outcome rows and three year columns, 2024–2026. Each miniature horizontal bar extends from its own zero: rightward red is a positive contrast and leftward blue a negative one. The upper rows give within-paper rating-point differences for marked versus unmarked co-reviews and, hatched, a separate longest-review benchmark. Lower rows give length-adjusted differences in the share citing "et al." in percentage points, number of rebuttal replies, and confidence points on the five-level scale. Scaling is shared within a row but differs between outcomes, so bar lengths across rows do not measure comparable effect sizes. Within papers, marker-positive reviews score approximately 0.14 points above unmarked co-reviews in 2024, 0.17 in 2025, and 0.11 in 2026. They are 58–123 words longer, yet a paper's longest review tends to score 0.17–0.35 points *lower* than its peers. This weakens a length-only explanation but does not rule out selection. Text-profile comparisons use length deciles, and the quality-stratified score contrasts permute marker labels within paper 200 times, for the mean-reversion reason given above. The adjusted excess is approximately 0.07–0.28 points over the permutation baseline across strata.

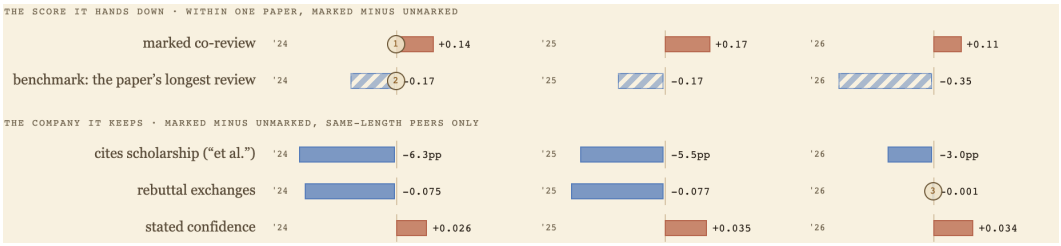


Figure 73: Within-paper score differences and length-adjusted associates of the marker. Inferences concern a changing vocabulary-defined group, not identified AI authorship.

The marker-positive group uses extractable scholarship references less often (the 2024 "et al." gap is about -6.3 pp) and has fewer rebuttal exchanges (approximately -0.08); both differences attenuate,

with the rebuttal gap near -0.0009 by 2026. Confidence is mildly higher, about $+0.03$ after length adjustment, differing from the lower-confidence profile discussed in Liang et al. [27]. The within-paper score direction resembles Latona et al. [24], using a transparent word list instead of detector labels. Using a held-out human rating as a quality proxy, Sharma et al. [33] finds that apparent interactions can reflect greater leniency toward lower-rated submissions. That finding underlines the need to separate selection into a group from an AI-use effect, and agreement across these studies does not turn our marker into an AI-use treatment. Changing group composition is itself a plausible reason that the correlates fade.

The marker set, counterfactual, placebo, and decision rules precede computation. Fixed-vocabulary similarity, length stratification, permutation controls, and some quality strata are post-run additions recorded as such. Fine-grained aspect-coverage work [43] and broader prevalence estimation [11] provide neighboring approaches; this report makes no exhaustive priority claim for co-review similarity measurement.

10 Null results, inspection material, and reproducibility

The previous sections report observed patterns together with the comparisons that weaken their interpretation. This section gathers those checks and adds the material needed to inspect how the measurements were produced. A result can be numerically reproducible and still fail to establish the proposed explanation: the same pattern may appear under a null model, or the representation may omit the information that would distinguish competing readings. Conversely, an illustrative source example can clarify a category without establishing its accuracy across the corpus. The section therefore treats four things separately: the status of an interpretation, the inspection of individual records, the development history, and the reproducibility of the computations.

10.1 Results that did not support their initial interpretation

The proposed reviewer archetypes would have described different emphases in evaluation, such as reliance on different reasoning standards. The actual observations, however, are profiles of one reviewer on one paper: a vector recording how that profile's extracted units are distributed across the standard categories. The profiles describe evaluations on particular papers; they do not track a person's preferences across papers. A short profile, moreover, can be dominated by a category simply because few units were observed. Before reading groups of such profiles as reviewer types, we need to know whether comparable groups appear when every profile is generated from the same underlying mix.

The analysis proceeds in three steps. We cluster the real reviewer–paper standard-frequency vectors, generate synthetic vectors from one shared standard mix at each observed profile's unit count, and apply the same clustering to the synthetic set. Comparing explained variance and standard-wise dispersion between the two then shows whether the visual types require genuine heterogeneity between profiles. Matching each profile's unit count is essential, since sparse counts alone produce extreme proportions.

K-means groups profiles around centers chosen to reduce within-group dispersion, and it will produce such groups even when the data contain no genuine types. Visually distinct centers are therefore insufficient evidence. The multinomial null makes that alternative explicit by sampling categories from one common mix with the unit count of each real profile. Explained variance measures how much variation the clustering captures, and overdispersion compares the observed spread of a standard's frequency with its spread under the common-mix baseline. What counts as evidence is the excess structure in the real profiles over the null, rather than the mere existence of clusters.

Figure 74 shows five cluster profiles from real reviewer–paper records in the upper row and five from the shared-mixture synthetic null below. Each radial spoke identifies one of the twelve standards, and its length is the square-root-compressed lift of that standard in the cluster relative to the corpus frequency; the dashed ring marks a lift of one. Labels give the dominant standard and the percentage of profiles assigned to the cluster. Neither the spokes nor the cluster percentages count people with a validated reviewer type. The clearest rejected interpretation is a set of five reviewer archetypes. K-means applied to 98,513 reviewer–paper reasoning-standard profiles with at least five units produces apparently strong specializations, although the median profile has only six units. Running the same

clustering on multinomial profiles generated from one common field mix, matched to each profile's unit count, produces similar apparent specialists. Explained variance is 35.8% in the real profiles and 31.0% in the null. Median per-standard overdispersion is 1.15, at most 1.58. A small real signal remains and grows with profile length, but it does not support the initial strong typology. The archetype analysis was withdrawn as a principal finding on 21 August 2026.

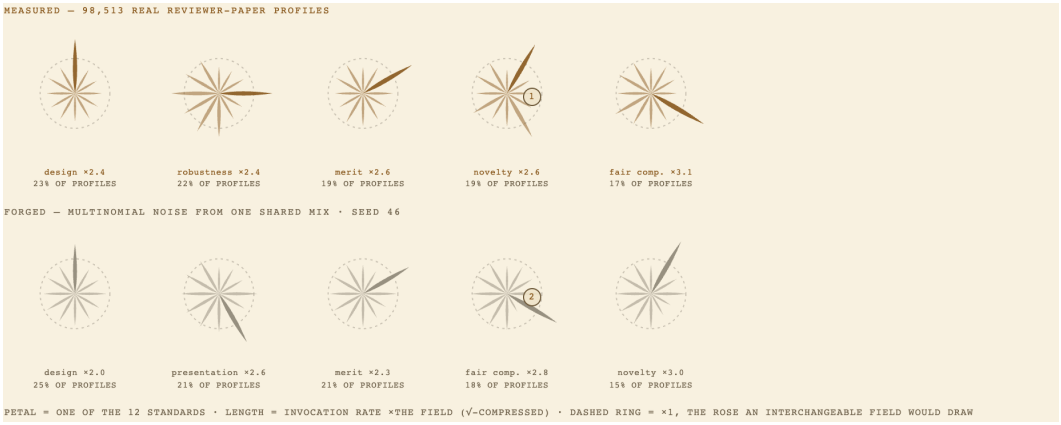


Figure 74: Real and synthetic reviewer-profile clusters. Synthetic profiles contain no true reviewer types, demonstrating that visually strong specializations can arise from limited unit counts and clustering.

A second retained description asks whether breadth of criticism accompanies rejection. Starting from the same twelve paper-level criticism indicators used in Figure 35, we count the distinct objects criticized on each paper, group decided papers by that count, and calculate each group's acceptance rate. The exercise re-summarizes the existing decision join. It adds no measurement of paper quality and cannot separate breadth from the severity of the underlying problems.

Figure 75 shows how acceptance varies with the breadth of criticism recorded for a paper. The horizontal categories are the number of distinct objects receiving criticism on a paper, with nine or more combined in the last bin. Vertical position is the percentage accepted among papers in that breadth group. Larger circles indicate groups with more papers, and the separate gray bars below summarize paper counts; their display scale is distinct from the acceptance-percentage axis. The connected points describe grouped proportions, not what happens when an objection is added to the same paper. Acceptance declines with breadth of criticism: approximately 69% at two objects, 57% at three, and 27% at nine or more, with an approximately five-point decline per additional object across much of the range. The modal decided paper is criticized on seven objects. No decided paper has zero criticized objects, and the two with one are omitted from the display. The association is partly expected: papers with more underlying problems may also attract criticism of more objects. We retain it as description; breadth is neither an intervention nor a separately established reviewer mechanism.

Other weak results are retained where they arise: form transitions remain close to their within-review null; object adjacency loses most of its apparent strength under that null; question-field placement does not meaningfully separate recorded post-response fates; most detected authorial moves lose their association after length stratification; and the highest- versus lowest-rater echo contrast is too small and tie-dominated to support a strong chair-borrowing rule. A preliminary "advocate reviewers soften more" interpretation was discarded because the advocate category was defined by praise-oriented reasoning. Because the category is defined by praise-oriented reasoning, the pattern provides no independent evidence for a reviewer type.

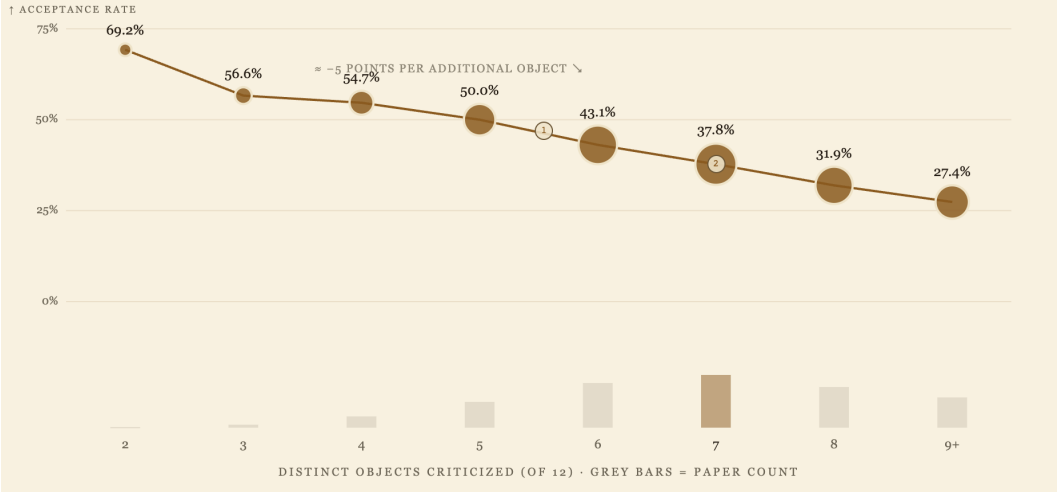


Figure 75: Acceptance associated with the number of distinct objects criticized. This is the same 13,704-paper join used in the object-specific acceptance comparison, summarized by breadth.

10.2 Lexical change in the extracted criticism

Stable category labels could conceal a change in the language assigned to them. To inspect that possibility, the lexical comparison holds the object fixed, contrasts extracted criticism from 2018–2020 and 2024–2026, and compares the early-to-late centroid displacement with typical year-to-year movement. Characteristic terms describe the direction of that displacement. Because the input is extracted text, the check concerns the representation's language and cannot separate reviewer wording from memo restatement.

Holding the object fixed prevents a change in the mixture of criticized objects from being mistaken for a change in the language used about an object. A centroid summarizes the represented language of the group; its displacement describes how far that summary moves between periods. The year-to-year comparison supplies a scale for judging whether the longer-period movement is large relative to ordinary annual variation. Characteristic terms help interpret what distinguishes the periods, but neither they nor the displacement identify why the language changed. Unlike the raw-review vocabulary tracer in Section 9, this comparison includes a model's restatement in the text being compared.

Figure 76 has three text columns for novelty, empirical-scope, and clarity criticism. Inside each column, period labels separate characteristic terms from 2018–2020 and 2024–2026 extracted language. The display is a vocabulary comparison: it has no numerical axes, and neither word position nor column width encodes token frequency, effect size, or a count of reviewers. Within extracted novelty, scope, and clarity criticisms, the comparison of 2018–2020 with 2024–2026 gives a centroid displacement of approximately 7–14 times typical year-to-year variation. Early discriminative vocabulary includes verdict terms such as "fails" and "weak"; later vocabulary includes "limited," "concerns," and "generalizability." These are the pipeline's paraphrases, and the introduction of structured fields is a plausible contributor. The analysis therefore documents a lexical shift under changed writing and extraction conditions; it is not independent evidence of a change in evaluative substance or of AI use.

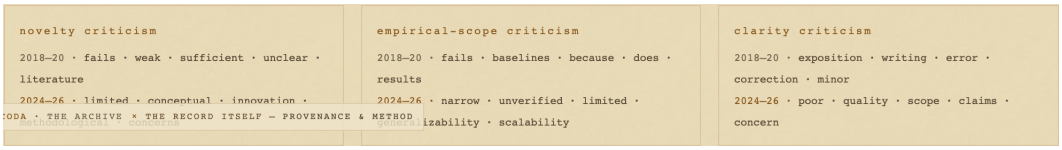


Figure 76: Characteristic early- and late-era criticism vocabulary. The comparison is on extracted language, and its interpretation is qualified by form changes and common model phrasing.

10.3 Inspection against the original record

Aggregate tables cannot show whether a particular label plausibly represents its source sentence. The specimen collection addresses that inspection task by selecting reviews that cover the taxonomy and linking their unit fields back to the public record. It deliberately favors interpretable, confidently assigned examples; accordingly, it supports examination of the transformation from source to label, rather than an estimate of population-wide accuracy.

The Atlas supplies 232 specimen reviews chosen for taxonomy coverage rather than by random sampling: for each object, up to twenty reviews with an assignment similarity of at least 0.75 are selected and then deduplicated. Their summaries and unit fields are model-generated representations linked to the public review, and they serve the inspection task just described rather than error-rate estimation. Fourteen deliberation cases and five reply trees play the same illustrative role. The report preserves the specimen-selection analysis and worked examples, while the full navigable case records remain at the linked source.

Table 8 puts object codes in rows and standard codes in columns. Each interior cell is an unnormalized count of 2026 review-track evaluation units with both labels; right and bottom margins give row and column totals, and the corner gives the total unit count. Dividing by the row total produces the conditional standard share used in Figure 10; dividing by the grand total gives a corpus share. Neither normalization is the negative-share calculation in Figure 9, which also needs valence counts.

Table 8: Full object–standard unit counts in the 2026 review-level track. The row and column codes are defined in the taxonomy tables in the methods section.

	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	Total
O1	10283	10485	7396	3690	6182	19729	9411	3149	3220	2324	2217	11376	89462
O2	3757	1582	10188	1992	764	5998	1882	655	443	402	626	4448	32737
O3	9779	4460	2390	2501	1745	13518	6704	1849	1092	1320	3236	8562	57156
O4	10113	4010	4178	2231	4179	14135	4867	3042	1450	1251	1601	9070	60127
O5	2600	1397	3248	1319	846	5824	2051	10883	466	521	460	3912	33527
O6	3629	610	760	823	383	3241	623	213	710	1070	15688	5246	32996
O7	7656	202	349	186	164	1562	140	83	110	36	149	803	11440
O8	5157	195	423	137	132	1010	84	50	100	187	1823	1008	10306
O9	3832	2893	4662	3567	1705	5701	3914	1227	5071	1081	2231	5231	41115
O10	1594	840	1666	1136	944	3292	6922	1191	475	668	431	1811	20970
O11	695	346	345	645	387	837	374	375	199	2665	559	1169	8596
O12	3186	506	469	377	464	1969	337	204	280	113	769	3480	12154
Total	62281	27526	36074	18604	17895	76816	37309	22921	13616	11638	29790	56116	410586

Inspecting a specimen means following a proposed interpretation back to what the reviewer actually wrote: which passage supports the assigned object, whether the claimed standard is explicit or inferred, and whether the extracted judgment keeps the scope of the original. A high assignment similarity means only that the extracted representation lies close to a category center, which does not certify that the source was read correctly. Because the specimens were selected for coverage and similarity, inspecting them successfully cannot replace a representative accuracy audit, and an ambiguous case should remain open to an alternative reading.

10.4 Development sequence and inferential status

Report provenance. The analysis is consolidated from the Atlas main page and Method page [40, 41], fixed to repository commit `7bf84e1`. About, Resources, API documentation, and the dataset card additionally inform the account of the instrument and its public artifacts [36, 37, 39, 42]; these were checked on 8 November 2026. Preparing the report neither extended the collection window nor re-executed the original analyses. The development ledger below separates the source dates and the analyses specified before computation from later exploratory checks.

The explanation attached to each analysis describes the question it addresses and why its comparisons are informative. Such explanations do not claim that every question or control was specified

before the results were available; the record of when those decisions were made is summarized separately here.

The report is ordered for exposition. The original analyses were developed iteratively, and their later arrangement into acts was not a preregistered sequence. The Method page distinguishes foundational work, questions stated before computation, partly specified work, exploratory descriptions, and analyses with no reliable motivational record. In its 34-row ledger, which also includes navigation and scene-setting material, the counts are 17 stated-first, four partly stated, three foundational, one exploratory, and nine not recorded. Table 9 groups the source analyses by development status in rows. Its columns give that status, the original Atlas plate identifiers in Roman numerals, and the inferential interpretation. The identifiers refer to source analyses rather than to this report's renumbered figures, and the entries classify documentary timing rather than ranking evidential quality.

These statuses describe the strength of the documentary record, not whether an analysis is worth reporting. A question recorded before a run can still leave its estimator or comparison unspecified, and an informative control added after a pattern was seen remains a later check. Where no reliable record survives, the motivation given in the text cannot establish when the idea arose. Keeping the statuses visible lets the reader weigh each analysis without mistaking a coherent presentation for a prospectively fixed design.

Table 9: Development provenance of the thirty source analyses. The table records when each question was stated; it does not describe a preregistered or confirmatory testing regime.

Status	Source analyses	Interpretation
Foundational	I, II, III	Built with the taxonomy and representation.
Question stated first	IV, VI–IX, XIII, XVIII, XXII–XXVII, XXIX, XXX	Question recorded before computation; subsequent design choices and additions remain qualified in the text.
Partly specified	V, XIV, XXI	An initial direction is recorded; later components evolved.
Exploratory	XXVIII	Descriptive field comparison without a prior prediction.
Not recorded	X–XII, XV–XVII, XIX, XX	No reliable record of the specific motivating question. Later additions can have a stronger record.

Production followed pilots at each stage. The record includes 395 initial-blind pilot jobs, 327 trajectory jobs, 100 synthesis jobs, 100 direct jobs, and a 100-forum protocol comparison. A richer episode schema was piloted on 1,000 stratified reviews and 63 deeper cases. Full-scale attempts were stopped after shallow or templated extraction and projected runtime made the approach unsuitable; 876 completed outputs with 186 failures were retained as comparison artifacts, not merged into final units. The compact Qwen protocol was frozen only after batch-endpoint failures, realtime tests, rejection of two-review packing for lost coverage, and a 206-request load test with 203 successes. Direct pilots also progressed through failed and stalled versions before a 79-of-90 run froze the production method.

Three consequential corrections remain part of the record. First, the original claim that a verdict was readable by unit six compared it with an incorrectly stated median length of eleven. The actual median is five; the corrected long-review analysis saturates around eight to nine units. Second, object-set convergence in the LLM-era analysis was reinterpreted after a size-matched null explained its rise. Third, the apparent increase in individual attention entropy was retracted as evidence of broadening after an n-matched null absorbed most of it. The latter two early readings appeared on the live page for less than a day. Section 9 reports the corrected analyses, not the abandoned interpretations.

10.5 Computing environment, artifacts, and production cost

Reproduction covers two different tasks. One is to recompute a table or figure from the saved units and analysis inputs. The other is to regenerate those units from the original reviews by repeating the language-model stages. Saved outputs and scripts support the first task even where model behavior prevents exact repetition of the second. The software record and cost ledger below document how

the existing measurement was produced; neither independently validates that the units faithfully represent human judgment.

The original pipeline uses Python at least 3.12 managed with `uv`, OpenReview clients, OpenAI-compatible API clients, JSON-schema validation, local sentence embeddings, UMAP, and HDBSCAN. The documented synthetic-fixture suite has 69 tests; passing it checks software behavior and does not establish measurement validity on production text. Analysis directories preserve manifests, request ledgers, rendered prompts, provider responses, local validations, and outputs. Collection and normalization maintain separate immutable and processed databases. Unit extraction selects only strict-complete records.

The artifacts connect successive transformations: the request ledger identifies what was sent, the rendered prompt records the instructions, the provider response preserves what was returned, and local validation records whether the returned structure passed the software checks. Manifests record run configuration. Each transformation can thus be inspected, although passing a schema check is no guarantee of a correct interpretation. Table 10 lists paid production stages in rows and recorded charges in US dollars in the numeric column. The final row adds those four charges only. The document and memo quantities embedded in the stage names identify the workload and are not a second currency column. The displayed precision reproduces the ledger; it does not express uncertainty about total research cost, and the subtotal covers the listed production API charges rather than the full cost of conducting or reproducing the research.

Table 10: Recorded production API charges. The subtotal includes only the four listed paid stages; pilots, local hardware, labor, and later analyses are excluded.

Stage	USD
DeepSeek Layered memos, 151,193	314.416569
DeepSeek Direct memos, 51,813	194.753528
Qwen forum track, including retries	74.795158
Qwen compact 2026 track	28.793265
Subtotal of these four stages	612.758520

Collection, normalization, unit extraction, and taxonomy construction carry no API charge in this ledger, which does not mean they were free of computation or labor. Runs declare budgets and preserve transaction-safe spending state. Schema and protocol hashes cover sixteen Qwen-era manifests; older pilots rely on documentary references. Two hardcoded repairs affect earlier artifacts: 28 malformed episode identifiers and seven placeholder chain templates. A reserved-shell-variable crash occurred after the 32,000-token retry rows were committed and is recorded as having no data impact.

Exact regeneration of language-model output is unavailable: DeepSeek has no output seed, and Qwen specifies neither temperature nor seed. Input-sampling and clustering seeds do not remove this limit. Manual merges, prompt-freeze choices, model updates, upstream OpenReview edits, and unarchived historical provider behavior restrict reproduction further. The local taxonomy pipeline is deterministic conditional on its inputs, which is a narrower reproducibility claim than regenerating the unit databases from scratch.

The Method page identifies named rebuild scripts for downstream numbers. Additional persisted checks include paper-clustered uncertainty, score-dispersion comparisons at exact means, reviewer-clustered drift uncertainty, length-quartile remedy checks, echo-threshold sensitivity, and the multinomial reviewer-profile null. Where the same statistic is reused in another figure, it remains one measurement rather than multiple replications.

10.6 Published data, machine-readable claims, and reuse

Readers may want the released artifacts for different tasks: inspecting a particular interpretation, repeating an aggregation, or constructing another classification. Each task needs different inputs, so the public release separates the source record, labeled observations, and derived summaries [39, 42]. Code for collection, memo generation, structuring, taxonomy induction, and aggregation is available at `https://github.com/t46/atlas-of-judgment`, together with provenance documents and

sanitized agent-session history. The primary labeled tables are distributed as Parquet at `https://huggingface.co/datasets/t46/atlas-of-judgment`. Table 11 lists the released products as rows, with a second column describing their contents and what one record represents. Counts refer to the named table's records or units; the API-file count refers to files rather than to new observations. Dataset configuration names distinguish the unit tables from review metadata and from already aggregated chart inputs, each of which needs different joins or calculations for reuse.

Table 11: Public data products and their roles. Dataset configuration names identify the three Parquet tables. The two unit tables overlap in source material; the review metadata table is not a third extraction track.

Product	Contents and observation grain
`review_level_2026`	410,586 units from individual ICLR 2026 reviews.
`review_level_2026_reviews`	74,380 review records with summaries and unit counts.
`forum_level_2018_2026`	1,009,592 units from 50,861 forums, including discussion-phase and update fields.
Taxonomy and rhetoric references	Twelve object and twelve standard definitions; 600 analyst-labeled rhetoric units used by the classifier.
API aggregate files	50 JSON files underlying charts or preserved verification material, catalogued separately from the primary tables.

Unit text is the pipeline's decomposition and restatement of a passage, not a verbatim quotation. The released fields include the four unit components, valence, suggested improvement, support status, model confidence, and taxonomy assignments with cosine similarities. Review-level `review_id` and `paper_id` are OpenReview note identifiers, so a unit can be compared with its original source. Forum-level `reviewer_key` identifies an anonymous reviewer within a forum and should not be assumed to identify a person across papers. Scores, decisions, and sub-scores are absent from these labeled tables and must be joined separately for outcome-linked analyses. The released `analysis/` directory also contains intermediate cluster assignments and exemplars, so the category merges behind the reported summaries can be inspected. The 600 rhetoric labels remain the analyst reference set described in Section 3; they are not a new human validation set.

A claim-level audit interface. The claim-level interface connects a reported result to the aggregate fields and calculations that support it. Here an API means a set of addresses from which a program can retrieve structured records; calling it performs no new analysis. The machine-reader API is a read-only set of static JSON resources over the same aggregates used by the website [37]. Its index catalogues 33 plate and appendix records and 50 aggregate files in the checked release. Each file entry records its byte count, SHA-256 hash, and consuming plates (the website's numbered analysis displays). The hash identifies the file's exact bytes; it does not assess the scientific validity of its contents. The index also labels extraction track and display status. These are catalogue metadata, and the years and denominators specific to an analysis still have to be read from the relevant claim and aggregate. Records at `/api/v1/plates/<plate-id>.json` provide stable claim identifiers, structured quantities and units, denominators where supplied, source arrays or fields, derivations, rebuild scripts, audit tags, and linked caveats. The OpenAPI schema and machine-reader entry point describe this interface.

Figure 7 provides a worked example. The API identifies its clarity–presentation comparison as `plate-i#1c-clarity-presentation` and records a rating-adjusted difference of -0.3465 on the 1–4 scale, with a joined universe of 74,380 reviews. The value is stored in `construct-data.json`, under `matrix.clarity.presentation`; its rebuild script is `scripts/build_construct_data.py`. The deposited aggregate and its index hash were checked during this report update, and the displayed rounded value agrees with the deposited value. The check neither reruns the review-level calculation nor establishes the validity of the construct, since rebuilding the statistic requires the underlying units, the outcome join, and the script. A dated `verified` tag belongs to the source audit record and does not claim fresh independent validation.

Versions, corrections, and attribution. To cite a particular number, give its plate and claim identifier together with a saved release or file hash: an identifier survives presentation changes but does not by itself freeze the value. The public correction ledger preserves the earlier claim, the check that overturned it, and the corrected reading. Reports of further corrections can name the claim and calcu-

lation in the repository's public issue tracker. The core analyses in this report keep their fixed source snapshot; consulting these public interfaces neither replaces the earlier corpus nor recomputes its results.

The Resources page specifies MIT for executable code and page markup, and CC BY 4.0 for the project's text, figures, labeled tables, and derived aggregates [42]. Reuse should attribute Atlas of Judgment and keep its model-produced reading distinct from the original reviewers' words. The underlying review text remains the work of its authors and is available through OpenReview under the applicable source terms; the project's release confers no authorship of that text. Taken together, the artifacts support replotting, checking arithmetic, reclassifying released units, and returning to source reviews, each with its own input requirements and inferential limits.

11 Discussion and limitations

The report asks how written evaluation is organized, how it relates to numerical judgment, and what becomes visible when the discussion continues. These questions share no single denominator. A unit records a local evaluation; a review gathers the evaluations of one reviewer; a paper or panel combines several reviewers; and an exchange restricts attention to participants who write again. Bringing these levels together helps explain a process, but a pattern at one level cannot simply be carried over to another. A frequent criticism need not carry the largest rating association, and a busy discussion need not contain a changed judgment.

The analyses find structure at some levels without establishing it at every level. Object–standard coupling concerns which kinds of justification accompany a given subject of evaluation; these associations are pronounced, and their coarse conditional distribution is comparatively stable. Sequence comparisons ask instead whether order adds regularity beyond the categories already present in a review, and they find only small effects once the review's composition is held fixed. Within novelty criticisms, observation, rule, and request are likewise only weakly coupled at the instrument's resolution. The findings are compatible: a recurring association between a subject and a standard does not require a fixed order of argument or a single shared procedure for reaching a verdict. These regularities describe selected features of written judgment; they do not establish a complete grammar of reasoning.

Novelty stands out across several related summaries: it is often stated once, frequently lacks a suggested repair, is relatively specific in formulation, and has a large negative association with both the reviewer's rating and the paper's decision. These observations are compatible with novelty functioning as an overall evaluative judgment as well as a localized defect. They do not show that a reviewer first selects a verdict and then invents a reason, or that novelty criticism causes rejection. Establishing either mechanism would require evidence beyond text associations.

This interpretation rests on the descriptions taken together; none of them identifies a mechanism on its own. An objection without a repair may concern a feature of the work that is hard to change, but the absence of a written repair does not reveal whether the reviewer considered one privately. Likewise, one underlying concern can feed both a novelty criticism and a low rating. Because the summaries draw on overlapping reviews and extracted categories, their agreement forms a connected account of the record rather than a set of independent confirmations.

The interaction analyses distinguish visible activity from a changed judgment. A reply tree can show engagement without a semantically detected concession, and a reviewer can update a score without writing another argument. The denominator determines which part of this process a rate describes. Starting from all original objections asks how much initial criticism receives visible follow-up. Restricting attention to later judgments asks what happens among evaluations already selected for renewed discussion; exchange-level comparisons further require a reviewer return. Sparse revisitation can therefore coexist with reinforcement among the judgments that reappear. An absent later statement cannot resolve whether a judgment persisted privately or changed without being written down.

The evidence is limited first by the record itself. It covers one venue whose forms, policies, volume, and submission composition change over time. Missing extractions and imperfect identity matching may also depend on content. Neither a large corpus nor precise estimates remove these selection and comparability problems.

The representation introduces additional uncertainty. The two extraction tracks overlap in source material but differ in prompt, unit grain, and inferred-support share. The taxonomy is induced from a noisy cluster structure and transfers a 2026 representation to earlier years; stable assignment similarity cannot substitute for a temporal accuracy audit. Model restatement can increase lexical similarity, remove specific referents, smooth genuine argument structure, and preferentially retain criticism. Finally, many analyses are exploratory or incompletely recorded. The report does not assign confirmatory status to the full collection, and its comparisons do not carry a uniform familywise error guarantee.

These limits act at different points in the inferential chain. An extraction error changes what enters a profile before any statistic is calculated; a category boundary changes which evaluations are counted together; and a missing identity link changes which discussions can enter a longitudinal comparison. Resampling quantifies variation within the chosen representation and sampling unit; by itself, it does not quantify uncertainty introduced before those observations were constructed. A precise estimate can therefore remain an uncertain measurement of the intended construct. The preserved inputs, controls, and source links make those assumptions inspectable; they do not eliminate them.

The information available during extraction also limits the prediction results. An isolated-review memo can receive numerical fields embedded in the public review form, and a whole-forum memo can receive later discussion and decision reasoning. Filtering the resulting units to an initial position does not undo that exposure or recover an earlier unedited source. Cross-validation tests the fitted predictor on held-out records, but does not remove information already present during construction of those records. A prospective test would require archived input versions and an extraction run in which the relevant outcomes and later context are withheld.

Conditioning on paper, rating, length, or set size answers a particular question in each case and does not remove every confounder. Within-paper designs leave reviewer-specific selection and stance in place; fixed-rating construct comparisons condition on another evaluative outcome; length strata are coarse; and data-derived decision cuts summarize the very outcome being described. Point estimates, bootstrap intervals, representation choices, and null-model definitions therefore have to be read together.

The LLM-era results show a clearer change in wording than in the measured structure of criticism. Co-review wording converges within a fixed-vocabulary, constant-form window; the increase in object overlap does not exceed its size-matched null, and coarse reasoning summaries do not show a post-2023 narrowing. These findings do not identify who used a language model, establish its prevalence, or show that review substance is unaffected at finer resolution.

The frozen lexical marker answers a separate question from the similarity and overlap comparisons. Its correlates describe the reviews selected by that vocabulary in each year; they do not establish that marker-positive reviews form a stable AI-use treatment group. The structural nulls, in turn, ask whether observed overlap exceeds a specified size-and-frequency baseline; they do not test whether every meaningful aspect of evaluation stayed unchanged. Keeping these measurement targets apart preserves both the positive wording result and the narrower finding about measured criticism structure.

12 Conclusion

Atlas of Judgment provides a way to study the evaluative logic expressed in public ICLR reviews and discussions. Each unit records an interpretation of how an observation about an object supports a judgment under a standard. We use these units to examine written evaluations and their visible revisitation. Stable object–standard associations and unequal rating deficits coexist with weak sequence effects, much criticism without visible follow-up after the response, and changing institutional and linguistic conditions. The framework makes these patterns available for inspection alongside the null comparisons and measurement limits. Further evidence is needed to connect them to private reasoning or causal mechanisms.

Further validation needs to address several gaps. Representative annotation would check how often the transformations from source text to units succeed, including difficult and omitted cases. Alternative taxonomies would test which descriptions depend on the present partition of evaluation. Temporal validation would ask whether the same instrument measures earlier and later reviews comparably,

and repeated reviewing would supply evidence about reproducibility beyond an observational redraw of the existing ratings. The present report supplies a record on which those checks can be built, and it keeps the distinction between expressed evaluation and private thought explicit.

AI assistance and authorship record. Language models produced the analytic memos and structured units. Model assistance also contributed to code, category interpretation, figures, and the preparation of this report under researcher direction. Reference annotations and audits described as machine readings are not presented as independent human ground truth. The source Method record documents failed pilots, human judgment points, and corrected interpretations. Responsibility for the research and the final text rests with the author.

References

[1] Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Position: Stop automating peer review without rigorous evaluation. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 169349–169380. PMLR, 2026. URL https://proceedings.mlr.press/v306/baumann26a.html.

[2] 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. https://arxiv.org/abs/2306.03262, 2023.

[3] Lutz Bornmann, Christophe Weymuth, and Hans-Dieter Daniel. A content analysis of referees' comments: how do comments on manuscripts rejected by a high-impact journal and later published in either a low- or high-impact journal differ? *Scientometrics*, 83(2):493–506, 2010. doi: 10.1007/s11192-009-0011-4.

[4] Lutz Bornmann, Hanna Herich, Hanna Joos, and Hans-Dieter Daniel. In public peer review of submitted manuscripts, how do reviewer comments differ from comments written by interested members of the scientific community? a content analysis of comments written for atmospheric chemistry and physics. *Scientometrics*, 93(3):915–929, 2012. doi: 10.1007/s11192-012-0731-8.

[5] Souvic Chakraborty, Pawan Goyal, and Animesh Mukherjee. Aspect-based sentiment analysis of scientific reviews. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 207–216, 2020. doi: 10.1145/3383583.3398541. URL https://arxiv.org/abs/2006.03257.

[6] Nuo Chen, Qian Wang, Qingyun Zou, and Bingsheng He. Articulate intuition or genuine analysis? benchmarking epistemic reliability in LLM-as-a-judge peer reviews. https://arxiv.org/abs/2607.10511, 2026.

[7] Liying Cheng, Lidong Bing, Qian Yu, Wei Lu, and Luo Si. APE: Argument pair extraction from peer review and rebuttal via multi-task learning. In *Proceedings of EMNLP 2020*, pages 7000–7011, 2020. doi: 10.18653/v1/2020.emnlp-main.569. URL https://aclanthology.org/2020.emnlp-main.569/.

[8] Corinna Cortes and Neil D. Lawrence. Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment. https://arxiv.org/abs/2109.09774, 2021.

[9] Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of NAACL-HLT 2019*, pages 1274–1290, 2019. doi: 10.18653/v1/N19-1129. URL https://aclanthology.org/N19-1129/. See also the authors' Additional Notes on correlation and causality: https://aclanthology.org/N19-1129v2.pdf.

[10] Tirthankar Ghosal, Sandeep Kumar, Prabhat Kumar Bharti, and Asif Ekbal. Peer review analyze: A novel benchmark resource for computational analysis of peer reviews. *PLOS ONE*, 17(1):e0259238, 2022. doi: 10.1371/journal.pone.0259238. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0259238.

[11] Andrew Gray. Estimating the prevalence of LLM-assisted text in scholarly writing. https://arxiv.org/abs/2512.01560, 2025.

[12] Joshua Guetzkow, Michèle Lamont, and Grégoire Mallard. What is originality in the humanities and the social sciences? *American Sociological Review*, 69(2):190–212, 2004. doi: 10.1177/000312240406900203.

[13] Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In *Proceedings of NAACL-HLT 2019*, pages 2131–2137, 2019. doi: 10.18653/v1/N19-1219. URL https://aclanthology.org/N19-1219/.

[14] Junjie Huang, Win-bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427, 2023. doi: 10.1016/j.joi.2023.101427. URL https://www.sciencedirect.com/science/article/pii/S1751157723000524.

[15] Sven E. Hug. How do referees integrate evaluation criteria into their overall judgment? evidence from grant peer review. *Scientometrics*, 129(3):1231–1253, 2024. doi: 10.1007/s11192-023-04915-y.

[16] Sven E. Hug and Michael Ochsner. Do peers share the same criteria for assessing grant applications? *Research Evaluation*, 31(1):104–117, 2022. doi: 10.1093/reseval/rvab034.

[17] Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 1647–1661, 2018. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

[18] Amir Hossein Kargaran, Nafiseh Nikeghbal, Jing Yang, and Nedjma Ousidhoum. Insights from the ICLR peer review and rebuttal process, 2025. URL https://arxiv.org/abs/2511.15462. Preprint.

[19] Neha Nayak Kennard, Tim O'Gorman, Rajarshi Das, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Hamed Zamani, and Andrew McCallum. DISAPERE: A dataset for discourse structure in peer review discussions. In *Proceedings of NAACL 2022*, pages 1234–1249, 2022. doi: 10.18653/v1/2022.naacl-main.89. URL https://aclanthology.org/2022.naacl-main.89/.

[20] Seungone Kim, Dongkeun Yoon, Kiril Gashteovski, Juyoung Suk, Jinheon Baek, et al. On the limits and opportunities of AI reviewers: Reviewing the reviews of Nature-family papers with 45 expert scientists. https://arxiv.org/abs/2605.20668, 2026.

[21] 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), 2025. doi: 10.1126/sciadv.adt3813. URL https://arxiv.org/abs/2406.07016.

[22] Sandeep Kumar, Tirthankar Ghosal, and Asif Ekbal. When reviewers lock horn: Finding disagreement in scientific peer reviews. In *Proceedings of EMNLP 2023*, pages 16693–16704, 2023. doi: 10.18653/v1/2023.emnlp-main.1038. URL https://aclanthology.org/2023.emnlp-main.1038/.

[23] Sandeep Kumar, Yash Kamdar, Abid Hossain, Bharti Kumari, Tanik Saikh, and Asif Ekbal. When reviews disagree: Fine-grained contradiction analysis in scientific peer reviews. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 38267–38290, 2026. doi: 10.18653/v1/2026.findings-acl.1908. URL https://aclanthology.org/2026.findings-acl.1908/.

[24] Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. https://arxiv.org/abs/2405.02150, 2024.

[25] Grit Laudel. Where do field-specific notions of research quality come from? *Research Evaluation*, 33:rvae027, 2024. doi: 10.1093/reseval/rvae027.

[26] Lingyao Li, Junjie Xiong, Changjia Zhu, Runlong Yu, Chen Chen, Junyu Wang, Renkai Ma, and Zhicong Lu. LLM-as-a-reviewer: Benchmarking their ability, divergence, and prompt injection resistance as paper reviewers. https://arxiv.org/abs/2605.25415, 2026.

[27] 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*, pages 29575–29620. PMLR, 2024. URL https://proceedings.mlr.press/v235/liang24b.html.

[28] Ryan Liu, Steven Jecmen, Vincent Conitzer, Fei Fang, and Nihar B. Shah. Testing for reviewer anchoring in peer review: A randomized controlled trial. *PLOS ONE*, 19(11):e0301111, 2024. doi: 10.1371/journal.pone.0301111. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0301111.

[29] Mathieu Louis, Tibo Vanleke, Vincent Ginis, and Andres Algaba. Rebuttals move peer-review scores, but initial-review structure bounds the movement, 2026. URL https://arxiv.org/abs/2606.22166. Preprint.

[30] Sheng Lu, Ilia Kuznetsov, and Iryna Gurevych. Identifying aspects in peer reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6145–6167, 2025. doi: 10.18653/v1/2025.findings-emnlp.326. URL https://aclanthology.org/2025.findings-emnlp.326/.

[31] Elizabeth L. Pier, Markus Brauer, Amarette Filut, Anna Kaatz, Joshua Raclaw, Mitchell J. Nathan, Cecilia E. Ford, and Molly Carnes. Low agreement among reviewers evaluating the same NIH grant applications. *Proceedings of the National Academy of Sciences*, 115 (12):2952–2957, 2018. doi: 10.1073/pnas.1714379115. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5866547/.

[32] Henri Savall, Véronique Zardet, Marc Bonnet, and Michel Péron. The emergence of implicit criteria actually used by reviewers of qualitative research articles: Case of a european journal. *Organizational Research Methods*, 11(3):510–540, 2008. doi: 10.1177/1094428107308855. URL https://recherche.iseor.com/pdf/articles_de_fond/4Article.pdf.

[33] Vibhhu Sharma, Thorsten Joachims, and Sarah Dean. Do LLMs favor LLMs? quantifying interaction effects in peer review. https://arxiv.org/abs/2601.20920, 2026.

[34] Daniel Scott Smith, Neha Nayak Kennard, Tianyu Du, and Daniel A. McFarland. How values and uncertainty shape scientific advance in peer review. *American Sociological Review*, 90 (5):879–915, 2025. doi: 10.1177/00031224251362254. URL https://journals.sagepub.com/doi/10.1177/00031224251362254.

[35] 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. doi: 10.1371/journal.pone.0287443. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0287443.

[36] Shiro Takagi. Atlas of Judgment: About the Atlas. https://atlas-of-judgment.pages.dev/about, 2026. Project purpose and interpretive scope; accessed 8 November 2026.

[37] Shiro Takagi. Atlas of Judgment: Machine Reader API. https://atlas-of-judgment.pages.dev/api/, 2026. Static API v1; index generated 4 September 2026; accessed 8 November 2026.

[38] Shiro Takagi. Atlas of Judgment: Source Code and Provenance. https://github.com/t46/atlas-of-judgment/tree/7bf84e172e7dc2df24d9a813cdb961b53f9ed358, 2026. Pinned source; input preparation and normalization inspected 8 November 2026.

[39] Shiro Takagi. Atlas of Judgment: ICLR Peer-Review Logic Units. https://huggingface.co/datasets/t46/atlas-of-judgment, 2026. Labeled unit tables, taxonomy, and analysis artifacts; dataset card accessed 8 November 2026.

[40] Shiro Takagi. Atlas of Judgment. https://atlas-of-judgment.pages.dev/, 2026. Main page; source snapshot at repository commit 7bf84e172e7dc2df24d9a813cdb961b53f9ed358.

[41] Shiro Takagi. Atlas of Judgment: How the Data Was Made. https://atlas-of-judgment.pages.dev/method.html, 2026. Method and reproduction record; source snapshot at commit 7bf84e1.

[42] Shiro Takagi. Atlas of Judgment: Resources. https://atlas-of-judgment.pages.dev/resources, 2026. Code, data, citation, and licensing record; accessed 8 November 2026.

[43] Wenqing Wu, Chengzhi Zhang, Yi Zhao, and Tong Bao. Impact of large language models on peer review opinions from a fine-grained perspective: Evidence from top conference proceedings in AI. https://arxiv.org/abs/2604.19578, 2026.

[44] Jing Yang, Qiyao Wei, and Jiaxin Pei. Paper copilot: Tracking the evolution of peer review in AI conferences. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6f951a09ec37047d46773ccf81816e29-Abstract-Conference.html.