REVTRACK: TRACKING REVIEW CONCERNS THROUGH SCIENTIFIC REVISIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

After a manuscript changes, a scientific assistant must decide which reviewer concerns still require action. Generating a plausible critique does not directly evaluate this ability. We introduce REVTRACK, a benchmark for tracking the status of review concerns from author responses and revision summaries. The task distinguishes *fixed*, *partially fixed*, *unresolved*, and *regressed* concerns using evidence tied to the original issue. Evaluation spans ICLR and NeurIPS discussions, with 301 primary-annotator-confirmed validation records and separate source-set diagnostics. A paired-manuscript audit finds agreement on 60 of 84 evaluable issues, relating discussion-based status to implemented changes while making their differences explicit. On an 80-issue ICLR 2025 stress frontier, six prompted LLMs reach only 0.066–0.161 four-label macro-F1, below the majority reference of 0.226. Against the recorded labels, their errors frequently promote unresolved concerns to fixed or partially fixed; voting and simple post-processing leave this failure largely intact. These actively sampled slices diagnose failures rather than estimate their venue-wide prevalence. RevTrack makes evidence-grounded status updates a distinct evaluation target for scientific assistants.

1 INTRODUCTION

A useful scientific assistant must know when to stop repeating a criticism. If a reviewer requests a computational-cost comparison and the revision supplies it, the original concern may no longer apply. Conversely, a detailed response that promises an experiment does not establish that the experiment was performed. Both cases require the assistant to connect an earlier concern to later evidence and decide whether further action is needed. This decision is central to revision support, but it is not directly measured by the quality of a newly generated review.

Peer-review corpora make drafts and reviews available for computational study (Kang et al., 2018; Dycke et al., 2023a). Recent work studies critique generation and reviewer assistance (D'Arcy et al., 2024; Du et al., 2024; Thakkar et al., 2026), while revision resources connect feedback to responses and edits (D'Arcy et al., 2024; Zhang et al., 2025). These resources motivate a complementary evaluation unit: *the status of a particular concern after revision evidence arrives*. An answer can address the right topic yet leave the requested comparison missing; an added experiment can resolve one issue while introducing a new problem on the same technical axis. Measuring these distinctions requires more than response relevance or paper-level quality scores.

We introduce REVTRACK, a benchmark for *discussion-evidence-based revision-status tracking* (Figure 1). Given a reviewer concern, paper context, author response, and revision summary, a model predicts *fixed*, *partially fixed*, *unresolved*, or *regressed*. The rubric checks whether the evidence addresses the same issue, resolves its material content, and describes a concrete change. Each confirmed label is accompanied by an evidence span and a boundary note. This formulation treats the response as evidence to assess, rather than accepting an author's reassurance as proof of implementation.

The benchmark combines evidence-linked validation records with complementary annotation and manuscript audits. Its 301 standard validation records span four venue–year settings; the 148-row ICLR 2024 source evaluation contains a separately identified 40-row locked validation subset. Independent blind reannotation covers 96 records, with agreement reported for its archived annotation version (Section 3). A paired-manuscript audit agrees with the discussion-based status in 60/84

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

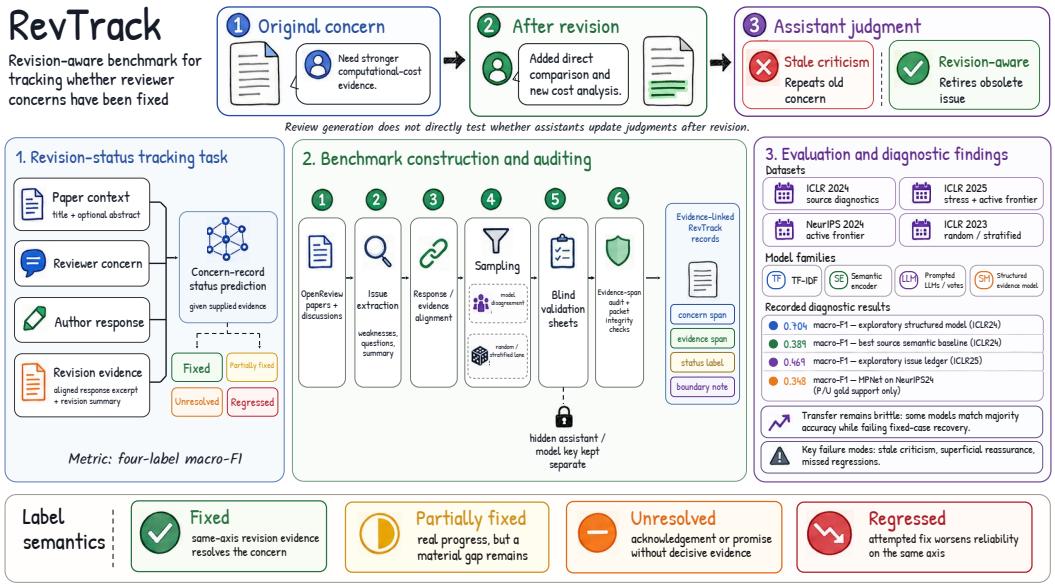

Figure 1: **RevTrack follows concerns through revision.** A schematic example (top) motivates the concern-record task (left), evidence construction and validation (center), and diagnostic evaluation (right). Labels refer to the supplied discussion evidence; Table 1 walks through an actual record. Numerical summaries are recorded diagnostics: structured-model results remain exploratory (Section 4), and NeurIPS supports only partial/unresolved evaluation. Enlarged panels appear in Appendix A.7.

evaluable cases. This comparison exposes a construct boundary: a status justified by a discussion is not a certificate that every manuscript change was implemented.

Our experiments examine which status decisions current systems recover and which errors aggregate scores conceal. Majority accuracy can hide complete failure to recover fixed concerns. On an actively sampled ICLR 2025 frontier, all six prompted LLMs fall below the majority macro-F1 reference, and their confusion matrices reveal systematic promotion of unresolved issues. Paper-grouped cross-encoders and a separate NeurIPS frontier provide additional diagnostics, alongside exploratory structured-model references (Section 4). The evidence supports a need to evaluate revision tracking directly, without assuming that a fluent critique implies reliable follow-up.

Our contributions are threefold:

- **Task:** an issue-level status rubric that distinguishes resolution, partial progress, insufficient evidence, and regression.
- **Benchmark:** evidence-linked validation slices, a bounded independent annotation audit, and a paired-manuscript audit that makes the task's construct boundary explicit.
- **Diagnosis:** evaluations of semantic models, grouped cross-encoders, and prompted LLMs that expose missed resolutions and premature promotion of unresolved concerns.

2 Task Definition

Prediction unit and evidence. An input $x = (p, c, a, r)$ contains paper context p (title and optional abstract), a review-concern excerpt c, author-response candidates and an aligned excerpt a, and a revision summary r. The output $y \in \mathcal{Y} = \{\mathrm{F}, \mathrm{P}, \mathrm{U}, \mathrm{R}\}$ describes the status of that concern given this evidence. Different records from the same paper may receive different labels. Some extracted review-field excerpts bundle multiple requests, so record-level prediction does not guarantee atomic concern segmentation. RevTrack v1 operates on the supplied text; it does not directly verify full manuscripts, figures, equations, or code.

Label semantics. *Fixed* (F) requires concrete evidence that resolves the original concern on the same technical axis. *Partially fixed* (P) requires substantive progress with a material gap remaining. *Unresolved* (U) covers missing or insufficient evidence, including acknowledgement and deferred promises. *Regressed* (R) applies when the attempted change worsens the same issue or introduces a reliability problem tied to it. The label concerns the evidence available to the assistant, not an acceptance decision.

Decision rubric. Annotators check **axis match** (does the evidence address the original technical issue?), **materiality** (does it resolve the scientifically consequential part?), and **evidence sufficiency** (is there a concrete revision rather than reassurance?). Passing all three checks supports F; a substantive but incomplete change supports P; insufficient or deferred evidence supports U. Evidence of deterioration on the same axis supports R. For example, adding an auxiliary ablation while omitting the requested core comparison is P, whereas promising both experiments without supplying evidence is U. Table 1 applies these checks to an existing record: new experiments answer several requests, while an explicitly deferred robustness concern keeps the recorded status at P. The walkthrough paraphrases the supplied discussion evidence; it does not independently verify the reported experiments. Appendix A.5 provides a complementary fixed example and label coverage.

Table 1: **Substantive progress can leave a concern open.** An abridged walkthrough of the existing ICLR 2024 record `9k0krNzvlV__r02`. The label applies to the bundled record, not to each subquestion separately.

Stage	Concern, evidence, and interpretation
Original concern	The record questions watermarking's practical contribution, robustness to fine-tuning, spoofing sample efficiency, changing keys, and possible defenses.
Reported revision	The response and summary report sample-efficiency experiments (source Appendix D) and two-key experiments (source Appendix E); robustness to fine-tuning remains future work.
Axis match	The additions address the requested sample-size and changing-key questions.
Materiality	Acknowledged robustness limitations remain despite the experimental progress.
Evidence sufficiency	The packet identifies concrete reported additions, but does not establish resolution of the entire concern.
Recorded label	**Partially fixed (P):** substantive progress with a material gap remaining.

Metrics. We report accuracy and per-label F1 alongside *fixed-four-label macro-F1*,

$$\mathrm{MacroF1}_4 = \frac{1}{4} \sum_{\ell \in \mathcal{Y}} \mathrm{F1}_\ell, \tag{1}$$

with undefined class F1 set to zero. This convention preserves the four-way task definition, but its scale depends on label coverage: the NeurIPS slice contains only P/U gold labels and therefore has a maximum $\mathrm{MacroF1}_4$ of 0.5. We also report supported-label macro-F1 for the grouped cross-encoders. Comparisons use the same split and metric; scores across slices are not a common-scale ranking of difficulty.

3 Benchmark Construction and Validation

From discussions to issue records. We collect public submissions and discussion notes from OpenReview (OpenReview, n.d.), extract concerns from review fields, and align each with response candidates and revision-summary text. Model disagreement prioritizes difficult candidates; a separate random/stratified lane provides a non-active comparison. Each record retains its source identifiers, concern, aligned response, revision summary, label, evidence span, confidence, and annotation note. Blind validation sheets omit labels, assistant suggestions, and model predictions; separate key files retain these fields for auditing. Checks cover row identity, duplicates, forbidden fields, and evidence completeness before labels enter the standard validation set.

Table 2: **Evaluation scope and validation provenance.** F/P/U/R are label counts. "Standard" counts primary-annotator-confirmed records, totaling 301. The 40 locked ICLR 2024 records are a subset of the 148 source rows.

Slice	n	Standard	F / P / U / R	Sampling / role
ICLR 2024 source	148	40	50 / 86 / 11 / 1	Source diagnostics
ICLR 2025 stress	21	21	5 / 16 / 0 / 0	Small stress set
ICLR 2025 active	80	80	4 / 4 / 66 / 6	Active frontier
NeurIPS 2024 active	80	80	0 / 44 / 36 / 0	Active frontier
ICLR 2023 random	80	80	5 / 49 / 26 / 0	Random/stratified

Evaluation slices and validation provenance. Table 2 distinguishes the source evaluation from the standard validation records. The source manifest records assistant-adjudicated labels for the 148 ICLR 2024 evaluation rows. A 40-row subset has a separate locked-validation confirmation; the other 108 rows do not inherit that subset's status. The 301 standard records comprise these 40 rows, 21 ICLR 2025 stress rows, and three 80-row slices: ICLR 2025 active, NeurIPS 2024 active, and ICLR 2023 random/stratified. Primary standard labels were confirmed by one annotator with evidence spans. The two active frontiers intentionally emphasize model disagreement, so their results characterize selected failure cases rather than venue-wide prevalence. The ICLR 2023 slice offers a bounded non-active comparison, not a representative sample of all review discussions. Candidate-pool accounting and release boundaries appear in Appendix A.1.

Independent reannotation and label versions. A completed independent human blind audit covers 96 selected records from 96 papers; the second annotator saw neither existing labels nor model outputs. We report pre-adjudication agreement and Cohen's κ (Cohen, 1960; Artstein & Poesio, 2008). The archived annotation columns yield 72/96 agreement (75.0%) and $\kappa = 0.551$; neither column contains a regressed label. The audit labels differ from the current evaluation manifest, and no version mapping is available; these statistics therefore describe the archived sample, not reliability of the current evaluation labels. Appendix A.2 reports the label-version comparison.

Does discussion evidence track manuscript changes? An independent human auditor inspected original/revised full-PDF pairs and locked implementation judgments before seeing the discussion-based labels. Of 99 selected issues from 48 papers, PDF pairs were retrievable for 85 issues and 84 received an evaluable judgment; 15 could not be assessed. The paired audit makes the relationship between discussion-based status and manuscript implementation observable: the two judgments agree on 60/84 evaluable issues (71.4%, $\kappa = 0.498$) and differ on 24. This comparison provides an empirical check on the task's evidence boundary, while leaving the causes of individual disagreements unresolved.

4 Models and Evaluation Protocols

We use the following systems as diagnostic probes of concern-status tracking, interpreting each result under its own evaluation protocol.

References and semantic models. The split-specific majority-label reference exposes the effect of class imbalance. Lexical and representation-based baselines use TF-IDF, sentence encoders, MPNet, or ModernBERT features with linear classifiers (Pedregosa et al., 2011; Reimers & Gurevych, 2019; Warner et al., 2025). These recorded baselines test how far generic text representations recover issue status. Their source-set results are kept separate from paper-grouped cross-encoder evaluation.

Paper-grouped cross-encoders. DeBERTa-v3-base and ModernBERT-base jointly encode the concern, response evidence, and revision summary. We use five-fold cross-validation over the 57 ICLR 2024 papers and aggregate three seeds, keeping all issues from each paper in the same fold. Transferred models are evaluated on the ICLR 2025 and NeurIPS 2024 frontiers. The small source set, its uneven labels, and its single regressed issue limit what these results establish about model ca-

Table 3: **Accuracy can hide zero fixed-case recovery.** The ICLR 2025 rows refer to the 21-issue stress slice, not the 80-issue active frontier. All scores are higher-is-better.

Slice	System	Accuracy	Macro-F1$_4$	Fixed F1
ICLR 2024 source	Majority	0.581	0.184	0.000
	TF-IDF	0.561	0.376	0.510
ICLR 2025 stress	Majority	0.762	0.216	0.000
	TF-IDF	0.762	0.216	0.000

pacity. An earlier, separately fine-tuned ModernBERT result is excluded: an identity audit found all 148 evaluation issues and normalized texts in its training data. This excluded experiment is distinct from both the representation-based baseline and the grouped runs reported here.

Prompted LLMs and voting. We evaluate six prompted systems: GPT-5.5, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, GPT-4o-mini, and Qwen2.5-72B. The task asks them to assign the four-way status from the supplied evidence. Three- and six-model votes, unresolved guards, fixed-label suppression, and prompt/temperature variants probe whether simple aggregation or caution changes the error pattern. Post-processing searches are retrospective diagnostics on these slices, not independently validated transfer improvements. Appendix A.4 reports all variants and the limits of replaying closed-model API results.

Exploratory issue-ledger models. The issue ledger encodes explicit revision cues, including added experiments, narrowed claims, deferrals, and response-only fixes. The structured calibrator combines text cues, an issue-ledger label/rule, and semantic-model disagreement, with optional hard overrides. Its recorded source evaluation uses row-wise leave-one-out fitting; transfer uses the labeled source sheet. However, the available lineage record does not establish that ledger features are independent of adjudicated labels or that learned disagreement features were regenerated inside each training fold. We therefore retain these results as exploratory diagnostics, not leakage-controlled evidence of a learned advantage. The paper-grouped cross-encoders and prompted systems provide separate evaluations of the task.

5 Experiments

We ask four questions: whether accuracy captures concern resolution, whether paper-grouped models recover the concern states, how models transfer across selected slices, and whether prompting or voting avoids over-crediting revision progress. All macro-F1 values below use the fixed four-label convention in Equation 1. We interpret scores together with label support and per-label recovery (Sokolova & Lapalme, 2009). Bootstrap intervals quantify sampling variability within each slice (Efron & Tibshirani, 1994; Dror et al., 2018); they do not account for annotation uncertainty or remove selection bias.

5.1 RQ1: Does Accuracy Capture Concern Resolution?

A model can attain substantial accuracy without recovering a single fixed concern (Table 3). On the 21-issue ICLR 2025 stress slice, TF-IDF matches the majority reference at 0.762 accuracy while fixed-case F1 is zero. On ICLR 2024, majority accuracy is 0.581, identical to the recorded MPNet accuracy, although their macro-F1 scores are 0.184 and 0.389 respectively. The difference matters operationally: predicting the dominant class cannot tell an assistant which resolved concerns to retire. Accuracy therefore needs a class-level account of the decisions it rewards.

5.2 RQ2: Do Paper-Grouped Models Recover the Concern States?

Under paper-grouped evaluation, ModernBERT-base and DeBERTa-v3-base reach 0.232 ± 0.008 and 0.209 ± 0.014 macro-F1 (Table 4). Neither recovers unresolved source cases; regression performance

Table 4: **Paper-grouped cross-encoder evaluation.** Mean ± standard deviation across three seeds; source predictions use five folds over 57 papers. $F1_4$ averages all four labels; $F1_{sup}$ averages gold-supported labels. NeurIPS supports only P/U.

	ICLR24 source	ICLR25 active	NeurIPS24 active	
Model	**$F1_4$**	**$F1_4$**	**$F1_4$**	**$F1_{sup}$**
DeBERTa-v3-base	0.209 ± 0.014	0.091 ± 0.116	0.163 ± 0.025	0.326 ± 0.051
ModernBERT-base	0.232 ± 0.008	0.073 ± 0.021	0.180 ± 0.012	0.359 ± 0.023

Table 5: **Two active frontiers, different label support.** Both slices have $n = 80$. $F1_4$ uses four labels; F/P/U/R columns report class F1. NeurIPS has no F/R gold cases, so those class metrics are omitted. † marks exploratory ledger-dependent rows. Compare methods within each slice.

	ICLR 2025: F/P/U/R				NeurIPS 2024: P/U			
Model	**Acc.**	**$F1_4$**	**F**	**R**	**Acc.**	**$F1_4$**	**P**	**U**
TF-IDF	0.050	0.026	0.000	0.000	0.550	0.177	0.710	0.000
ModernBERT features	0.125	0.092	0.056	0.000	0.200	0.137	0.229	0.320
MPNet features	0.312	0.276	0.174	0.500	0.550	0.348	0.516	0.875
Structured†	0.150	0.298	0.229	0.800	0.362	0.197	0.545	0.244
Issue ledger†	0.850	0.469	0.429	0.500	0.500	0.211	0.167	0.679

cannot be estimated stably from one source example. These results demonstrate difficulty under the specified small-data protocol, without isolating architecture, data size, or optimization effects.

Appendix Table 8 reports the historical semantic and structured diagnostics. Their row-wise evaluation and unresolved structured-feature lineage require separate interpretation from the paper-grouped results above (Section 4).

5.3 RQ3: What Transfers Across Years and Venues?

The two 80-issue active frontiers expose different failures (Table 5). On ICLR 2025, 66 of 80 labels are unresolved. TF-IDF recovers none of the four fixed or six regressed issues, whereas MPNet reaches 0.276 macro-F1. The exploratory issue ledger reaches 0.469, but this result remains subject to the lineage qualification; it does not establish a general solution. On NeurIPS, the supported labels are 44 partially fixed and 36 unresolved issues. TF-IDF obtains 0.550 accuracy by predicting the dominant partial class while recovering no unresolved cases. MPNet's 0.348 four-label macro-F1 reflects recovery within this two-label slice; it should not be read on the same absolute scale as a four-label frontier.

These slices were selected partly through model disagreement. Their results establish failures on difficult transferred issues, not a causal effect of venue or an estimate of how often such errors occur in natural review traffic. The ICLR 2023 random/stratified slice broadens the validation coverage, but we do not report it as an additional model-transfer result. Appendix Table 9 provides within-slice uncertainty intervals.

5.4 RQ4: Do Prompting and Voting Avoid Over-crediting Progress?

Prompted systems improve over the majority reference on the ICLR 2024 source set, yet all six single-model variants fall below it on the ICLR 2025 frontier (Figure 2; Table 6). Their macro-F1 values range from 0.066 to 0.161, compared with 0.226 for majority. The three-model vote reaches 0.071; its best reported U/F post-processing variant reaches only 0.094. On NeurIPS, selected prompted systems and votes remain close to the 0.177 majority reference, with paired-bootstrap intervals for their differences spanning zero (Appendix Table 14).

Table 6: **All prompted models and voting variants.** Four-label macro-F1 on the 148-row source set and two 80-row active frontiers. U/F denotes unresolved-guard plus fixed-suppression post-processing. Its rule selection is retrospective. Values should be compared within columns; no cross-slice mean is used because support and sampling differ.

Family	System	ICLR24 source	ICLR25 active	NeurIPS24 active
Reference	Majority	0.184	0.226	0.177
Single LLM	GPT-5.5 (v2)	0.350	0.081	0.171
	GPT-4.1	0.326	0.070	0.166
	GPT-4o-mini	0.299	0.077	0.178
	GPT-4.1-nano	0.210	0.161	0.126
	Qwen2.5-72B	0.334	0.072	0.165
	GPT-4.1-mini	0.347	0.066	0.181
Vote	Vote-3 (top)	0.352	0.071	0.179
	Vote-3 (diverse)	0.347	0.088	0.174
	Vote-3 (U-guard)	0.345	0.092	0.170
	Vote-3 (U/F)	0.350	0.094	0.167
	Vote-6 (all)	0.344	0.064	0.177

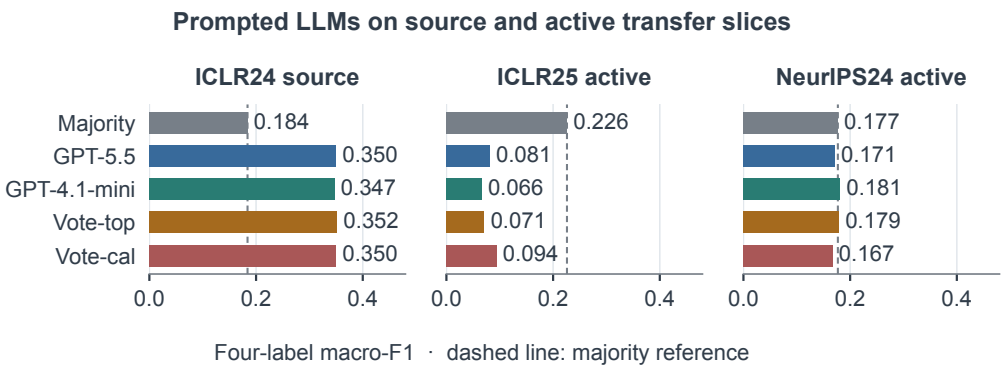

Figure 2: **Prompted-model performance depends on the evaluation slice.** Dashed lines mark the majority reference. All displayed prompted/vote systems fall below it on ICLR 2025. NeurIPS has only P/U gold support, so compare systems within each panel. Vote-cal denotes retrospective U/F post-processing, not a separately validated model.

The confusion matrices identify the practical failure behind these scores (Figure 3). GPT-5.5 assigns 63 of 66 unresolved issues to F or P, while the post-processed vote does so for 62. Relative to the recorded evaluation labels, these predictions give unresolved concerns excessive resolution credit. The opposite error also matters: majority preserves unresolved cases but misses every fixed, partial, and regressed case in this slice. A useful assistant must distinguish these outcomes rather than choose a globally optimistic or conservative default. Figure 4 shows the recorded error decomposition; Appendix Table 10 connects these errors to revision-support decisions. Post-processing searches and decoding ablations appear in Appendix A.4.

6 Related Work

Peer-review corpora and revision resources. Peer-review datasets began by making drafts, reviews, scores, and decisions computationally accessible (Kang et al., 2018; Dycke et al., 2023a). Other resources study propositions in reviews (Hua et al., 2019), rebuttal effects (Gao et al., 2019), and review–rebuttal argument pairs (Cheng et al., 2020; Bao et al., 2021). Complementary resources annotate discourse and pragmatic structure in review discussions (Kennard et al., 2022; Dycke et al.,

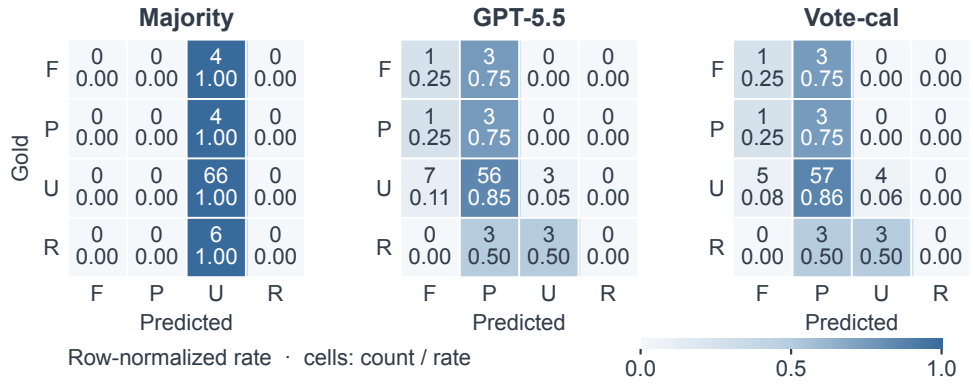

Figure 3: **Premature promotion of unresolved concerns.** ICLR 2025 active-frontier confusion matrices; cells show counts and gold-row-normalized rates. F/P/U/R denote the four status labels. GPT-5.5 promotes 63 of 66 unresolved cases to F/P; the post-processed vote promotes 62. Majority retains unresolved cases but misses every other class.

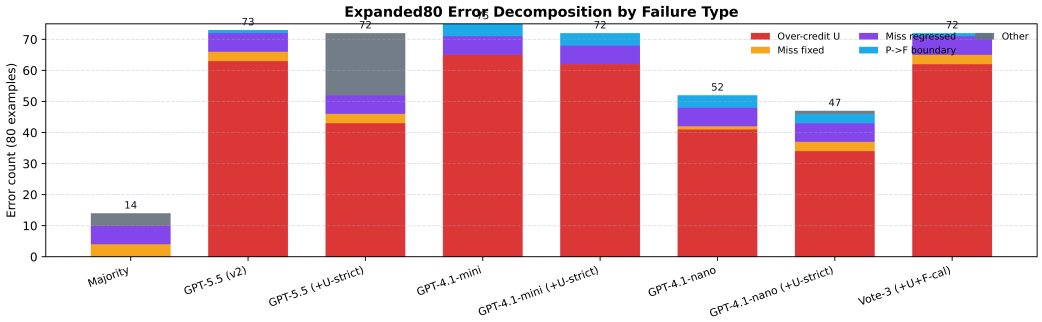

Figure 4: **ICLR 2025 active-frontier error decomposition.** Premature promotion of unresolved concerns dominates the prompted variants. Counts describe this selected slice; they do not estimate error prevalence across a venue.

2023b). Recent datasets cover multi-turn peer-review dialogue and review aspects (Tan et al., 2024a; Lu et al., 2025). PeerQA, LazyReview, and ClaimCheck add question-answering, heuristic, and claim-grounded views of the same ecosystem (Baumgärtner et al., 2025; Purkayastha et al., 2025; Ou et al., 2025). ARIES links review comments to manuscript edits, while Re2 provides full-stage peer-review and multi-turn rebuttal discussions (D'Arcy et al., 2024; Zhang et al., 2025). RevTrack defines a complementary prediction target over a concern and its subsequent evidence: whether the concern is fixed, partially fixed, unresolved, or regressed. A response or edit can be relevant to a concern while leaving its material request unmet; the status rubric makes that distinction explicit.

Automatic review generation and review-feedback tools. Earlier systems decomposed reviewing into score prediction, weakness identification, comment generation, review understanding, and meta-review support (Leng et al., 2019; Wang et al., 2020; Staudinger et al., 2024). Neural and LLM-based work studies whether models can automate scientific reviewing or provide useful feedback to authors and reviewers (Yuan et al., 2022; Liang et al., 2024; Du et al., 2024; Chamoun et al., 2024). Recent systems add focused feedback, multi-agent review generation, specialized review models, and review simulation (D'Arcy et al., 2024; Jin et al., 2024; Yu et al., 2024; Idahl & Ahmadi, 2025). Other lines explore deeper review frameworks, tree-structured review, and multimodal review benchmarks (Zhu et al., 2025; Chang et al., 2025; Gao et al., 2025). Other work evaluates the strengths and limitations of generated reviews (Li et al., 2025), trains review models with reinforcement learning (Zeng et al., 2025), or studies reviewer-facing interventions (Thakkar et al., 2026). RevTrack asks

what happens after such feedback is acted on, when the assistant must decide whether the original criticism should remain live.

Evidence grounding, claim verification, and correction. Evidence-verification benchmarks ask whether claims are supported, refuted, or unverifiable with respect to text or scientific evidence (Thorne et al., 2018; Wadden et al., 2020). Neighboring work studies behavioral testing, conflicting contexts, and adaptive QA (Ribeiro et al., 2020; Liu et al., 2025; Shaier et al., 2024). Truthful answering, hallucination analysis, and factual precision metrics test related forms of evidence discipline (Lin et al., 2022; Ji et al., 2023; Min et al., 2023). Self-checking and attribution repair study whether models can inspect or ground their own outputs (Manakul et al., 2023; Gao et al., 2023). Self-refinement, verbal feedback, and automatic correction ask whether models can revise outputs through iterative feedback (Madaan et al., 2023; Shinn et al., 2023; Pan et al., 2024). RevTrack is adjacent, but its proposition is a historical review concern whose status can move after revision evidence arrives.

LLM judges, assisted annotation, and calibration under skew. LLM-as-judge and LLM-assisted annotation are attractive for scalable evaluation (Zheng et al., 2023; Liu et al., 2023; Tan et al., 2024b; Gu et al., 2026), but documented judge biases warrant validation against human assessments (Wang et al., 2024). Separately, research on neural-network calibration and distribution shift motivates reporting uncertainty and checking transfer (Guo et al., 2017; Ovadia et al., 2019; Koh et al., 2021). RevTrack separates blind validation sheets from hidden assistant/model keys. The evaluated prompted systems supply predictions, not replacement gold labels. Appendix Section A.6 summarizes the distinction between these settings and longitudinal issue tracking after revision.

7 Discussion

RevTrack separates three questions that a revision assistant can otherwise conflate: whether an author responded, whether the response supplies evidence addressing the concern, and whether the manuscript implements the reported change. The task measures the second question; the paired-PDF audit relates it to the third without equating the two. This separation also clarifies the role of the evidence span: it allows a reader to inspect why a concern was promoted, retained, or marked as worse.

The observed errors suggest a concrete interface requirement. An assistant should preserve the original concern, identify the evidence that changes its status, and leave the issue open when that evidence is insufficient. The results motivate evaluating these decisions separately from critique generation.

Limitations. The standard validation set is primarily confirmed by one annotator; the separate audit does not establish current-version label reliability (Section 3). The 40 locked ICLR 2024 validation rows must also be distinguished from the full 148-row source evaluation. Active sampling, small slices, and uneven label support preclude venue-level prevalence estimates and stable regression metrics. The historical structured diagnostics have an unresolved feature-lineage limitation, and API-based results may change with model versions. Some concern excerpts bundle multiple requests, without a separately validated aggregation rule. Discussion summaries can also omit changes visible only in manuscripts, figures, equations, or code. These boundaries make the present benchmark most informative as a diagnostic of evidence-based status decisions within the evaluated slices.

8 Conclusion

RevTrack evaluates whether a scientific assistant can update the status of a reviewer concern when new evidence arrives. Its four-way rubric, evidence-linked records, and paired-manuscript audit make the decision and its limits explicit. The evaluated systems can miss resolved issues or over-promote unresolved ones, particularly on selected transfer cases. Scientific-assistant evaluation should therefore test which concerns remain open, alongside the quality of the critiques an assistant can generate.

Ethics Statement

RevTrack uses public OpenReview discussions for research on assistant behavior, not for automated acceptance decisions. Public availability does not remove privacy concerns: the intended release should retain only necessary excerpts, provenance, evidence-linked labels, and audit artifacts rather than reproduce complete review threads. Blind validation packets remain separate from assistant/model keys. Prematurely closing a concern can mislead authors and reviewers; repeatedly raising a resolved concern can also waste effort. Unequal credit for polished versus terse responses is a further risk. These considerations favor evidence-linked recommendations subject to human review.

Reproducibility Statement

Section 2 defines the labels and metric convention; Section 3 documents sampling and validation; Section 4 distinguishes the evaluation protocols. Appendix A.1 documents artifact roles and feature provenance; Appendix A.4 reports prompted-model variants, uncertainty analyses, and available execution metadata. Incomplete archived execution settings and changes to hosted APIs limit exact reruns; the missing records are specified in Appendix A.4.

AI Use Statement

Generative AI tools assisted manuscript organization and editing, literature search and reference cross-checking, interpretation of reported results, and preparation of figures and LaTeX artifacts. Model predictions also enter candidate prioritization and the evaluated baselines, as described in Sections 3 and 4; their role is distinguished from the human-validation protocols reported in Section 3. The authors are responsible for the final text, references, claims, and artifacts, including material prepared with AI assistance.

References

Ron Artstein and Massimo Poesio. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008. doi: 10.1162/coli.07-034-R2. URL https://aclanthology.org/J08-4004/.

Jianzhu Bao, Bin Liang, Jingyi Sun, Yice Zhang, Min Yang, and Ruifeng Xu. Argument pair extraction with mutual guidance and inter-sentence relation graph. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3923–3934, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.319. URL https://aclanthology.org/2021.emnlp-main.319/.

Tim Baumgärtner, Ted Briscoe, and Iryna Gurevych. PeerQA: A scientific question answering dataset from peer reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 508–544, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.22. URL https://aclanthology.org/2025.naacl-long.22/.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604, 2018. doi: 10.1162/tacl_a_00041. URL https://aclanthology.org/Q18-1041/.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9742–9763, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.580. URL https://aclanthology.org/2024.findings-acl.580/.

Yuan Chang, Ziyue Li, Hengyuan Zhang, Yuanbo Kong, Yanru Wu, Hayden Kwok-Hay So, Zhijiang Guo, Liya Zhu, and Ngai Wong. TreeReview: A dynamic tree of questions framework for deep and efficient LLM-based scientific peer review. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15651–15682, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.790. URL https://aclanthology.org/2025.emnlp-main.790/.

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 the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7000–7011, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.569. URL https://aclanthology.org/2020.emnlp-main.569/.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi: 10.1177/001316446002000104.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers, 2024. URL https://arxiv.org/abs/2401.04259.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. ARIES: A corpus of scientific paper edits made in response to peer reviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6985–7001, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.377. URL https://aclanthology.org/2024.acl-long.377/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Jesse Dodge, Suchin Gururangan, Dallas Card, Roy Schwartz, and Noah A. Smith. Show your work: Improved reporting of experimental results. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2185–2194, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1224. URL https://aclanthology.org/D19-1224/.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1383–1392, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1128. URL https://aclanthology.org/P18-1128/.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Ranran Haoran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5081–5099, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.292. URL https://aclanthology.org/2024.emnlp-main.292/.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. NLPeer: A unified resource for the computational study of peer review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5049–5073, Toronto, Canada, July 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.277. URL https://aclanthology.org/2023.acl-long.277/.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. Overview of PragTag-2023: Low-resource multi-domain pragmatic tagging of peer reviews. In *Proceedings of the 10th Workshop on Argument Mining*, pp. 187–196, Singapore, December 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.argmining-1.21. URL https://aclanthology.org/2023.argmining-1.21/.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1994. doi: 10.1201/9780429246593.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16477–16508, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.910. URL https://aclanthology.org/2023.acl-long.910/.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. MMReview: A multidisciplinary and multimodal benchmark for LLM-based peer review automation, 2025. URL https://arxiv.org/abs/2508.14146.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1274–1290, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1129. URL https://aclanthology.org/N19-1129/.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92, 2021. doi: 10.1145/3458723. URL https://cacm.acm.org/research/datasheets-for-datasets/.

Feng Gu, Zongxia Li, Carlos Rafael Colon, Benjamin Evans, Ishani Mondal, and Jordan Lee Boyd-Graber. Large language models are effective human annotation assistants, but not good independent annotators. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 71–89, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.4. URL https://aclanthology.org/2026.findings-acl.4/.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Xinyu Hua, Mitko Nikolov, Nikhil Badugu, and Lu Wang. Argument mining for understanding peer reviews. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 2131–2137, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1219. URL https://aclanthology.org/N19-1219/.

Maximilian Idahl and Zahra Ahmadi. OpenReviewer: A specialized large language model for generating critical scientific paper reviews. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, pp. 550–562, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. doi: 10.1145/3571730. URL https://dl.acm.org/doi/10.1145/3571730.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024*

Conference on Empirical Methods in Natural Language Processing, pp. 1208–1226, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.70. URL https://aclanthology.org/2024.emnlp-main.70/.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1149. URL https://aclanthology.org/N18-1149/.

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 the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1234–1249, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.89. URL https://aclanthology.org/2022.naacl-main.89/.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton A. Earnshaw, Imran S. Haque, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

Youfang Leng, Li Yu, and Jie Xiong. DeepReviewer: Collaborative grammar and innovation neural network for automatic paper review. In *2019 International Conference on Multimodal Interaction*, pp. 395–403. ACM, 2019. doi: 10.1145/3340555.3353766. URL https://doi.org/10.1145/3340555.3353766.

Ruochi Li, Haoxuan Zhang, Edward Gehringer, Ting Xiao, Junhua Ding, and Haihua Chen. Unveiling the merits and defects of LLMs in automatic review generation for scientific papers. In *2025 IEEE International Conference on Data Mining (ICDM)*, pp. 1370–1379. IEEE, 2025. doi: 10.1109/ICDM65498.2025.00146. URL https://doi.org/10.1109/ICDM65498.2025.00146.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8), 2024. doi: 10.1056/AIoa2400196. URL https://doi.org/10.1056/AIoa2400196.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/.

Siyi Liu, Qiang Ning, Kishaloy Halder, Wei Xiao, Zheng Qi, Phu Mon Htut, Yi Zhang, Neha Anna John, Bonan Min, Yassine Benajiba, and Dan Roth. Open domain question answering with conflicting contexts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1838–1854, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-naacl.99. URL https://aclanthology.org/2025.findings-naacl.99/.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-EVAL: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46534–46594, 2023. doi: 10.52202/075280-2019. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91edff07232fb1b55a505a9e9f6c0ff3-Abstract-Conference.html.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 220–229. Association for Computing Machinery, 2019. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

OpenReview. About OpenReview, n.d. URL https://openreview.net/about. Accessed September 21, 2026.

Jiefu Ou, William Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel Khashabi, and Benjamin Van Durme. CLAIMCHECK: How grounded are LLM critiques of scientific papers? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 21712–21735, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1185. URL https://aclanthology.org/2025.findings-emnlp.1185/.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. Automatically correcting large language models: Surveying the landscape of diverse automated correction strategies. *Transactions of the Association for Computational Linguistics*, 12:484–506, 2024. doi: 10.1162/tacl_a_00660. URL https://aclanthology.org/2024.tacl-1.27/.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85): 2825–2830, 2011. URL https://www.jmlr.org/papers/v12/pedregosa11a.html.

Sukannya Purkayastha, Zhuang Li, Anne Lauscher, Lizhen Qu, and Iryna Gurevych. LazyReview: A dataset for uncovering lazy thinking in NLP peer reviews. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3280–3308, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.165. URL https://aclanthology.org/2025.acl-long.165/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Sagi Shaier, Ari Kobren, and Philip V. Ogren. Adaptive question answering: Enhancing language model proficiency for addressing knowledge conflicts with source citations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17226–17239, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.956. URL https://aclanthology.org/2024.emnlp-main.956/.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8634–8652, 2023. doi: 10.52202/075280-0377. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1b44b878bb782e6954cd888628510e90-Abstract-Conference.html.

Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009. doi: 10.1016/j.ipm.2009.03.002. URL https://doi.org/10.1016/j.ipm.2009.03.002.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pp. 257–268, Bangkok, Thailand, August 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.sdp-1.24/.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. Peer review as a multi-turn and long-context dialogue with role-based interactions, 2024a. URL https://arxiv.org/abs/2406.05688.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 930–957, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.54. URL https://aclanthology.org/2024.emnlp-main.54/.

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

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1074. URL https://aclanthology.org/N18-1074/.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://papers.neurips.cc/paper/7181-attention-is-all-you-need.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pp. 384–397, Dublin, Ireland, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.inlg-1.44. URL https://aclanthology.org/2020.inlg-1.44/.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard, and Iacopo Poli. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2526–2547. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.127. URL https://aclanthology.org/2025.acl-long.127/.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, Renjing Cui, Chengcheng Han, Qiushi Sun, Zhiyong Wu, Yunshi Lan, and Xiang Li. Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 10164–10184, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.595. URL https://aclanthology.org/2024.findings-emnlp.595/.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212, 2022. doi: 10.1613/jair.1.12862. URL https://www.jair.org/index.php/jair/article/view/12862.

Sihang Zeng, Kai Tian, Kaiyan Zhang, Yuru Wang, Junqi Gao, Runze Liu, Sa Yang, Jingxuan Li, Xinwei Long, Jiaheng Ma, Biqing Qi, and Bowen Zhou. ReviewRL: Towards automated scientific review with RL. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 16931–16943, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.857. URL https://aclanthology.org/2025.emnlp-main.857/.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. Re2: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions, 2025. URL https://arxiv.org/abs/2505.07920.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623. Curran Associates, Inc., 2023. doi: 10.52202/075280-2020. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. DeepReview: Improving LLM-based paper review with human-like deep thinking process. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29330–29355, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1420. URL https://aclanthology.org/2025.acl-long.1420/.

A Additional Analyses and Audit Details

This appendix documents sample provenance, annotation and manuscript audits, historical model diagnostics, and the full prompted-model analysis. It records the protocols and evidence boundaries needed to interpret each result.

A.1 Data Accounting and Release Boundaries

Evaluation rows versus standard validation. The 148-row ICLR 2024 source evaluation contains 57 papers and includes the 40-row locked validation subset. Only those 40 source rows contribute to the stated 301 standard validation records. The remaining components are 21 ICLR 2025 stress records and 80 records each from ICLR 2025 active, NeurIPS 2024 active, and ICLR 2023 random/stratified slices. Thus, the 148-row source evaluation and the 301-row validation collection are overlapping objects, not interchangeable descriptions of one dataset. The full source result must not inherit a validation claim established only for its locked subset.

Candidate pools. The ICLR 2025 expanded pool contains 322 candidates from 80 submissions, with complete-field rate 0.963 and 244 model-disagreement rows. The NeurIPS 2024 pool contains 393 candidates, with complete-field rate 1.000 and 316 disagreement rows. The evaluated 80-row frontiers are selected subsets of these pools. In the source artifacts, these slices are named `expanded80` and `limit100`; the latter is a collection identifier, not the number of evaluated issues. The ICLR 2023 `random80` slice provides a separate random/stratified validation lane.

Artifact roles. The construction workflow separates public source text, blind validation packets, assistant/model keys, confirmed labels, and metric summaries. Dataset and model documentation motivate recording collection scope, annotation provenance, and evaluation conditions (Bender & Friedman, 2018; Mitchell et al., 2019; Dodge et al., 2019; Gebru et al., 2021). The intended release includes minimally necessary excerpts, source identifiers, evidence-linked labels, prediction summaries, and regeneration scripts; it should exclude complete review threads and hidden validation keys. This manifest describes the planned release rather than current public availability.

Training overlap and feature lineage. The row-identity audit found all 148 source evaluation issue IDs and normalized texts in the 180-row train-v8 sheet, spanning 57 shared papers. The separately fine-tuned ModernBERT result obtained from that overlap is withdrawn. The transfer manifest records no shared issue or paper IDs between train-v8 and the ICLR 2025/NeurIPS frontiers; that check is not equivalent to a complete normalized-text contamination audit. For the structured model, independence of issue-ledger labels/rules from adjudicated labels and fold-local generation of learned disagreement features remain unverified. These are substantive limits on interpretation, even when a metric table is internally consistent.

A.2 Annotation and Manuscript Audits

Completed blind reannotation and label-version comparison. The completed second pass contains 96 issues from 96 papers, sampled with split stratification and independently annotated without existing labels or model output. Agreement before adjudication was 72/96 (75.0%); Cohen's $\kappa = 0.551$, with a paper-cluster bootstrap 95% interval of $[0.401, 0.691]$. The 24 disagreements consist of 16 F/P and 8 P/U cases. Neither pre-adjudication annotation column contains a regressed label. A row-identity join finds that 59 of the 96 first-pass labels differ from the current evaluation manifest: 19/46 ICLR 2024 issues, 24/25 ICLR 2025 issues, and 16/25 NeurIPS 2024 issues. Of these mismatches, 46 occur where the two archived annotators agreed, so the 24 annotation disagreements alone cannot explain the difference. No label-version mapping is available, so the 75.0% estimate describes agreement between the recorded pre-adjudication annotations rather than agreement with the current evaluation labels. Version disagreement does not identify which labels are incorrect; both label versions are retained without substitution. Table 7 distinguishes version differences from annotator disagreement.

Earlier consistency checks. The mini60 check records 60/60 agreement and $\kappa = 1.0$. The boundary160 check records 160/160 agreement and $\kappa = 1.0$, but involved prelabel assistance before

Table 7: **Annotation-version reconciliation remains open.** The archived first-pass label is compared with the current evaluation manifest after joining by issue ID. These mismatches are not the 24 disagreements between the two archived annotators. No labels were replaced during this check.

Slice in annotation packet	Matched records	First-pass / manifest mismatches
ICLR 2024	46	19
ICLR 2025	25	24
NeurIPS 2024	25	16
Total	96	59

confirmation. These records are reported separately from the blind 96-issue audit: their selection and assistance conditions do not support an equivalent inter-annotator-reliability claim.

Paired full-PDF audit. The manuscript audit selected 99 issues from 48 papers and locked implementation judgments before discussion-based labels were revealed. PDF pairs were retrievable for 85 issues; 84 had evaluable judgments and 15 selected issues could not be assessed. The evaluable implementation judgments were 13 full, 65 partial, and 6 none, mapped to F/P/U respectively. This implementation-derived status is the comparison target used by the archived reconciliation script. Of the 84 evaluable rows, 61 also have an explicit original PDF-status label consistent with the mapping; 23 have an unknown PDF-status field but an evaluable implementation judgment. The mapping has no regressed branch and therefore does not test four-way PDF-status annotation. Exact status agreement was 60/84 (71.4%), with $\kappa = 0.498$. The comparison tests agreement with implementation evidence in selected manuscripts, not scientific sufficiency or regression detection. Unassessable issues remain in the selected-sample accounting rather than disappearing from the denominator description.

A.3 Historical Source and Transfer Diagnostics

Table 8 preserves the recorded source-set results, including the structured model's 0.704 macro-F1 and its 0.424 ablation without hard overrides. These values do not resolve the feature-lineage limitation in Section 4. The single source regressed example also makes four-label macro-F1 sensitive to one observation: any one class contributes up to 0.25 to the average. The representation-based semantic baselines use standard text-classification and Transformer representations (Vaswani et al., 2017; Devlin et al., 2019); they are distinct from the separately trained, withdrawn ModernBERT run.

Table 8: **Historical ICLR 2024 source diagnostics.** Values reproduce the recorded 148-row evaluation. † marks exploratory ledger-dependent models with unresolved feature lineage (Section 4). These rows are not directly comparable to paper-grouped cross-encoders, and the 148 rows are not all part of the 40-row locked validation subset.

Model	Accuracy	Macro-F1$_4$	Fixed F1
Majority label	0.581	0.184	0.000
TF-IDF + LinearSVC	0.561	0.376	0.510
ModernBERT features	0.561	0.387	0.509
MPNet features	0.581	0.389	0.554
Structured, no hard rules†	0.655	0.424	0.602
Structured†	0.682	0.704	0.624

Table 9 reports the existing stratified-bootstrap intervals for these model/slice pairs. Fixed label counts explain the degenerate intervals of constant majority predictions. These intervals quantify variability conditional on the recorded labels and sampling scheme; they cannot establish feature independence, annotation validity, or representativeness. Differences are retained from the original metric records and may differ slightly from subtraction of displayed rounded values.

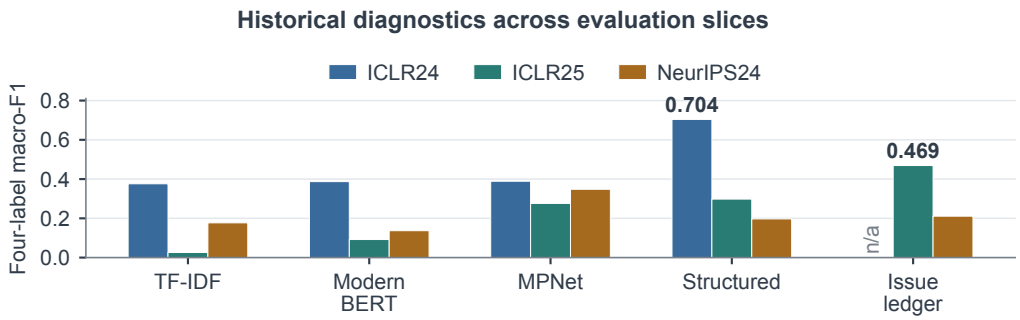

Figure 5: **Recorded source and transfer diagnostics.** Structured and issue-ledger results are exploratory. Slice-dependent label support and different source evaluation protocols prevent interpreting bar height as a common-scale ranking across datasets. No in-domain issue-ledger value is reported.

Table 9: Recorded stratified-bootstrap macro-F1 intervals. Structured and issue-ledger rows remain exploratory. The 148-row source set is distinct from its 40-row locked validation subset. Δ is the gap to the displayed reference model within each block, calculated before rounding.

Split	Model	F1 [95% CI]	Δ	Reading
ICLR24 clean	Structured	0.704 [0.630, 0.775]	0.000	Exploratory source.
	Struct. no overrides	0.424 [0.352, 0.494]	0.280	Recorded ablation.
	MPNet	0.389 [0.320, 0.461]	0.314	Source diagnostic.
	ModernBERT	0.387 [0.311, 0.464]	0.317	Source diagnostic.
	TF-IDF	0.376 [0.305, 0.447]	0.328	Lexical limit.
	Majority	0.184 [0.184, 0.184]	0.520	Skewed floor.
ICLR25 exp80	Issue ledger	0.469 [0.323, 0.577]	0.000	Hard frontier.
	Structured	0.298 [0.218, 0.352]	0.171	Transfer brittle.
	MPNet	0.276 [0.140, 0.372]	0.193	Uncertain gap.
	Majority	0.226 [0.226, 0.226]	0.243	Skewed ref.
	ModernBERT	0.092 [0.054, 0.134]	0.377	Misses U/R.
	TF-IDF	0.026 [0.025, 0.028]	0.443	Collapses.
NeurIPS24 limit100	MPNet	0.348 [0.299, 0.390]	0.000	U/P frontier.
	Issue ledger	0.211 [0.173, 0.253]	0.136	Above majority.
	Structured	0.197 [0.141, 0.250]	0.150	Weaker transfer.
	Majority	0.177 [0.177, 0.177]	0.170	No F/R rows.
	TF-IDF	0.177 [0.177, 0.177]	0.170	Majority-like.
	ModernBERT	0.137 [0.082, 0.193]	0.211	Below majority.

A.4 Prompted LLM Transfer Diagnostics

Full results and failure modes. Table 6 in the main text reports all six single-model systems and the vote variants. The ICLR 2025 active slice is dominated by unresolved labels, whereas the NeurIPS slice contains only partial/unresolved labels. Interpretation therefore follows within-slice differences and class-level errors, not the mean rank across slices. Figure 3 in the main text shows the mechanism for GPT-5.5 and the U/F vote: most unresolved issues are promoted to fixed or partial. Figure 4 in the main text shows the recorded error decomposition; Table 10 summarizes the corresponding decision errors.

Table 10: **Errors as revision-support decisions.** Illustrative counts come from the ICLR 2025 active frontier. The issue-ledger example is exploratory; regression support is only six issues.

Error	Decision affected	Recorded example
Stale criticism	A resolved concern remains active.	TF-IDF misses all 4 fixed issues.
Premature promotion	Insufficient evidence receives F/P credit.	GPT-5.5 promotes 63/66 U issues.
Partial/full boundary	An incomplete change is treated as a full fix.	Issue ledger: 4 boundary errors.
Missed regression	A harmful change is not recognized.	TF-IDF misses all 6 R issues.

Post-processing and decoding. The unresolved guard prioritizes unresolved votes; fixed suppression reduces fixed predictions under disagreement. Table 11 reports the retrospective rule search. Its best displayed ICLR 2025 macro-F1 is 0.094, with unresolved recall 0.061. Selection on these same diagnostics means this improvement is not an independent generalization estimate. Temperature and unresolved-strict prompting variants appear in Table 12; their trade-offs do not remove the observed failure on this frontier.

Table 11: **Retrospective post-processing search.** Macro-$F1_4$ and ICLR 2025 unresolved recall for three-model vote rules. "Fixed suppression" avoids a fixed prediction under vote disagreement. The selected rule has not been evaluated on an independent tuning/test split.

Rule	ICLR24 source	ICLR25 active	ICLR25 U recall
U-guard + fixed suppression	0.350	0.094	0.061
Any unresolved vote wins	0.345	0.092	0.061
Suppress fixed on disagreement	0.357	0.073	0.015
Base vote	0.352	0.071	0.015

Table 12: Expanded80 decoding and prompt ablation. The tested higher temperatures lower GPT-4o-mini macro-F1. Unresolved-strict prompting changes recall/F1 trade-offs without removing the observed frontier failure.

Model	Prompt	Temp	Macro	U rec.	Reading
GPT-4o-mini	default	0.0	0.077	0.015	Weak baseline.
GPT-4o-mini	default	0.3	0.020	0.015	Lower recorded F1.
GPT-4o-mini	default	0.7	0.021	0.015	Lower recorded F1.
GPT-4o-mini	U-strict	0.0	0.032	0.015	Strictness insufficient.
GPT-5.5(v2)	default	0.0	0.081	0.045	Over-credits U rows.
GPT-5.5(v2)	U-strict	0.0	0.059	0.091	U recall up; F1 down.
GPT-4.1-mini	default	0.0	0.066	0.015	U recall near zero.
GPT-4.1-mini	U-strict	0.0	0.086	0.061	Best mini; still low.
GPT-4.1-nano	default	0.0	0.161	0.379	Best default; below maj.
GPT-4.1-nano	U-strict	0.0	0.171	0.485	U up; fixed falls.

Uncertainty and comparisons. Table 13 reports example-level bootstrap intervals. Table 14 separately reports paired differences described in the original summary as stratified bootstrap. The majority intervals in these two tables imply different resampling conditions; the interval-generation records are unavailable, so the complete protocols have not been independently reconciled. For the selected systems, the paired difference intervals are above zero on ICLR 2024, below zero on ICLR 2025, and overlap zero on NeurIPS. They characterize the saved predictions on the evaluated slices, conditional on the recorded labels; they do not include annotation uncertainty, post-selection uncertainty, or future API changes.

Table 13: Bootstrap intervals for selected prompted LLM and vote baselines, with the intended claim reading made explicit. Intervals capture sample instability only; they do not include annotator uncertainty or API stochasticity.

Split	Method	n	F1 [95% CI]	Reading
ICLR24	Majority	148	0.184 [0.167, 0.198]	Reference floor.
ICLR24	GPT-5.5	148	0.350 [0.279, 0.430]	In-domain gain; split-specific.
ICLR24	GPT-4.1-nano	148	0.210 [0.155, 0.263]	Close to reference.
ICLR24	GPT-4.1-mini	148	0.347 [0.272, 0.437]	In-domain gain; split-specific.
ICLR24	Vote-3 (U+F)	148	0.350 [0.282, 0.428]	Source-specific.
ICLR25	Majority	80	0.226 [0.212, 0.237]	Transfer reference.
ICLR25	GPT-5.5	80	0.081 [0.021, 0.158]	Below majority.
ICLR25	GPT-4.1-nano	80	0.161 [0.116, 0.207]	Best prompted; below majority.
ICLR25	GPT-4.1-mini	80	0.066 [0.017, 0.114]	Does not transfer.
ICLR25	Vote-3 (U+F)	80	0.094 [0.026, 0.184]	Below majority.
NeurIPS24	Majority	80	0.177 [0.152, 0.199]	Cross-venue reference.
NeurIPS24	GPT-5.5	80	0.171 [0.144, 0.193]	Overlaps reference.
NeurIPS24	GPT-4.1-nano	80	0.126 [0.075, 0.179]	Lower, uncertain.
NeurIPS24	GPT-4.1-mini	80	0.181 [0.154, 0.203]	Overlaps reference.
NeurIPS24	Vote-3 (U+F)	80	0.167 [0.141, 0.189]	Overlaps reference.

Table 14: Reported paired-bootstrap differences against majority (the original summary specifies stratification; complete generation records are unavailable). Δ is macro-F1(method) minus macro-F1(majority); the reading column states the claim supported by each interval.

Split	Method	F1	Δ	95% CI	Reading
ICLR24	GPT-5.5 (v2)	0.350	0.167	[0.093, 0.250]	Above majority.
ICLR24	GPT-4.1-mini	0.347	0.164	[0.091, 0.245]	Above majority.
ICLR24	Vote-3 (+U+F)	0.350	0.167	[0.097, 0.244]	Source-specific.
ICLR25	GPT-5.5 (v2)	0.081	-0.145	[-0.200, -0.068]	Below majority.
ICLR25	GPT-4.1-mini	0.066	-0.160	[-0.176, -0.136]	Below majority.
ICLR25	Vote-3 (+U+F)	0.094	-0.132	[-0.198, -0.047]	Below majority.
N24	GPT-5.5 (v2)	0.171	-0.006	[-0.019, 0.004]	Overlaps.
N24	GPT-4.1-mini	0.181	0.003	[-0.012, 0.018]	Overlaps.
N24	Vote-3 (+U+F)	0.167	-0.011	[-0.024, 0.001]	Overlaps.

Available records and reproducibility. The available metric records identify grouped cross-encoder seeds 17, 29, and 43, the model families, and the evaluated slices. The archive lacks complete per-example model outputs, original prompted input templates, vote membership and tie-breaking rules, versioned API configurations, and all per-run optimization, checkpoint-selection, and truncation settings. The summaries therefore support analysis of the recorded evaluations but do not

fully specify their execution. Exact reproduction is limited by these missing records and changes to hosted APIs; model aliases do not identify immutable service versions. Prompted outputs are evaluated as predictions, not promoted to human labels.

A.5 Label Coverage and Worked Boundaries

Table 15 lists the label support used to interpret each slice. An absent gold class has no recoverable examples; a zero class F1 in that case is a metric convention, not evidence of failed recovery. NeurIPS consequently supports P/U comparisons, while the six ICLR 2025 regressed cases remain a small diagnostic sample.

Table 15: **Label support and interpretation.** F/P/U/R denotes the count of each label. The source set includes only 40 locked validation rows; NeurIPS has no fixed or regressed gold examples.

Slice	n	F/P/U/R	Evaluation scope
ICLR24 source	148	50/86/11/1	Source diagnostics; one regressed case.
ICLR25 stress	21	5/16/0/0	Small F/P stress slice.
ICLR25 active	80	4/4/66/6	Four-way active-frontier diagnostics.
NeurIPS24 active	80	0/44/36/0	P/U diagnostics; not four-way recovery.
ICLR23 random	80	5/49/26/0	Bounded random/stratified validation.

Companion fixed example. The ICLR 2024 source record `RzNlECeoOB__r03` requests experiments for a hierarchical model and minor presentation corrections. Its response and revision summary report a newly implemented two-layer t^3HVAE comparison with Gaussian HVAE, locate the results in Section 4.2, Figure 5, and Table C3 of the source paper, and report the presentation corrections. The current manifest records F: the supplied evidence describes concrete changes on the requested axes. This contrasts with the deferred material concern in the recorded P example in Table 1. Both walkthroughs paraphrase evidence packets to illustrate the rubric; they are not additional annotations, independent manuscript verification, or estimates of label frequency.

A.6 Benchmark Positioning

PeerRead, NLPeer, and Re2 make peer-review material accessible at corpus scale. ARIES aligns review comments with edits; PeerQA studies reviewer information needs; FEVER and SciFact evaluate support for claims. Review-generation and review-feedback work measures the content of critiques or assistance. RevTrack instead labels the status of a historical concern given subsequent evidence. This is a complementary prediction target: the present experiments do not establish how strongly performance on those other tasks correlates with revision-status accuracy.

A.7 Enlarged Framework Panels

These views enlarge the original framework layout in Figure 1. The pipeline describes record construction; it does not imply that every historical audit validates the current labels.

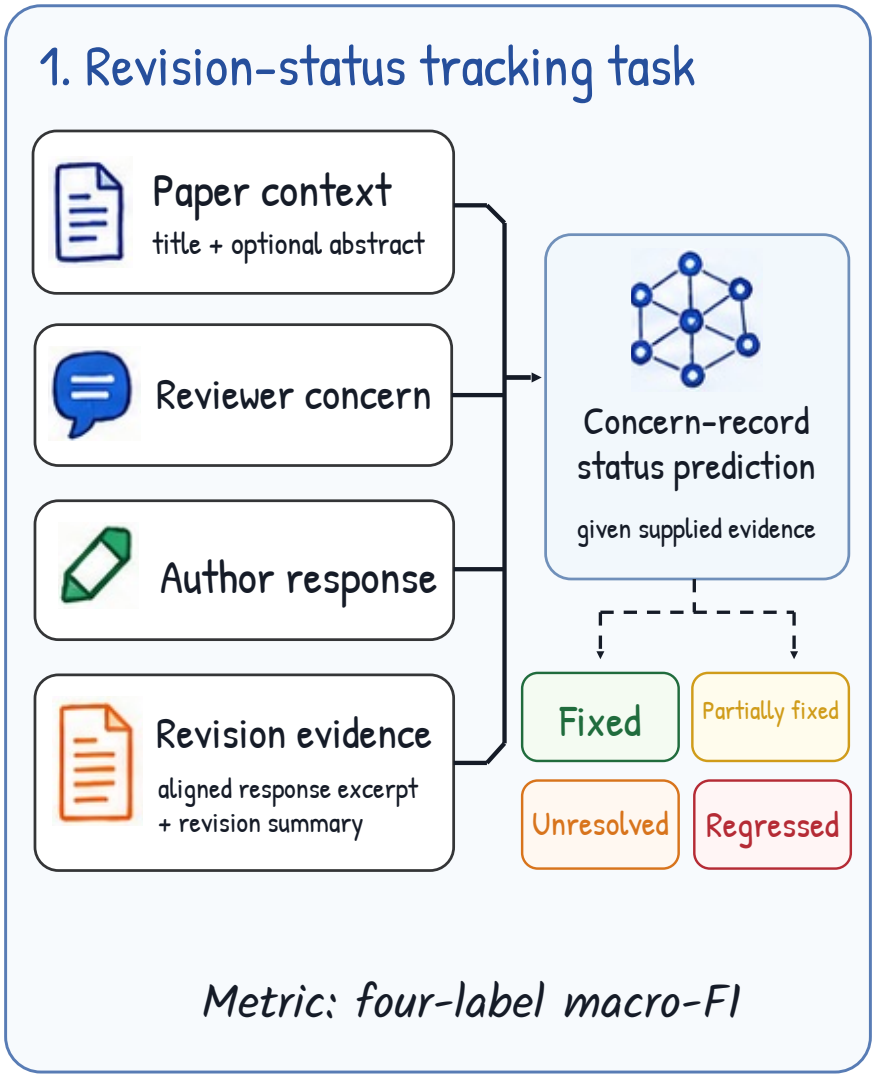

Figure 6: **Task and evidence inputs.** Labels describe the supplied evidence for a concern record; some excerpts bundle multiple requests (Section 2).

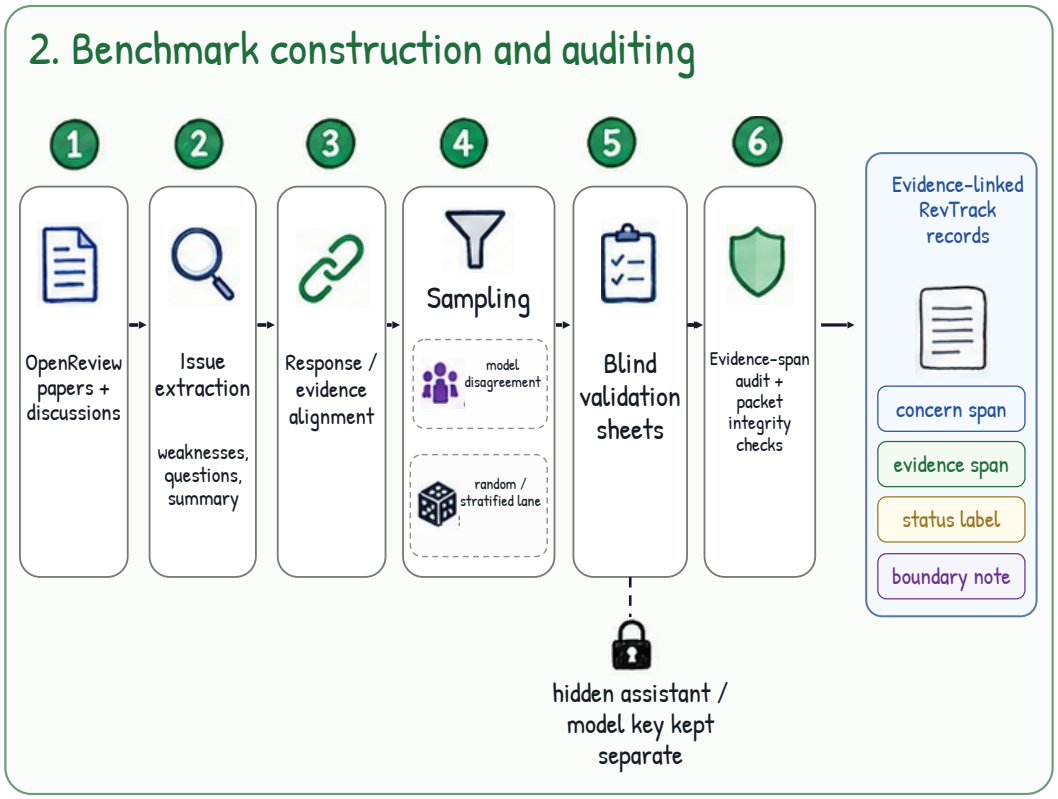

Figure 7: **Construction and audit workflow.** Source collection, evidence alignment, sampling, and validation packets are distinct stages. Primary confirmation and independent version-matched agreement are distinct evidence requirements (Section 3).

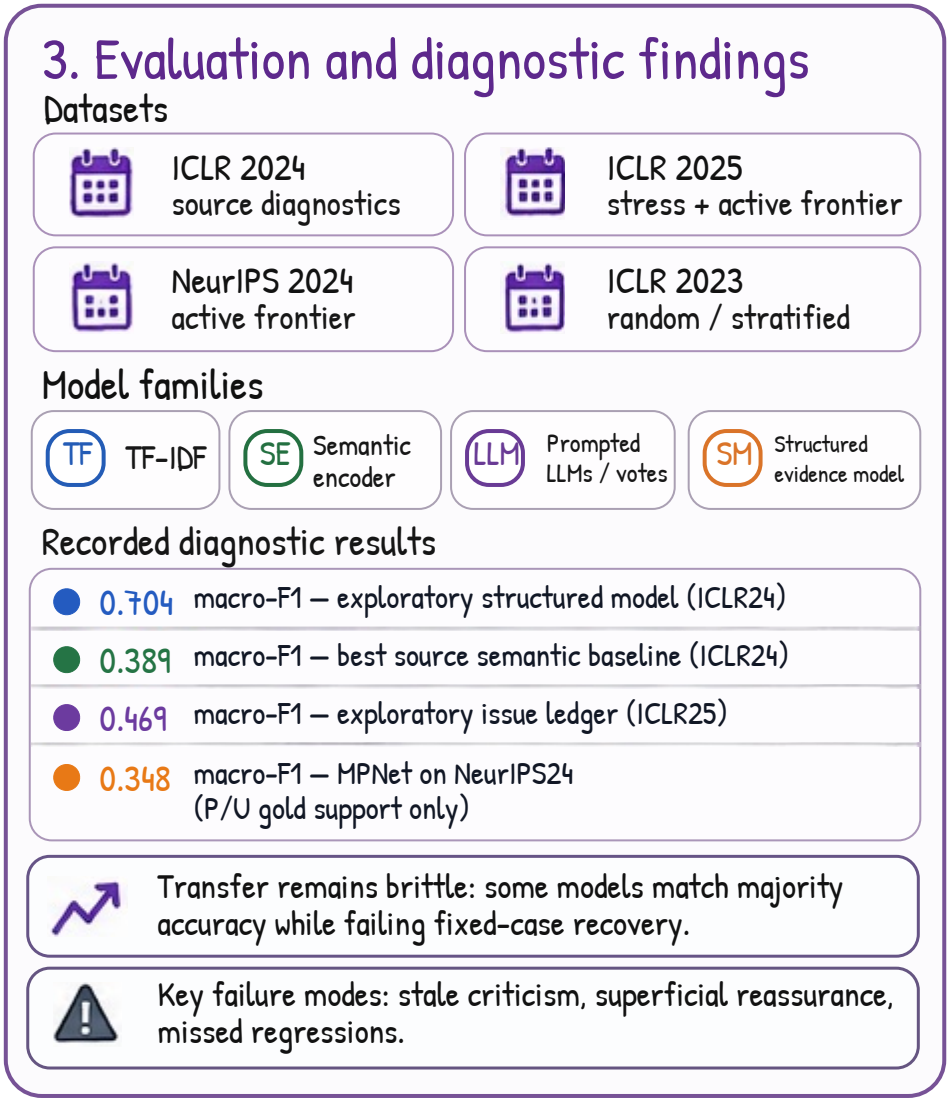

Figure 8: **Evaluation scope and recorded diagnostics.** The structured and issue-ledger scores remain exploratory because feature lineage is unresolved. NeurIPS supports P/U labels only; numerical comparisons must respect each slice's protocol and label support.