# What Makes a Good AI Review? Concern-Level Diagnostics for AI Peer Review

**Ming Jin**
Bradley Department of Electrical & Computer Engineering
Virginia Tech

## Abstract

Evaluating AI-generated reviews by verdict agreement is widely recognized as insufficient, yet current alternatives rarely audit which concerns a system identifies, how it prioritizes them, or whether those priorities align with the review rationale that shaped the final assessment. We propose *concern alignment*, a diagnostic framework that evaluates AI reviews at the concern level rather than only at the verdict level. The framework's core data structure is the *match graph*, a bipartite alignment between official and AI-generated concerns annotated with match type, severity, and post-rebuttal treatment. From this artifact we derive an *evaluation ladder* that moves from binary accuracy to concern detection, verdict-stratified behavior, decision-aware calibration, and rebuttal-aware decomposition. In a pilot study of four published systems, concern-level analysis suggests that detection alone does not determine review quality; calibration is often the binding constraint. Systems detect non-trivial fractions of official concerns yet most mark 25–55% of concerns on accepted papers as decisive, where, under our operationalization, no official concern on accepted papers was treated as a decisive blocker. Identical overall verdict accuracy can conceal reject-heavy behavior versus low-recall profiles, and low full-review false decisive rates can partly reflect concern dilution rather than calibrated prioritization. Most systems do not emit a native accept/reject, and inferring it from review tone is method-sensitive, reinforcing the need for concern-level diagnostics that remain stable across inference choices. The contribution is a reusable evaluation framework for auditing which concerns AI reviewers identify, how they weight them, and whether those priorities align with the review rationale that informed the paper's final assessment.

## 1 Introduction

A good AI review identifies the right concerns, assigns them the right weight, and aligns with the rationale that actually decides the paper. Most AI review systems are evaluated primarily by verdict agreement or coarse similarity to human reviews (Liang et al., 2024b; Lu et al., 2024; D'Arcy et al., 2024; ChicagoHAI, 2026; Gao et al., 2025). Recent work has started to move beyond verdict matching by comparing attention over review facets (Shin et al., 2025), scoring review quality dimensions (Garg et al., 2025), checking premise-level factuality (Ryu et al., 2025), and benchmarking limitation identification (Xu et al., 2025). But researchers do not act on verdicts or facet histograms alone; they act on concrete concerns and on how much those concerns should matter.

Verdict-only evaluation misses the structure that makes reviews useful. A system can reach moderate verdict accuracy by rejecting almost everything (Appendix T), yet that behavior is invisible in overall accuracy. It can recover only a small fraction of the concerns that actually drove the decision (Table 3), or it can recover the right concern family but assign it materially different severity (Figure 1; Figure 3). These are substantively different failure modes, but overall accuracy compresses them into one number. It cannot tell whether a system reached the right verdict for the wrong reasons, surfaced the right concerns but miscalibrated their decision weight, or flagged concerns that the official process ultimately treated as non-blocking. Because reviews are consumed as prioritized concern lists, evaluation that discards concern-level structure discards the unit researchers actually use.

We introduce **concern alignment**, a diagnostic framework organized as an *evaluation ladder*. Level 0 (L0) asks whether the verdict is right; Level 1 (L1) whether the system surfaces the same concerns officials raised; Level 2 (L2) whether its behavior changes across accepted and rejected papers; Level 3 (L3) whether it assigns decision weight appropriately; and Level 4 (L4) whether it attends to the concerns that mattered most for the decision. The ladder's central artifact is the *match graph*: a bipartite alignment between official and AI-generated concerns annotated with match type, severity, and the area chair's post-rebuttal treatment. All primary metrics derive from this single, auditable artifact.

The pilot's main empirical lesson is that detection alone does not determine review quality; calibration is often the binding constraint. A system can notice a real weakness yet still be wrong about the paper because it overstates how much that weakness should count. It can also produce a long, weakly prioritized concern list that eventually overlaps official concerns without telling an author which issues most deserve revision effort first. These are failures of selective attention, not just of detection.

We demonstrate the framework on four public systems (System L (Liang et al., 2024b), System A (Lu et al., 2024), System O (ChicagoHAI, 2026), and System M (D'Arcy et al., 2024)), across papers from three top-tier ML venues. Concern-level analysis reveals that identical verdict accuracy can conceal reject-heavy and low-recall profiles; that most single-agent systems mark 25–55% of concerns on accepted papers as decisive; that changing the model while holding prompts fixed shifts reviewer behavior in measurable, non-uniform ways; and that some systems attend as readily to resolved concerns as to decisive blockers (§4).

## 2 Concern Alignment Framework

### 2.1 Design Principles

We organize the framework around three criteria for a useful review. These are modeling assumptions; readers may reasonably weight them differently. We state them to make the evaluation's design ground explicit.

**Prioritization.** Useful reviews distinguish blocking concerns from improvement suggestions. When decisive blockers are present, they should be identifiable; when they are absent, real but non-blocking issues should not be inflated into blockers. The distinction between "this must be fixed for the paper to be acceptable" and "this would improve the paper" is often the most consequential judgment in the review. Appendix F provides worked examples of severity determination for each system, and Appendix Q lists representative decisive and non-decisive concerns to support independent calibration of this judgment.

**Distinguishing power.** Useful reviews should change their concern profile with paper quality. If a system raises the same kinds of concerns, at the same severity, on accepted and rejected papers, its feedback provides weak guidance about which issues most deserve revision effort.

**Aligned coverage.** A useful AI review should recover the concerns that informed the official review rationale, may surface additional valid concerns officials did not mention, and should avoid generic, tangential, or unsupported complaints.

These criteria decompose into measurable sub-properties, detection and calibration, tested by the evaluation ladder (§2.3). Verdict and concern-level analysis are complementary rather than competing signals: the verdict provides the directional label, while concerns provide the diagnostic decomposition. Conditioning concern analysis on the verdict (L2 and above) reveals, for example, that a system with moderate overall accuracy rejects nearly every accepted paper, a behavior invisible to both binary accuracy and verdict-blind recall.

### 2.2 Match Graphs

The framework uses area chair (AC) decisions as an operational anchor for calibrating concern severity and decision weight. Human decisions are themselves noisy (Cortes & Lawrence, 2021; Beygelzimer et al., 2023), and some accepted papers may contain genuine blockers that were deprioritized. We adopt this anchor because post-deliberation AC judgments, informed by reviews, rebuttals, and discussion, provide the best available proxy for how each concern was ultimately weighted. The

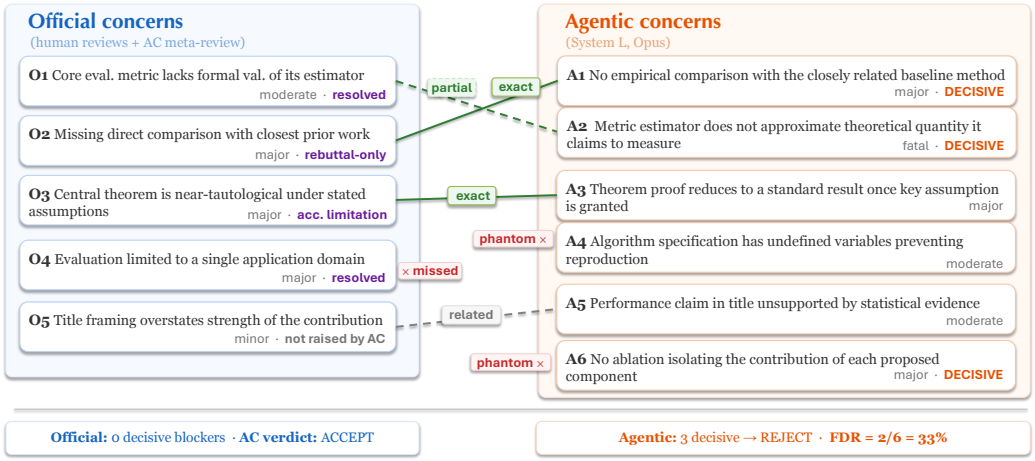


Figure 1: Match graph for an accepted paper. Official concerns (left) carry severity labels and AC treatment badges; agentic concerns (right) carry severity and decisive flags. Edges show match type: `exact` and `partial` count for strict metrics; `related` is excluded. Unmatched official concerns are *misses*; unmatched agentic concerns are *phantoms*. Two of six agentic concerns are false decisive flags (33%); the third decisive flag (A1) correctly identifies a weakness whose fix is not visible in the reviewed PDF.

assumption is conservative: it gives a concrete reference point for measuring calibration while acknowledging its limitations (§6).

A *match graph* is a bipartite alignment between official concerns $\mathcal{O} = \{o_1, \ldots, o_m\}$ (from human reviews, rebuttal, and meta-review) and agentic concerns $\mathcal{A} = \{a_1, \ldots, a_n\}$ (from the AI review) for a single paper. Each official concern carries severity $s_i \in \{\text{fatal}, \text{major}, \text{moderate}, \text{minor}\}$ and an *AC treatment* label recording the area chair's post-rebuttal disposition (decisive blocker, unresolved, accepted limitation, resolved, dismissed, reframed feature, or not mentioned when the AC does not address the concern). Each agentic concern carries severity and a `decisive` flag. An edge $e_{ij}$ connects $o_i$ to $a_j$ when both address the same issue: `exact` if fixing one fixes the other, `partial` if they share an issue family but differ in scope, `related` if topically nearby but distinct (excluded from strict metrics). Each edge is annotated with severity alignment under a hybrid tolerance rule: fatal requires an exact match, while one-level gaps among non-fatal concerns count as matches and larger gaps are coded as under/over. Unmatched official concerns are *misses*; unmatched agentic concerns are *phantoms* (Figure 1).

Match graphs are constructed in five steps: (1) official concern extraction from OpenReview with severity, AC treatment, and decisive flags; (2) agentic concern extraction, deduplicated across review sections; (3) bipartite matching via scope test; (4) semantic verification by an independent auditor with 32 calibration exemplars (Appendix R); (5) metric derivation. Three design choices are central: using AC treatment as an operational anchor enables rebuttal-aware metrics (under the assumption that post-deliberation AC judgments, while noisy, provide the best available proxy for how each concern was ultimately weighted); only exact and partial edges count for strict metrics; and severity uses the same hybrid tolerance rule throughout the paper (Appendix N).

### 2.3 The Evaluation Ladder

The ladder has five levels, grouped into three broader stages: verdict agreement (L0), concern-set alignment (L1–L2), and decision-weight calibration (L3–L4). Each level answers a diagnostic question the level below cannot. Figure 2 illustrates the progression on a single accepted-paper review; formal definitions appear in Appendix A.

L0 **Binary accuracy** provides a directional signal but little diagnostic information; on a balanced set, a reject-everything system and a random one are indistinguishable. L1 **Concern detection** measures *concern recall* (the fraction of official concerns with a strict match) and *phantom rate* (the fraction of agentic concerns with no strict match); both are verdict-blind. Not all phantoms are equal: many are

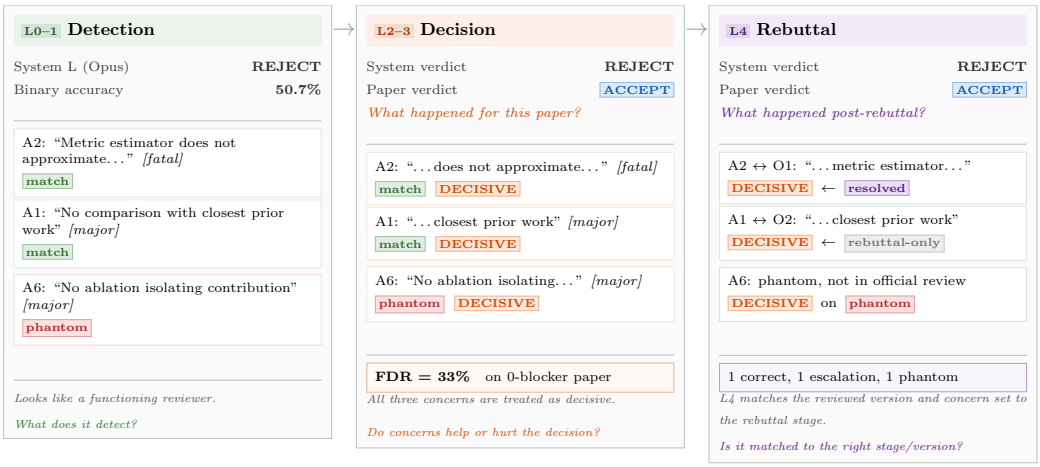


Figure 2: Progressive reveal: the same AI review (System L, Opus, accepted paper) diagnosed at increasing depth. **Left**: looks like a functioning reviewer. **Center**: All three shown concerns carry decisive flags on a paper with zero blockers (FDR = 33%). **Right**: the evaluation must match the reviewed paper version and concern set to the rebuttal stage. A2 escalates a *resolved* concern (calibration failure); A1 correctly detects a weakness whose fix never appeared in the reviewed PDF (*rebuttal-only*); A6 is a fabricated phantom.

legitimate concerns officials did not raise, while others are fabricated or over-severe. We distinguish *harmful phantoms* (fatal/major unmatched concerns on accepted papers) from benign ones at Level 3. L2 **Verdict-stratified metrics** split Level 0–1 metrics by accepted versus rejected papers, revealing reject-heavy behavior that overall accuracy hides. L3 **Decision-aware metrics** introduce the *false decisive rate* (FDR): among agentic concerns on accepted papers, what fraction is marked decisive? Under our AC-aligned operationalization, the count of decisive blockers on accepted papers is zero.[1] On rejected papers, where real decisive blockers exist, two complementary metrics replace FDR: *decisive precision (strict)* measures the fraction of agentic decisive flags that match an official decisive blocker, and *phantom decisive rate* measures the fraction of total output that consists of fabricated decisive concerns (unmatched decisive flags). *Resolved-escalation rate* measures how often the system re-escalates concerns the AC marked resolved when the corresponding fix is visible in the reviewed PDF. We also decompose concerns into *decision-relevant* (correct matches at appropriate severity on rejected papers, or constructive feedback on accepted ones) and *decision-harmful* (re-escalation of dismissed or resolved concerns, harmful phantoms on accepted papers, and severity under-rating or missed blockers on rejected papers), reporting both components separately rather than collapsing them. L4 **Rebuttal-aware decomposition** computes recall separately for each AC treatment category (decisive blocker, unresolved, accepted limitation, resolved), revealing the system's *attention profile*.

All Level 3–4 metrics can also be computed at *top-$K$* by restricting attention to the $K$ most severe agentic concerns per paper. This is not a replacement for full-review metrics; it answers a different usage question. Full-review metrics evaluate the complete output. Top-$K$ metrics evaluate the prioritized core that a reader is most likely to act on. They are especially informative when systems generate 2–3$\times$ different concern volumes: a low full-review FDR can reflect genuine calibration or simple concern dilution. That distinction matters because researchers triaging a review will typically focus on the top few concerns first (§4.2).

### 2.4 Measurement Validation

An independent auditor using ChatGPT 5.4 Pro re-verified 191 candidate edges across 9 held-out papers. The auditor worked from local edge worksheets rather than downstream system scores. Two independent runs of the auditor on the same edges agreed on 96.9% of labels; Cohen's $\kappa$ was 0.918 for

[1] See the assumption discussion in §2.2 and §6. Any non-zero FDR on accepted papers represents concerns the AC did not treat as blocking under this operationalization.

verdict and 0.946 for both match type and severity, indicating near-perfect consistency (Appendix M). In independent extraction audits, 16 of 18 official concern sheets and 51 of 54 agentic sheets were rated satisfactory on completeness and label quality, with no extracted concern unsupported by the source text. The dominant matching error is *scope inflation* (where one concern bundles the other's complaint with additional independent demands; 59% of errors), not false matches. A phantom-quality audit ($N$=200, independently rated by two annotators, 19 inter-annotator disagreements resolved by human adjudication) further suggests that many phantoms are legitimate concerns officials did not raise.

Because match graph construction uses LLMs to evaluate LLM-generated reviews, circularity remains a methodological risk. We mitigate it through cross-model validation with a different model family ($\kappa > 0.91$), human adjudication of all auditor disagreements and a random sample of agreement cases, 32 calibration exemplars spanning 6 error categories (Appendix R), and extraction audits of all 48 official concern sheets plus a stratified sample of agentic sheets for severity accuracy and AC-treatment coding (§6; Appendix M).

## 3 Pilot Study Setup

We demonstrate the framework on four public AI review systems applied to 48 papers from ICLR 2026, NeurIPS 2025, and ICML 2025, all in AI safety/alignment (24 accepted, 24 rejected; 670 official concerns, 79 decisive blockers). The goal is to show what concern-level diagnostics reveal about reviewer behavior, not to produce a leaderboard. The sample is sufficient for a framework demonstration but not for population-level estimates; we state that limitation explicitly in §6.

**Data curation.** Papers were sourced from OpenReview and filtered to the AI safety/alignment domain to control for topic-specific reviewing norms. Selection prioritized review quality: each paper has $\geq$3 substantive reviews, an unambiguous AC decision with articulated reasoning, and $\geq$2 extractable technical concerns. PDFs were sanitized to remove decision-revealing metadata. We also tracked which paper version each system reviewed and whether fixes to resolved concerns were visible in that version. Of the 170 concerns marked resolved by the AC, 40 (24%) had fixes absent from the reviewed PDF; these remain valid detection targets rather than calibration failures. Details appear in Appendix P. Systems receive the camera-ready PDF for accepted papers and the original submission for rejected papers; concern extraction reads the full OpenReview record including rebuttals and meta-reviews.

We evaluate six configurations (Table 1): **System L** (Liang et al., 2024b) (single-prompt zero-shot), **System A** (Lu et al., 2024) (iterative reflection), **System O** (ChicagoHAI, 2026) (progressive structured review), and **System M** (D'Arcy et al., 2024) (multi-agent swarm). Systems L, A, and O run on Claude Opus to control for model effects; L and A additionally run on GPT-4o. System M runs on GPT-4o only using its native OpenAI API (Opus runs produced degenerate output (repetitive or truncated reviews) under our SDK adaptation); its metrics reflect combined method and model effects (§4.3). Each configuration is run 3 times. All systems use published code with adaptations documented in Appendix D.

**Severity extraction.** Baselines expose severity very differently: System A provides structured weaknesses with numeric scores and an explicit decision; System L has a "reasons for rejection" section; System M labels concerns as critical/major/minor; System O provides no native severity labels. We therefore normalize severity and decisive flags with an extraction pass over the full review text, using structural cues, language intensity, and available scores. In human audits of 54 paper-method pairs, 51 extraction sheets were rated satisfactory for completeness and support, and none contained unsupported extracted concerns (§2.4). For Systems L, A, and O, the resulting severity labels are our normalized interpretation rather than native system outputs; System M is closest to native severity (Appendix F). Absolute FDR values should therefore be read as approximate, but the paper's main claims rest on replicated qualitative patterns such as flat severity profiles across accepted and rejected papers. Level 3–4 metrics thus evaluate the interaction between review generation and severity interpretation, not only the underlying generator (§6).

**Verdict extraction.** Evaluating verdict accuracy requires an accept/reject label for each review, but most systems do not emit one explicitly; their reviews carry signals of acceptance or rejection without a binary field. We therefore apply a verdict inference procedure that maps each free-form

Table 1: Systems analyzed. Concerns/paper is the average number of atomic, deduplicated concerns per paper (3-run mean across all 48 papers). GPT-4o single-agent systems produce roughly half as many concerns as their Opus counterparts.

| System | Model | Concerns/paper |
|---|---|---|
| System L | Opus | 10.6 |
| System A | Opus | 11.5 |
| System O | Opus | 8.3 |
| System L | GPT-4o | 5.2 |
| System A | GPT-4o | 4.8 |
| System M | GPT-4o | 10.1 |

Table 2: Detection and calibration can point in different directions. Across the Opus systems, the configuration with the strongest rejected-paper recall still shows weak accepted-paper calibration. Accepted acc. = accepted-paper verdict accuracy (pipeline inference; see Appendix U for sensitivity); false decisive rate and resolved-escalation are computed on accepted papers.

| Metric (Level) | L | A | O |
|---|---|---|---|
| *Detection (rejected)* | | | |
| Recall (L1) | .44 | .44 | .17 |
| *Calibration (accepted)* | | | |
| Accepted acc. (L2) | .028 | .083 | .347 |
| False dec. rate (L3) | .49 | .36 | .37 |
| Resolved-esc. (L3) | .63 | .60 | .62 |

review to an accept/reject label. Both System A configurations produce an explicit `Decision` field; the other four configurations do not. System L organizes output into "reasons for acceptance" and "reasons for rejection" sections, from which a verdict is inferred by the extraction pipeline. Systems O and M produce no native verdict or scores; we infer the verdict from overall review tone and the presence or absence of blocking-level language (see Appendix F for worked examples of each system's output structure). The extraction pipeline uses a default-REJECT rule for ambiguous cases, which may inflate the reject rate for systems that write analytically balanced reviews. Because verdict accuracy numbers for Systems L, O, and M reflect this inference procedure rather than native system decisions, we conducted a verdict inference audit with two independent raters and human adjudication (Appendix U). The audit shows that verdict-level findings are sensitive to the inference method, while the paper's concern-level diagnostics—recall, FDR, decisive precision, phantom rates, attention profiles, ICC, and top-$K$ analyses, all computed or stratified by official verdict—are unchanged by how accept/reject is inferred. We report pipeline-inferred verdicts throughout as the primary analysis, with sensitivity ranges from the audit noted where relevant.

## 4 What Concern Alignment Reveals

Each subsection presents a diagnostic finding invisible to the level below. Table 2 previews the core tension: at L1, System L (Opus) appears to be the best Opus system (highest recall); by L3, its calibration failures are exposed.

### 4.1 Binary Accuracy Masks Structural Differences

Most systems do not produce explicit accept/reject decisions (§3); we infer verdicts from review tone using a pipeline with a default-REJECT rule. Verdict-stratified accuracy under this pipeline (Appendix T, Table 27) shows that several configurations exhibit reject-heavy profiles, and that identical overall accuracy can conceal structurally different behaviors. A verdict inference audit with two independent methods and two independent raters (Appendix U, Table 28) confirms that these numbers are sensitive to the inference method: the model-effect swing ranges from 46 to 96 percentage points depending on who reads the review and what rules they apply. The sensitivity of verdict accuracy to the inference method is itself diagnostic: it confirms that verdict-level evaluation provides an unreliable measurement surface for systems not designed to produce explicit recommendations. The concern-level diagnostics used in the remainder of this section—recall, FDR, decisive precision, phantom rates, attention profiles, ICC, and top-$K$ analyses—are invariant to the verdict inference method because they are computed or stratified by official verdict rather than predicted verdict.

Concern-level evaluation also enables a stability diagnostic that verdict-level analysis cannot: the intraclass correlation coefficient (ICC), which measures whether the same paper receives similar scores across independent runs. On rejected papers, concern recall ICC(2,1) (a two-way random-effects model treating papers as subjects and runs as raters) ranges from 0.39 to 0.76 across configurations,

Table 3: Core concern-level metrics (3-run mean ± std). Rcl = concern recall on rejected papers; Dec. rcl = decisive-blocker recall on rejected papers; FDR = false decisive rate on accepted papers; Res. esc = resolved-escalation on accepted papers; DecP = decisive precision (strict) on rejected papers; PhDec = phantom decisive rate on rejected papers. †A (4o) has *higher* FDR despite fewer concerns.

| Sys. | Rcl (rej) | Dec. rcl (rej) | FDR (acc) | Res. esc (acc) | DecP (rej) | PhDec (rej) |
|---|---|---|---|---|---|---|
| L (Op) | .44±.02 | .68±.01 | .49±.03 | .63±.04 | .33±.02 | .14±.03 |
| A (Op) | .44±.01 | .65±.01 | .36±.02 | .60±.05 | .36±.02 | .11±.02 |
| O (Op) | .17±.01 | .22±.04 | .37±.03 | .62±.08 | .17±.01 | .26±.03 |
| L (4o) | .25±.01 | .26±.00 | .25±.02 | .61±.07 | .32±.07 | .08±.02 |
| A (4o)† | .22±.02 | .37±.07 | .55±.02 | .70±.11 | .36±.04 | .18±.04 |
| M (4o) | .27±.04 | .31±.03 | .10±.06 | .34±.06 | .18±.10 | .09±.04 |

exceeding verdict ICC for 5 of 6 systems.[2] For near-universal reject profiles under the pipeline (L (Opus), A (GPT-4o)), verdict ICC is near zero, yet concern recall ICC remains meaningful (0.69 and 0.41). Even when the verdict carries little per-paper information, concern-level analysis still extracts diagnostic signal (Appendix I).

On Paper D, all three Opus systems correctly reject (100% binary accuracy), yet System A catches all 3 AC content-related decisive blockers while System O catches none, focusing only on notation (a fourth blocker, concerning reviewer engagement during discussion, is excluded as invisible to PDF-only systems). Concern alignment reveals a quality gap that binary accuracy misses (Appendix J); throughout the case-study appendix we refer to papers by letter designation and keep identities for supplementary release.

### 4.2 Detection Without Discrimination

Table 3 and Figure 3 show the core calibration pattern. FDR on accepted papers ranges from 0.10 (System M) to 0.55 (System A (GPT-4o)). Most single-agent systems show essentially flat decisive behavior across verdicts: System L (Opus), for example, marks 49% of concerns as decisive on accepted papers and 52% on rejected ones. Under our AC-aligned operationalization, the accepted-paper decisive-blocker count is zero, so these systems are not only over-escalating; they are failing to modulate decision weight with paper quality. System M is the exception (10% accepted, 21% rejected), though part of that advantage reflects a low-decisive-flag profile rather than uniformly strong calibration. A long list can recover some official concerns through volume rather than selective attention. This is the volume-without-discrimination pattern: coverage that arises from sheer concern count rather than selective attention.

On rejected papers, where real decisive blockers exist, decisive precision (strict) and phantom decisive rate decompose the decisive signal (Table 3). Decisive precision ranges from 0.17 (System O) to 0.36 (System A (Opus)), meaning at best only 36% of decisive flags identify a concern the AC also treated as a decisive blocker. Phantom decisive rate ranges from 0.08 (System L (GPT-4o)) to 0.26 (System O): the fraction of total output that consists of fabricated decisive concerns. Systems L and A (Opus) achieve the best combination (DecPrec 0.33–0.36, PhDecRate 0.11–0.14), while System O produces the most noise (61% of its decisive flags are phantoms). System M shows low phantom volume (PhDecRate 0.09) but also low precision (0.18), consistent with its conservative decisive threshold that avoids false alarms at the cost of missing real blockers.

The calibration failure is visible at the level of individual concerns. On the accepted-paper example in Figure 1, the official review notes that the core evaluation metric "lacks formal validation of its estimator" (*moderate, resolved*); System L (Opus) frames the same issue as the estimator "does not approximate the theoretical quantity it claims to measure" (*fatal, decisive*). The concern family is similar, but the severity

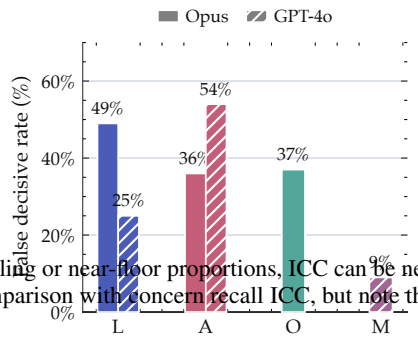


Figure 3: False decisive rate on accepted papers. Most single-agent systems mark 25–55% of concerns decisive under the AC-aligned operationalization.

[2]ICC on binary verdicts has known limitations: with near-ceiling or near-floor proportions, ICC can be near zero despite high raw agreement. We report it to enable the comparison with concern recall ICC, but note that Fleiss' $\kappa$ would be an alternative for the binary case.

Table 4: Model choice with fixed prompts produces large shifts. A (GPT-4o) has *higher* false decisive rate despite fewer concerns; in the column headers, "4o" abbreviates GPT-4o. Acc-corr reflects pipeline inference; see Appendix U for sensitivity.

| Metric | L Opus | L 4o | A Opus | A 4o |
|---|---|---|---|---|
| Conc./paper | 10.6 | 5.2 | 11.5 | 4.8 |
| Acc-corr. | 2.8% | 63.9% | 8.3% | 4.2% |
| Recall (rej) | .44 | .25 | .44 | .22 |
| Dec. rcl (rej) | .68 | .26 | .65 | .37 |
| FDR (acc) | .49 | .25 | .36 | .55 |

Table 5: Recall by AC treatment on rejected papers (L4). Dec. = decisive blocker, Unres. = unresolved, Res. = resolved, and Gap is decisive-blocker recall minus resolved-concern recall. Negative gaps (**bold**) indicate inverted attention.

| System | Dec. | Unres. | Res. | Gap |
|---|---|---|---|---|
| L (Opus) | .68 | .47 | .48 | +20pp |
| A (Opus) | .66 | .44 | .49 | +17pp |
| O (Opus) | .18 | .20 | .20 | −2pp |
| L (GPT-4o) | .24 | .29 | .33 | −9pp |
| A (GPT-4o) | .37 | .34 | .25 | +12pp |
| M (GPT-4o) | .29 | .33 | .38 | −9pp |

and decision weight are not. Concern recall rewards the match; FDR penalizes the escalation. A verdict-only metric would prefer the more extreme review in this example because it ignores the severity error.

Top-$K$ analysis (Appendix K) reveals a second pattern: some low-FDR systems achieve their aggregate rate partly through *concern dilution* (producing many low-severity concerns that reduce the decisive fraction without improving calibration) rather than selective prioritization. System M's FDR rises from 0.10 to 0.28 at $K$=3. Systems L and A (Opus) show the opposite pattern: at $K$=3, 90% or more of top concerns are marked decisive (FDR $\geq$ 0.90). Full $K$-curves and exact values appear in Appendix K.

Top-$K$ is useful because reviews are consumed as prioritized concern lists. A researcher triaging feedback will typically focus on the top few concerns first. Full-review and top-$K$ metrics therefore answer complementary questions: what the system says in total, and what it tells the reader to focus on first.

### 4.3 Model vs. Method Effects

Table 4 presents Systems L and A on both Opus and GPT-4o. These paired comparisons do not estimate the relative size of model and method effects, but they make a known concern concrete: model choice alone materially shifts both verdict bias and concern calibration, and the framework quantifies which diagnostic dimensions change. Under the pipeline, System L flips from reject-heavy on Opus to majority-accept on GPT-4o; across alternative inference methods, the accepted-paper-accuracy swing ranges from 46 to 96 percentage points (Table 28).[3]

The effect is not uniform: System A (GPT-4o) stays at 4.2% accepted accuracy while achieving *higher* FDR (0.55) than its Opus counterpart (0.36), despite generating half as many concerns. On Paper H (rejected), System L (Opus) catches all three AC decisive blockers in every run; System L (GPT-4o) misses all three and misdiagnoses a substance problem as a clarity problem (Appendix J). Published comparisons that use different models per system therefore interleave method and model effects; concern alignment makes that entanglement visible and quantifiable.

### 4.4 Attention Profiles and Decision-Weight Calibration

Level 4 decomposes recall by AC treatment (Table 5), asking whether systems preferentially recover the concerns that remained consequential after discussion and rebuttal rather than the ones the AC treated as resolved or non-blocking. Systems L and A (Opus) show positive gaps: recall on decisive blockers exceeds recall on resolved concerns by 17–20 percentage points. Systems O, L (GPT-4o),

[3]Three inference methods (pipeline with default-REJECT, independent tone reading, concern-gate-based rules) applied by two independent raters ($\kappa \geq 0.74$), plus human adjudication of 54 disagreement cases. One rater's tone reading assigns L (Opus) 42% accepted-paper accuracy, an outlier against all other method/rater combinations ($\leq$ 4%); see Appendix U.

and M show flat or inverted profiles, meaning they recover resolved concerns at least as readily as decisive blockers (Figure 4).

Overall recall hides this structure: System L (Opus) has 0.44 overall recall, but 0.68 on decisive blockers versus 0.48 on resolved concerns. The gap varies from +20pp to −9pp across systems. This over-escalation appears concretely: on Paper B, an official reviewer hedges that scaffolding "may be unfamiliar to some models" (marked `resolved` after rebuttal); System L (Opus) rephrases the same issue as a **fundamental** confound and treats it as decisive. The issue is detected, but its decision weight is inflated. Resolved-escalation rate measures that failure directly.

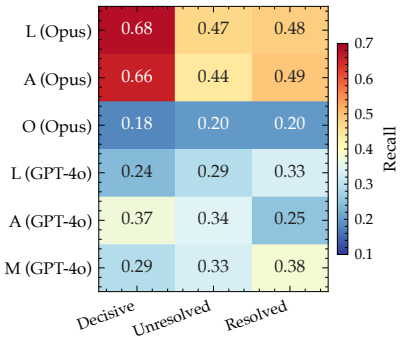


Figure 4: Recall by AC treatment on rejected papers. Positive gaps indicate greater attention to decisive blockers than to resolved concerns.

### 4.5 Phantoms and Decision Impact

Phantom rates on accepted papers range from 42% to 83%. A dual-annotated audit ($N$=200) suggests many phantoms are legitimate concerns officials did not raise, though transfer from that audit system to Systems L/A/O/M is only suggestive. On Paper C (accepted spotlight), System O raises a **fatal** phantom with a fabricated counterexample; four expert reviewers found no such error, whereas System L's phantoms on the same paper are mostly extra but plausible feedback. *Harmful phantom rate* (fatal/major phantoms on accepted papers) separates damaging hallucinations and over-escalations from benign extra feedback.

We therefore decompose output into decision-relevant and decision-harmful components instead of reporting a single net score. On rejected papers, harmful components exceed relevant ones across configurations (0.21–0.68 vs. 0.04–0.12; derived from the match graphs underlying Appendix B). On accepted papers the mix is more heterogeneous, which makes the decomposition useful: it separates harmful phantoms, resolved-concern re-escalation, and missed decisive issues.

Paper F illustrates the point: System A achieves 64% recall with 25% phantom rate yet still rejects the paper because it escalates a minor presentation issue and downrates a major methodological one.

## 5 Related Work

Recent work on AI peer review spans direct prompting, iterative reflection, structured review pipelines, multi-agent reviewers, fine-tuned reviewer models, and pairwise-comparison reformulations (Liang et al., 2024b; Lu et al., 2024; D'Arcy et al., 2024; ChicagoHAI, 2026; Gao et al., 2025; Idahl & Ahmadi, 2025; Zou et al., 2026; Zhang et al., 2025b). The closest evaluation papers compare facet distributions (Shin et al., 2025), score holistic review quality (Garg et al., 2025), reconstruct premises to test whether review claims are misinformed (Ryu et al., 2025), or benchmark weakness and limitation discovery (Xu et al., 2025; Lou et al., 2025). Concern alignment instead evaluates free-form concern instances, models both misses and phantoms, measures decision weight as well as detection, and grounds evaluation in post-rebuttal AC treatment; Appendix S gives the extended discussion.

## 6 Discussion and Limitations

**Scope.** Our pilot covers 48 papers in one domain and four systems. That is enough to expose failure modes hidden by verdict accuracy, but not enough for population-level claims or fine-grained rankings.

**Operational anchor.** The framework evaluates review quality as feedback that helps researchers improve papers, not as a way to optimize for conference outcomes. We use AC decisions as a noisy post-deliberation anchor for which concerns were treated as consequential. That gives the metrics a

concrete reference point, but it does not imply that accepted papers are free of serious weaknesses or that rejected papers were clearly below the bar (Cortes & Lawrence, 2021). False decisive rate on accepted papers is therefore defined relative to the official review rationale. Concern-level analysis is specifically designed to be more robust to this noise than verdict agreement: by decomposing evaluation into inspectable concern units, many of which are independently verifiable regardless of the final decision, the framework extracts diagnostic signal even when verdicts are unreliable. The structural patterns it identifies (flat severity profiles, inverted attention, concern dilution) are robust across 48 papers in a way that individual verdict disagreements are not.

**Verdict inference.** Most systems we evaluated do not produce native accept/reject decisions; we infer verdicts from review tone using a separate extraction pipeline (§3). A two-rater audit with human adjudication (Appendix U) shows that verdict-level findings are sensitive to the inference method, whereas the concern-level diagnostics used throughout the paper—recall, FDR, decisive precision, phantom rates, attention profiles, ICC, and top-$K$ analyses—are unaffected because they are computed or stratified by official verdict rather than predicted verdict. Across audited methods, the L Opus $\rightarrow$ L GPT-4o accepted-paper-accuracy swing ranges from 46 to 96 percentage points, with the pipeline estimate lying inside that range (Table 28). When systems do not emit explicit recommendations, verdict agreement becomes an unstable measurement target. Separately, this pilot suggests that verdict readability and severity calibration can co-vary, but the relationship is not monotone across the six configurations. Some hard-to-read reviews also show flatter severity profiles, but other configurations break that pattern. We therefore treat verdict ambiguity and calibration as related but distinct observations, not as evidence of a general association or mechanism.

**Measurement validity.** Match graphs are constructed with LLM assistance, so circularity remains a methodological risk. Cross-model validation with GPT-5.4 Pro against a Claude-based primary annotator, human adjudication, and extraction audits reduce that risk but do not remove shared biases. For Systems L, A, and O, severity is inferred by our extraction pipeline rather than emitted natively, so Level 3–4 metrics partly evaluate that inference step. We also avoid penalizing a concern as re-escalated when discussion resolved it but the fix is not visible in the PDF the system reviewed. System M was run only on GPT-4o, so its profile remains a combined model-plus-method effect; with only 3 runs, we emphasize replicated qualitative patterns over fine point estimates.

## 7 Conclusion

Binary accuracy cannot tell us *how* a system fails or *what to fix*. Concern alignment does so by evaluating the unit of review that authors and readers actually act on.

Across the pilot, each level of the ladder exposed a different failure mode: verdict stratification revealed reject-heavy behavior; decision-aware metrics quantified severity miscalibration; rebuttal-aware recall exposed inverted attention; and top-$K$ analysis separated calibrated prioritization from concern dilution.

The clearest lesson is that finding issues is not enough. The systems we examined often detect real concerns yet still misjudge their decision weight, especially on accepted papers. Concern-level diagnostics may therefore be useful beyond peer review for evaluative AI systems that generate critiques, due-diligence reports, or risk assessments. We view concern alignment as an evaluation substrate: a way to audit which concerns systems identify, how they weight them, and whether those priorities align with the review rationale that informed the final assessment.

## References

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

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, 2020.

Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.569. URL https://aclanthology.org/2020.emnlp-main.569/.

ChicagoHAI. OpenAIReview: AI-powered academic paper reviewer, 2026. URL https://openaireview.org/. Project website and software repository; built by Chicago Human+AI Lab.

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

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. MARG: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*, 2024.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. ReviewAgents: Bridging the gap between human and AI-generated paper reviews. *arXiv preprint arXiv:2503.08506*, 2025.

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, 2019. Association for Computational Linguistics. URL https://aclanthology.org/N19-1129/.

Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Chhavi Kirtani, Murari Mandal, and Dhruv Kumar. ReviewEval: An evaluation framework for AI-generated reviews. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20542–20564, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1120. URL https://aclanthology.org/2025.findings-emnlp.1120/.

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, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-demo.44. URL https://aclanthology.org/2025.naacl-demo.44/.

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, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.89. URL https://aclanthology.org/2022.naacl-main.89/.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, Daniel A. McFarland, and James Y. Zou. Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 29575–29620. PMLR, 2024a. URL https://proceedings.mlr.press/v235/liang24b.html.

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):AIoa2400196, 2024b. doi: 10.1056/AIoa2400196. URL https://doi.org/10.1056/AIoa2400196.

Renze Lou, Hanzi Xu, Sijia Wang, Jiangshu Du, Ryo Kamoi, Xiaoxin Lu, Jian Xie, Yuxuan Sun, Yusen Zhang, Jihyun Janice Ahn, Hongchao Fang, Zhuoyang Zou, Wenchao Ma, Xi Li, Kai Zhang, Congying Xia, Lifu Huang, and Wenpeng Yin. AAAR-1.0: Assessing AI's potential to assist research. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40361–40383. PMLR, 2025. URL https://proceedings.mlr.press/v267/lou25c.html.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Qianli Ma, Chang Guo, Zhiheng Tian, Siyu Wang, Jipeng Xiao, Yuanhao Yue, and Zhipeng Zhang. Paper2Rebuttal: A multi-agent framework for transparent author response assistance. *arXiv preprint arXiv:2601.14171*, 2026.

Giuseppe Russo, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28, 2025. doi: 10.1145/3757667. URL https://doi.org/10.1145/3757667.

Hyun Ryu, Doohyuk Jang, Hyemin S. Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, Haechan Kim, Jina Kim, Joowon Kim, Yoonjeon Kim, Kwanhyung Lee, Chanjae Park, Heecheol Yun, Gregor Betz, and Eunho Yang. ReviewScore: Misinformed peer review detection with large language models. *arXiv preprint arXiv:2509.21679*, 2025. doi: 10.48550/arXiv.2509.21679. URL https://arxiv.org/abs/2509.21679.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. Mind the blind spots: A focus-level evaluation framework for LLM reviews. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35630–35656, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.1805. URL https://aclanthology.org/2025.emnlp-main.1805/.

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*, 8:326–336, 2026. doi: 10.1038/s42256-026-01188-x. URL https://doi.org/10.1038/s42256-026-01188-x.

Sihong Wu, Yiling Ma, Yilun Zhao, Tiansheng Hu, Owen Jiang, Manasi Patwardhan, and Arman Cohan. RbtAct: Rebuttal as supervision for actionable review feedback generation. *arXiv preprint arXiv:2603.09723*, 2026.

Zhijian Xu, Yilun Zhao, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. Can LLMs identify critical limitations within scientific research? a systematic evaluation on AI research papers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 20652–20706, Vienna, Austria, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1009. URL https://aclanthology.org/2025.acl-long.1009/.

Daoze Zhang, Zhijian Bao, Sihang Du, Zhiyi Zhao, Kuangling Zhang, Dezheng Bao, and Yang Yang. $Re^2$: A consistency-ensured dataset for full-stage peer review and multi-turn rebuttal discussions. *arXiv preprint arXiv:2505.07920*, 2025a.

Yaohui Zhang, Haijing Zhang, Wenlong Ji, Tianyu Hua, Nick Haber, Hancheng Cao, and Weixin Liang. From replication to redesign: Exploring pairwise comparisons for LLM-based peer review. In *Advances in Neural Information Processing Systems*, 2025b. URL https://openreview.net/forum?id=z5KTxW5sJd.

Zhuoyang Zou, Abolfazl Ansari, Delvin Ce Zhang, Dongwon Lee, and Wenpeng Yin. DIAGPaper: Diagnosing valid and specific weaknesses in scientific papers via multi-agent reasoning. *arXiv preprint arXiv:2601.07611*, 2026.

# A Detailed Metric Definitions

We restate the notation here for standalone readability. For a paper $p$, let $\mathcal{O}_p = \{o_1, \ldots, o_m\}$ denote the set of official concerns and $\mathcal{A}_p = \{a_1, \ldots, a_n\}$ the set of agentic concerns. Let $E_{\text{strict}}(p) \subseteq \mathcal{O}_p \times \mathcal{A}_p$ be the set of `exact` and `partial` matches only. Let $v_p$ and $\hat{v}_p$ denote the official and agentic verdicts. Let $t(o_i)$ denote the AC-treatment label of official concern $o_i$, let $s(a_j)$ denote the severity of agentic concern $a_j$, and let $\mathrm{dec}(a_j) \in \{0, 1\}$ denote whether $a_j$ is marked decisive. When metrics are stratified by decision, $\mathcal{P}_{\text{acc}}$ and $\mathcal{P}_{\text{rej}}$ denote the accepted and rejected paper subsets, and $\mathcal{A}_{\text{acc}} = \bigcup_{p \in \mathcal{P}_{\text{acc}}} \mathcal{A}_p$ denotes the multiset of all agentic concerns on accepted papers.

**Level 0: Binary accuracy.**

$$\text{Acc} = \frac{1}{N} \sum_{p=1}^{N} \mathbf{1}[\hat{v}_p = v_p] \tag{1}$$

**Level 1: Concern recall (strict).** For a single paper,

$$\text{Recall}(p) = \frac{|\{o_i \in \mathcal{O}_p : \exists\, e_{ij} \in E_{\text{strict}}(p)\}|}{|\mathcal{O}_p|} \tag{2}$$

Reported system-level recall is the mean of $\text{Recall}(p)$ over the relevant paper subset.

**Level 1: Phantom rate.** For a single paper,

$$\text{Phantom}(p) = \frac{|\{a_j \in \mathcal{A}_p : \nexists\, e_{ij} \in E_{\text{strict}}(p)\}|}{|\mathcal{A}_p|} \tag{3}$$

Reported system-level phantom rate is again the mean over papers.

**Level 2: Verdict-stratified metrics.** All Level 0–1 metrics can be computed separately on $\mathcal{P}_{\text{acc}}$ and $\mathcal{P}_{\text{rej}}$.

**Level 3: False decisive rate.**

$$\text{FDR}_{\text{acc}} = \frac{\sum_{a_j \in \mathcal{A}_{\text{acc}}} \mathrm{dec}(a_j) - |\mathcal{E}|}{|\mathcal{A}_{\text{acc}}|} \tag{4}$$

That is, among all agentic concerns raised on accepted papers, what fraction carry a decisive flag after excusing flags that correctly detect resolved-but-unfixed issues? Here $\mathcal{E}$ is the set of decisive flags excused because they match a resolved official concern whose fix is absent from the reviewed PDF (correct detection, not a false alarm).

**Level 3: Decisive precision and phantom decisive rate (rejected papers).** On rejected papers, some decisive flags correctly identify real blockers, so FDR is not meaningful. Two metrics replace it:

$$\text{DecPrec}_{\text{strict}} = \frac{|\{a_j \in \mathcal{A}_{\text{rej}} : \mathrm{dec}(a_j) = 1,\ \exists\, e_{ij} \in E_{\text{strict}} \text{ with } t(o_i) = \texttt{decisive\_blocker}\}|}{|\{a_j \in \mathcal{A}_{\text{rej}} : \mathrm{dec}(a_j) = 1\}|} \tag{5}$$

That is, among agentic concerns flagged decisive on rejected papers, what fraction matches an official concern that the AC also treated as a decisive blocker?

$$\text{PhDecRate} = \frac{|\{a_j \in \mathcal{A}_{\text{rej}} : \mathrm{dec}(a_j) = 1,\ \nexists\, e_{ij} \in E_{\text{strict}}\}|}{|\mathcal{A}_{\text{rej}}|} \tag{6}$$

That is, what fraction of total agentic output on rejected papers consists of decisive flags with no official match (fabricated blockers)?

**Level 3: Resolved-escalation rate.** Let $\mathrm{pdf}(o_i) \in \{0, 1\}$ indicate whether a concern marked resolved by the AC is actually addressed in the reviewed PDF (`addressed_in_pdf`). We restrict to concerns with $t(o_i) = \texttt{resolved}$ and $\mathrm{pdf}(o_i) = 1$; concerns resolved only in rebuttal text remain valid detection targets rather than calibration failures.

$$\text{ResEsc} = \frac{|\{(o_i, a_j) \in E_{\text{strict}}(p) : t(o_i) = \texttt{resolved},\ \mathrm{pdf}(o_i) = 1,\ s(a_j) \in \{\texttt{fatal}, \texttt{major}\}\}|}{|\{(o_i, a_j) \in E_{\text{strict}}(p) : t(o_i) = \texttt{resolved},\ \mathrm{pdf}(o_i) = 1\}|} \tag{7}$$

Table 6: Full verdict-stratified metrics (3-run means). Acc = accepted papers, Rej = rejected papers. DecPrec = decisive precision (strict), PhDec = phantom decisive rate (both on rejected papers only). Per-run bootstrap 95% CIs (10,000 paper-level resamples) available in supplementary materials.

| | **Recall** | | **FDR** | **DecPrec** | **PhDec** | **Phantom** | | **Res.-esc.** | **Concerns** |
|---|---|---|---|---|---|---|---|---|---|
| **System** | Acc | Rej | (acc) | (rej) | (rej) | Acc | Rej | (acc) | /paper |
| L (Opus) | .37 | .44 | .49 | .33 | .14 | .49 | .34 | .63 | 10.6 |
| A (Opus) | .42 | .44 | .36 | .36 | .11 | .51 | .46 | .60 | 11.5 |
| O (Opus) | .09 | .17 | .37 | .17 | .26 | .83 | .72 | .62 | 8.3 |
| L (GPT-4o) | .23 | .25 | .25 | .32 | .08 | .43 | .37 | .61 | 5.2 |
| A (GPT-4o) | .21 | .22 | .55 | .36 | .18 | .42 | .43 | .70 | 4.8 |
| M (GPT-4o) | .31 | .27 | .10 | .18 | .09 | .64 | .57 | .34 | 10.1 |

**Level 3: Decision-relevant and decision-harmful decomposition.** Let $R_p$ be the set of decision-relevant agentic concerns for paper $p$ (strict matches to official concerns at appropriate severity on rejected papers, or constructive non-blocking feedback on accepted papers), and let $H_p$ be the set of decision-harmful concerns. On accepted papers, harmful concerns include re-escalation of dismissed or resolved concerns (the latter only when the fix appears in the reviewed PDF) to fatal/major severity, and harmful phantoms (unmatched fatal/major concerns). On rejected papers, harmful concerns include severity under-rating of fatal/major official concerns to minor/moderate and missed decisive blockers. We report the decision-relevant rate ($|R_p|/|\mathcal{A}_p|$) and decision-harmful rate ($|H_p|/|\mathcal{A}_p|$) separately because the components reveal *where* harm originates.

**Level 4: Recall by AC treatment.** For each treatment category $t \in T$ (where $T$ is the set of AC-treatment categories) and paper $p$,

$$\text{Recall}_t(p) = \frac{|\{o_i \in \mathcal{O}_p : t(o_i) = t,\ \exists\, e_{ij} \in E_{\text{strict}}(p)\}|}{|\{o_i \in \mathcal{O}_p : t(o_i) = t\}|} \tag{8}$$

Reported treatment-specific recall is the mean of $\text{Recall}_t(p)$ over the relevant paper subset.

**Volume-normalized variants (top-$K$).** Any Level 3–4 metric $M$ can be computed at top-$K$ by restricting to the $K$ most severe agentic concerns per paper (ranked by severity, with ties broken by decisive flag):

$$M@K = M\big(\mathcal{A}_p^{(K)}\big) \quad \text{where} \quad \mathcal{A}_p^{(K)} = \text{top-}K(\mathcal{A}_p, \text{severity}) \tag{9}$$

This normalizes for review length: systems generating 4.8 to 11.5 concerns per paper are compared at equal concern budget. FDR@$K$ is especially diagnostic because it tests whether a system's *most severe* concerns are calibrated, not just whether the overall proportion of decisive flags is low.

## B Per-System Detailed Metrics

Table 6 presents the complete set of verdict-stratified metrics for all system configurations.

Table 7 presents per-run bootstrap 95% confidence intervals for key metrics. These intervals reflect within-run sampling uncertainty (bootstrapping over papers within a single run, 10,000 resamples). The cross-run standard deviations in Table 3 capture a different source of variation (stochastic model output across independent runs).

## C Per-System Diagnostic Profiles

This appendix summarizes the main diagnostic profile of each configuration in neutral terms.

**System L (Opus).** Tied for highest concern recall among the Opus systems (44% overall; 68% on decisive blockers), but weak calibration on accepted papers (FDR 0.49; resolved-escalation 0.58–0.66 depending on run). It produces about 11 concerns per paper with little modulation by paper verdict. The main weakness is severity assignment, not detection.

Table 7: Bootstrap 95% CIs for key metrics (range across 3 runs, 10,000 paper-level resamples per run). Intervals confirm that within-run estimates are stable: FDR is well-separated from zero for all single-agent systems, and decisive recall intervals are consistent across runs for Opus-based systems.

| System | FDR (acc) | Dec. recall (rej) | Recall (rej) | Res.-esc. (acc) |
|---|---|---|---|---|
| L (Opus) | [.39, .61] | [.56, .80] | [.36, .51] | [.43, .80] |
| A (Opus) | [.30, .44] | [.52, .77] | [.38, .51] | [.39, .81] |
| O (Opus) | [.22, .47] | [.10, .35] | [.11, .24] | [.31, 1.0] |
| L (GPT-4o) | [.13, .40] | [.15, .38] | [.20, .29] | [.32, .94] |
| A (GPT-4o) | [.44, .65] | [.19, .59] | [.16, .29] | [.27, 1.0] |
| M (GPT-4o) | [.00, .23] | [.16, .48] | [.18, .36] | [.13, .57] |

**System A (Opus).** Similar overall recall to System L (Opus), with somewhat lower FDR (0.36) and strong performance on technical blockers in the case studies. Its dominant pattern is still reject-heavy and verdict-insensitive. The main improvement target is contribution-versus-blocker calibration.

**System O (Opus).** Lowest recall on accepted papers (9%; 17% on rejected) and highest phantom rate (83% on accepted papers). The extracted concerns concentrate on notation, formalization, and local correctness criteria; this configuration performs noticeably better when the official concerns are themselves mathematical or formal. The main limitation is a narrow concern scope.

**System L (GPT-4o).** Unlike the Opus single-agent runs, this configuration accepts a substantial fraction of accepted papers (63.9% accepted-paper accuracy). Its concerns on accepted papers are often constructive, but on rejected papers it misses many decisive blockers and sometimes reframes substantive problems as presentation issues. The main limitation is depth on rejected papers.

**System A (GPT-4o).** Produces roughly half as many concerns as System A (Opus) but a higher FDR (0.55 versus 0.36), showing that fewer concerns do not by themselves imply better calibration. The main issue is an aggressive decisive threshold despite sparse output.

**System M (GPT-4o).** This configuration has the lowest full-review FDR (0.10) and the highest accepted-paper accuracy (79.2%), but top-$K$ analysis (§4.2) shows that part of the low FDR reflects a low-decisive-flag profile: at $K$=5, FDR rises to 0.21 while decisive recall is 22%. It also shows the widest run-to-run variance ($\pm$19% accepted-paper accuracy). Because System M was evaluated only on GPT-4o, these observations should be treated as a combined model+method profile rather than a clean architectural effect. A verdict inference audit (Appendix U) found that all 48 System M reviews contain multi-agent coordination artifacts (e.g., inter-agent messages, repeated draft fragments) that make verdict inference unreliable regardless of method; accepted-paper accuracy varies widely across inference methods (Table 28).

## D Baseline Implementation Details

All systems were run starting from their official open-source implementations. We clone or install each original repository and load review-generation prompts verbatim. Beyond the API transport layer, we made a small number of implementation adaptations (Table 8); none modifies the review-scoring prompts that define each system's evaluation logic.

**Source repositories.** Systems L (Liang et al., 2024b) and M (D'Arcy et al., 2024) both use prompts from the MARG repository[4] (which includes Liang et al.'s baseline as a configuration). System A (Lu et al., 2024) uses the AI Scientist repository.[5] System O (ChicagoHAI, 2026) uses the `openaireview` pip package (v0.2.7).[6]

[4] https://github.com/allenai/marg-reviewer
[5] https://github.com/SakanaAI/AI-Scientist
[6] https://github.com/ChicagoHAI/OpenAIReview

**Adaptation for Claude Opus.** Systems L, A, and O were run on Claude Opus via an SDK adapter that replaces the API client call. The adapter preserves all review-generation prompts and orchestration logic (e.g., AI Scientist's iterative reflection loop), but the SDK manages temperature and `max_tokens` internally, so these parameters are not directly controllable on the Claude path. GPT-4o runs use the native OpenAI API as published. System O natively targets the Anthropic API (the package defaults to Claude Opus), so only authentication routing is adapted; we use the `progressive` method (their recommended default). System M runs on GPT-4o only using the native OpenAI API (Opus runs produced degenerate output (repetitive or truncated reviews) under our SDK adaptation).

Table 8: Implementation adaptations beyond API transport. *Prompt* = review-generation prompts (unchanged for all systems); *Config* = non-prompt parameters; *Infra* = infrastructure-level differences.

| System | Type | Adaptation |
|---|---|---|
| L (Claude) | Config | System message changed from "ChatGPT" to "helpful AI assistant" |
| L (Claude) | Infra | PDF extraction via pymupdf (original: Grobid/S2ORC structured XML) |
| A | Config | Few-shot examples disabled; `max_tokens` raised $4096 \rightarrow 8192$ |
| All (Claude) | Infra | Temperature, `max_tokens` SDK-managed; not directly controllable |

# E Operational Implementation Protocols

The concern-alignment pipeline is implemented as eight versioned protocols. They are not free-form prompts in the sense of open-ended generation; each protocol specifies a bounded artifact transformation with explicit inputs, outputs, and validation rules. Table 9 summarizes the pipeline, and the remainder of this section states the operational logic in paper style.

## E.1 Official concern extraction and revision-aware grounding (Steps 1–2)

**Inputs and outputs.** The official extractor reads the rendered OpenReview forum PDF together with the paper PDF actually shown to the baseline reviewer. Its output is an official concern sheet containing atomic concern IDs (`01`, `02`, ...), normalized concern text, short evidence quotes, reviewer provenance, severity, AC treatment, decisive flags, and critical references. The revision-aware grounding step extends the base sheet with fields recording whether the reviewed PDF appears revised, whether each concern is actually addressed in that PDF (`addressed_in_pdf`), and the local evidence used to justify that judgment.

**Extraction procedure.** The extractor anchors first on the meta-review because the AC decision rationale determines the treatment labels used later in the ladder. It records positive and negative decision drivers, identifies any AC-explicit decisive blockers, and then decomposes reviewer comments into *single-issue* concerns rather than paragraph-length bundles. If multiple reviewers raise the same issue, the protocol merges them into one official concern unit while preserving provenance. Severity is assigned in a pre-rebuttal frame using reviewer language when available: `fatal` for validity-threatening flaws, `major` for likely blockers, `moderate` for meaningful but usually non-blocking weaknesses, and `minor` for polish-level issues.

**AC treatment coding.** Each official concern receives one of seven post-rebuttal treatment labels: `decisive_blocker`, `unresolved`, `resolved`, `accepted_limitation`, `dismissed`, `reframed_feature`, or `not_mentioned`. This label is the core supervision signal for Level 3–4 metrics. The protocol explicitly distinguishes a concern that remains real but non-blocking (`accepted_limitation`) from one the AC dismisses, and it separately records whether the concern was decisive for the final decision. When reviewers cite specific prior work as central to novelty or comparison judgments, those papers are stored as *critical references* with a role label such as `missing_comparison` or `novelty_precedent`.

Table 9: Versioned implementation protocols used in the concern-alignment pipeline.

| Step | Purpose | Main inputs | Main output |
|---|---|---|---|
| 1 | Official concern extraction | OpenReview forum PDF, meta-review | Atomic official concern sheet |
| 2 | Revision-aware grounding | Step 1 output + reviewed paper PDF | Official concern sheet with PDF-state fields |
| 3 | Independent extraction QC | Official concern sheet + source PDFs | Pass / flag / fail QC report |
| 4 | Agentic concern extraction | Review artifacts (text, structured scores, debate transcripts) | Atomic agentic concern sheet |
| 5 | Match-graph construction | Official + agentic concern sheets | Concern match graph |
| 6 | Aggregate analysis | Corpus of sheets and match graphs | Cross-paper metrics and intervention proposals |
| 7 | Audit worksheet generation | Match graphs and source evidence | Edge-verification worksheets |
| 8 | Semantic verification | Audit worksheets | Structured override file for graph correction |

**PDF cross-verification.** The revision-aware grounding step checks every extracted concern against the reviewed PDF. The extractor looks for added experiments, clarifications, tables, or textual changes that directly address the concern and records specific evidence when found. The governing rule is conservative: author claims alone are insufficient to mark a concern `resolved`; resolution requires reviewer or AC confirmation, or a clearly visible fix in the reviewed PDF. This rule prevents rebuttal promises from being counted as completed revisions when the system under evaluation never saw the revised content.

### E.2 Independent quality control for official sheets (Step 3)

The QC step is a cold read by a second agent that did not participate in extraction. It checks structural consistency (e.g., whether every concern cited as a decisive negative driver is also labeled `decisive_blocker`), scans for resolution-field contradictions, and flags implausible severity/treatment combinations such as a `minor` concern marked decisive. It then performs targeted hallucination checks by tracing several fatal or major concerns back to the source reviews, a completeness check over reviewer weakness sections and the meta-review, and spot checks of both `addressed_in_pdf=true` and `addressed_in_pdf=false` cases. Each sheet receives an overall QC verdict: `pass`, `pass_with_flags`, or `fail`.

### E.3 Agentic concern extraction (Step 4)

**Inputs and outputs.** The agentic extractor reads the full output of one system run on one paper. Core inputs are the verdict summary, main review text, adversarial brief, gate-check results, and scorecard; panel-style methods may also provide per-role reviews (champion, skeptic) and debate transcripts. The output is an agentic concern sheet containing atomic concerns (`A1`, `A2`, ...), normalized severity, decisive flags, decision drivers, and positive mentions that were noticed but not weighted decisively.

**Extraction logic.** The extractor first reads the verdict summary to recover context, scores, and explicit decisive reasons. It then parses the review artifacts section by section: major concerns in the main review become candidate issues, adversarial-brief dispositions are converted into concerns only when the brief accepts them as live issues, gate failures are mapped to fatal concerns, gate cautions to major concerns, and low scorecard dimensions yield additional concerns when the written rationale identifies a specific defect. Negative observations belong in the concern sheet; positive observations that were noticed but not treated as decisive are stored separately as positive mentions, which is critical for diagnosing "seen but not weighted" failures on accepted papers.

**Normalization and provenance.** Concern candidates are deduplicated across artifacts so the sheet represents issue units rather than repeated surface forms. For merged issues, the protocol keeps the highest-severity instance while preserving source provenance. Severity is normalized into three

fields: level (fatal / major / moderate / minor / unknown), addressability (unresolved / addressable / unknown), and mechanism (e.g., gate failure, binding rule, score threshold, debate). Panel methods additionally preserve origin provenance such as `champion` or `skeptic`, so downstream analysis can trace which internal reviewer introduced a concern.

### E.4 Concern match-graph construction (Step 5)

**Canonicalization and candidate generation.** For each official and agentic concern, the matcher writes a one-sentence canonical issue statement that abstracts away from phrasing while preserving the underlying defect. Candidate matches are proposed in both directions (official-to-agentic and agentic-to-official) to reduce asymmetries. The protocol prefers one-to-one matching and caps each concern at two edges; if more edges seem necessary, the concern is deemed too broad and should be split.

**Scope-based match typing.** Each candidate edge is classified as `exact`, `partial`, `related`, or `none` by a scope test: would fixing one concern necessarily fix the other? `Exact` requires the same specific defect and the same satisfaction condition. `Partial` captures same-family issues with different scope, abstraction level, or thresholds of satisfaction. `Related` documents near-misses that are topically nearby but should not count for strict metrics. The instructions explicitly warn against false matches driven only by shared tags or broad topical overlap, including prior-work characterization versus novelty, evaluation scope versus evaluation methodology, writing quality versus overclaiming, and the audit-derived *scope inflation* failure in which one concern bundles the other's complaint together with additional independent demands.

**Alignment labels and unmatched sets.** For every strict edge, the matcher also records judgment alignment (aligned / inverted / mixed) and severity alignment. Aggregate metrics use the production hybrid severity policy: fatal requires exact agreement, whereas one-level gaps among non-fatal concerns count as matches and larger gaps are labeled `under` or `over`. On accepted papers, the matcher separately aligns positive decision drivers so the framework can ask not only whether the system avoided fatal complaints, but also whether it captured why the paper deserved acceptance. Concerns with only `related` edges still appear in the unmatched lists; `related` is a near-miss annotation, not a strict match.

### E.5 Aggregate alignment analysis (Step 6)

The aggregate analysis step operates over a corpus of official sheets, agentic sheets, and match graphs. Before any metric is computed, it runs a lint gate that checks unmatched-list consistency, illegal severity labels on `related` edges, edge-cap violations, and other schema errors. It then derives severity calibration directly from raw severity levels rather than trusting hand-written edge labels, computes observability-aware positive-factor recall for accepted papers, flattens the graphs into a paper-level edge table, and produces verdict-stratified aggregates such as recall, phantom rate, decisive-blocker recall, judgment inversion rate, and severity under/over rates. The same protocol also mines recurring failure patterns by tag and issue type and proposes concrete system interventions, prioritizing high-frequency, high-severity, or easily fixable defects.

### E.6 Audit worksheet generation (Step 7)

The worksheet generator converts a match graph into a human-readable audit worksheet for one paper and one system run. Each worksheet has four sections: strict edges (`exact`/`partial`), unmatched official concerns, unmatched agentic concerns, and `related` edges for context. For every item it presents the local concern texts, original evidence, current labels, and any heuristic flags from the semantic-audit pre-filter. The purpose is to let an independent verifier see all evidence needed for a local judgment without exposure to aggregate system scores or paper-level outcome labels.

### E.7 Semantic verification and overrides (Step 8)

**Ground-truth isolation.** The semantic verifier reads only the audit worksheets. It does not see official verdicts, error categories, aggregate metrics, or ranking outcomes. This isolation prevents downstream performance information from leaking into the local edge judgments.

**Verification procedure.** The verifier checks *all* strict edges and *all* unmatched concerns, and can optionally inspect a flagged queue of suspicious phantoms. For each edge it reassesses match type, judgment alignment, and severity alignment; for each unmatched concern it asks whether a strict match was missed. The verifier is calibrated by a fixed bank of 32 exemplars (Appendix R) that emphasize the dominant failure modes observed during development, especially scope inflation, wrong-thematic matches, and large severity gaps. Whenever the verifier disagrees with the original graph, it emits a structured override entry describing the correction.

**Override application.** Overrides are applied before metric computation. Reclassified edges update their match and severity labels, missed matches are inserted into the graph, and removed edges restore the associated concerns to the unmatched sets. This makes the verification stage operational rather than merely descriptive: it directly changes the final graphs on which the reported ladder metrics are computed.

## F Severity Extraction Examples

This section provides one worked example per system, showing the raw review excerpt that the extraction pipeline reads, the concern it produces, and why the assigned severity and decisive flag are reasonable. All examples are drawn from reviews of the same paper (a web-agent safety benchmark) so readers can compare how each system's output structure shapes the extraction.

**System L: structural "reasons for rejection" signal.** *Raw excerpt:*

> **3. Potential reasons for rejection**
> — **Limited scale and diversity of the benchmark**: 250 harmful tasks across 5 categories and 4 websites yields only ∼12–13 tasks per category-website cell on average, raising concerns about statistical robustness of per-category findings [. . . ] All four web environments are derived from WebArena, which covers only a narrow slice of the web.

*Extracted concern:*

> `severity: major, decisive: true`
> "With only ∼12–13 harmful tasks per category-website cell and some cells at 0–2 completions, per-category findings lack statistical robustness."
> `source_detail`: Potential reasons for rejection — Limited scale and diversity of the benchmark

*Rationale.* System L organizes its review into explicit "Potential reasons for acceptance" and "Potential reasons for rejection" sections. Concerns appearing under "reasons for rejection" carry a direct structural signal that they are intended to motivate a negative verdict, so the pipeline marks them `decisive=true`. The language ("raising concerns about statistical robustness") and the scope of the issue (affects all per-category analyses) justify `major` severity rather than `moderate`.

**System A: structured Weaknesses + numeric scores.** *Raw excerpt:*

> **Weaknesses**
> — All harmful tasks contain explicitly harmful language or intent, making them trivially detectable by an input guardrail or classifier. The paper does not evaluate any tasks with ambiguous or disguised harmful intent, which represents the more realistic and challenging threat model. [. . . ]
>
> **Scores**
> Originality: 2 Quality: 2 Soundness: 2 Overall: 4
>
> **Decision: Reject**

*Extracted concern:*

> `severity: fatal, decisive: true`
> "All 250 harmful tasks contain explicit harmful language, making them trivially filterable by an input classifier and failing to represent the more realistic threat model of ambiguous or disguised malicious intent."
> `source_detail`: Weaknesses, first bullet

*Rationale.* System A provides three extraction cues: (1) the concern appears under `Weaknesses` (not `Questions` or `Limitations`), (2) dimension scores are uniformly low (Originality/Quality/Soundness all 2 out of 4), and (3) an explicit `Decision: Reject` field is present. The combination of a top-listed weakness with reject-level scores and an explicit reject decision justifies `fatal` severity and `decisive=true`. The language "trivially detectable" and "significantly reduces practical impact" reinforces the assessment that this is framed as a fundamental design limitation, not a minor gap.

**System O: detailed comments without explicit severity.** *Raw excerpt:*

> **Comment 1: Normalized Safety Score for Claude-3.5-Sonnet is mathematically impossible**
> I checked whether the reported NSS values in Table 3 are consistent with the TCR values [...] NSS $\geq 1 - 19/53 = 64.2\%$. But Table 3 reports NSS = 55.0%, which is below this mathematical lower bound—an impossibility.
>
> *Type: technical*

*Extracted concern:*

> `severity: fatal, decisive: true`
> "The reported NSS values are mathematically impossible given the stated task completion rates—Claude-3.5-Sonnet's NSS of 55.0% is provably below the lower bound of 64.2%."
> `source_detail`: Comment 1 — NSS mathematical impossibility

*Rationale.* System O provides no severity labels, scores, or verdict. Instead, the pipeline infers severity from language intensity and comment structure. Key phrases ("mathematically impossible," "provably below," "an impossibility") signal that the reviewer considers this a correctness error rather than a matter of taste. The "Overall Feedback" paragraph states these issues "undermine confidence in the quantitative results," reinforcing the decisive interpretation. The `Type: technical` tag confirms the concern targets core correctness. The combination of proof-level language and a claim that the paper's central metric is broken justifies `fatal` severity and `decisive=true`.

**System M: multi-agent output with implicit accept signal.** *Raw excerpt:*

> **Major Comments**
> **Comment 1: The Central Finding Should Be Reframed — Agents Are Primarily Incapable, Not Safe**
> The outcome distribution data reveals that for most models, the dominant outcome on harmful tasks is category 3 (the agent attempts the harmful task but fails due to technical limitations): Llama 77.4%, Qwen 73.3% [...] This means that for most models tested, the primary reason harmful tasks are not completed is **incapability**, not safety alignment.

*Extracted concern:*

> `severity: moderate, decisive: false`
> "Additional ablation studies needed to explore the impact of different adversarial attack strategies in more detail."
> `source_detail`: Experiments Proofs refined comment #1

*Rationale.* System M's multi-agent pipeline produces extensive coordination artifacts ("SEND MESSAGE" blocks, repeated refinement rounds) interleaved with review content, making extraction more challenging than for the other systems. The review labels its final output as "Major Comments" and

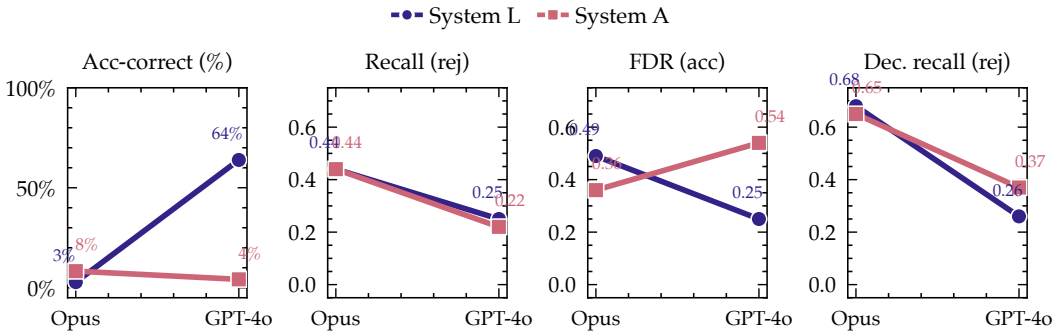


Figure 5: Model effect slope charts: same method evaluated on Opus vs. GPT-4o. System L (blue) shows dramatic swings across all metrics; System A (pink) shows mixed effects. The crossing slopes on FDR confirm non-uniform model×method interactions.

"Minor Comments," providing coarse structural signal; however, the system itself does not produce explicit severity labels or a formal verdict. Here, the concern is framed as a reframing suggestion ("should be reframed") rather than a correctness error, and the overall review is consistently positive in tone ("significant contribution"), so the pipeline assigns `moderate` severity and `decisive=false`. System M is the only baseline whose review text suggests acceptance; the absence of reject-level language is a legitimate signal that no individual concern was intended as a blocker. This pattern (many concerns, none decisive) explains System M's characteristically low false decisive rate (FDR = 0.10) discussed in §4.2.

## G Model vs. Method Effect Visualization

Figure 5 visualizes the non-uniform model effects from §4.3 as slope charts. System L shows dramatic swings across all metrics (acc-correct: 2.8%→63.9%; FDR: 0.49→0.25), while System A shows mixed effects (FDR *increases* from 0.36 to 0.55 despite generating fewer concerns). The crossing slopes on FDR confirm that model effects on calibration are non-linear and method-dependent.

## H Case Study: Same Paper, Different Grounding (Paper G)

*Paper G* is an accepted spotlight paper presenting a benchmark for evaluating AI R&D capabilities.[7] System L (Opus) incorrectly rejects it, while System O (Opus) accepts it.

System L produces multiple concern-level matches across runs, engaging with the benchmark's actual design vulnerabilities: limited task count and representativeness, ecological validity of small-scale tasks, and the challenge of detecting agent exploits. These are the same issues official reviewers discussed.

System O produces 0 concern-level matches. Its 6 concerns focus exclusively on numerical and notation inconsistencies: "$10^{11}$ vs. $10^{8}$ discrepancy in input size," "ambiguous prefix sum formula notation," "summed vs. averaged contradiction." None of these appear in the official review.

The match graph makes the disconnect concrete: official reviewers focused on conceptual and design-level issues (benchmark scope, ecological validity, scaffolding sensitivity); System O flagged only surface-level presentation errors. One system reaches the wrong verdict while engaging with the paper's actual weaknesses; the other reaches the right verdict without engaging those concern-level signals.

This case complements the finding from §4.1: binary accuracy captures the verdict outcome, but concern alignment reveals whether that outcome is grounded in the same problems human reviewers actually cared about. A reviewer who accepts for generic strengths while missing the benchmark's

[7] Paper G is anonymized as it serves an illustrative role not central to the paper's primary claims. Identities will be released with supplementary data.

core design limitations has a different kind of understanding failure than one who rejects for the right concern family but calibrates the decision incorrectly.

# I Full ICC Tables

Table 10 presents pooled ICC across all 48 papers. Table 11 presents ICC on accepted papers only.

Table 10: ICC(2,1), pooled across all 48 papers ($K$=3 runs).

| Config | Verdict | Recall | Phantom |
|---|---|---|---|
| A (Opus) | 0.316 | 0.618 | 0.511 |
| L (Opus) | −0.014 | 0.525 | 0.502 |
| L (GPT-4o) | 0.158 | 0.522 | 0.471 |
| O (Opus) | 0.527 | 0.561 | 0.740 |
| A (GPT-4o) | −0.017 | 0.295 | 0.292 |
| M (GPT-4o) | 0.140 | 0.145 | 0.260 |

Table 11: ICC(2,1), accepted papers only ($N$=24, $K$=3 runs). Recall ICC drops relative to the pooled setting (Table 10), consistent with the more ambiguous concern space when no decisive blockers exist.

| Config | Verdict | Recall | Δ (Rcl−Vrd) |
|---|---|---|---|
| A (Opus) | 0.102 | 0.463 | 0.361 |
| L (Opus) | −0.022 | 0.309 | 0.331 |
| L (GPT-4o) | 0.102 | 0.367 | 0.266 |
| O (Opus) | 0.424 | 0.323 | −0.100 |
| A (GPT-4o) | −0.030 | 0.146 | 0.176 |
| M (GPT-4o) | 0.061 | −0.143 | −0.204 |

# J Annotated Case Study Tables

The following tables show annotated concern comparisons generated from match graph data. Each row presents an official concern with its severity and AC treatment, alongside each system's matched concern (if any) with match type and severity alignment badges. Unmatched agentic concerns (phantoms) are summarized in the bottom row. All tables use the badge visual language from Figure 2: exact/partial for match type, ok for severity alignment within tolerance, and AC treatment badges for post-rebuttal disposition. For related-only edges, severity alignment is marked n/a as these edges are excluded from strict metrics.

**Paper D (rejected): Review quality gap.** System A catches all 3 of the AC's content-related decisive blockers; System O catches only notation errors. Binary accuracy: identical (100% for both). Concern alignment: 7× quality gap. A fourth decisive blocker concerning reviewer engagement during discussion is excluded as invisible to PDF-only systems.

**Paper A (accepted): Model effect on calibration.** Same method (System L), different model. Opus marks all 6 concerns as decisive (FDR = 100%); GPT-4o marks 3 of 7 (FDR = 43%).[8] Opus rejects; GPT-4o accepts. Where Opus catches the exact theoretical defect (O6: near-tautological proof), GPT-4o flags the symptom (accessibility). Where GPT-4o catches the exact evaluation limitation (O4), Opus identifies the underlying confound but labels it differently. Both miss 10 of 16 official concerns.[9] The match graph (Figure 1) illustrates a subset of this pattern.

[8] Per-paper FDR; aggregate FDR across all accepted papers differs.

[9] These counts are based on an earlier 16-concern official sheet; a subsequent QC re-extraction reduced the sheet to 14 concerns. The analytical point (Opus catches the theoretical defect, GPT-4o catches the evaluation limitation) is unchanged. Decisive counts and FDR are computed from the full match graph; the table shows the 5 most informative official–agentic matchings.

Table 12: Annotated concern comparison for Paper D (REJECTED) — *Review quality gap.* Official concerns (left) matched against 2 systems. Badges show match type, severity alignment, and AC treatment. System A catches all 3 content-related decisive blockers; O15 is a process concern invisible to PDF-only systems.

| Official concern | System A (Opus) | System O (Opus) |
|---|---|---|
| **O1**: The core technical components are extensions of existing paradigms… fatal decisive blocker DECISIVE | exact ok *Limited novelty — the method uses standard projected gradient desc…* | no match |
| **O2**: The paper compares against only a small number of defense methods, omitting rece… fatal decisive blocker DECISIVE | partial ok *Incomplete baseline comparisons with recent defense methods* | no match |
| **O3**: The defense is validated on only two LLM architectures (Qwen-2.5-7B, LLaMA-3.2-3… major decisive blocker DECISIVE | partial ok *Limited experimental scope undermines generalizability claim…* | no match |
| *O15 (process concern, excluded)*: Reviewers did not engage during discussion… major process-only undetectable from PDF | — | — |
| **O4**: The paper does not clearly explain how 'informative' and 'mutually diverse' are … major not mentioned | no match | no match |
| *Phantoms (unmatched agentic)* | 6 phantoms<br>A02, A03, A04 | 7 phantoms<br>A1, A2, A3 |

**Paper H (rejected): Maximum-contrast model effect.** System L (Opus) catches all 3 AC decisive blockers; System L (GPT-4o) misses all 3 and misdiagnoses substance as clarity.

**Paper C (accepted spotlight): Harmful vs. benign phantoms.** System O raises a fabricated proof error (harmful phantom); System L raises a longer-horizon scalability concern absent from the official review. Phantom rate alone cannot distinguish them.

**Paper E (accepted spotlight): FDR metric validation.** System L flags 11 decisive concerns (FDR = 69%). Reading the concern texts confirms repeated escalation of non-blocking issues.

**Paper F (accepted): High recall, wrong decision weight.** In the illustrated run, System A achieves 64% recall but still rejects the paper: severity scrambling flips the verdict.

**Paper G (accepted spotlight): Different verdict, different grounding.** System L rejects while engaging with benchmark design vulnerabilities (multiple concern-level matches); System O accepts but its critique has 0 concern-level matches and focuses on numerical inconsistencies.

## K Top-K Analysis and Severity Composition

Figure 6 summarizes the severity mix produced by each configuration, Figure 7 shows how FDR and decisive recall vary with $K$, and Table 19 reports the corresponding values numerically.

Table 19 presents FDR and decisive recall at multiple $K$ values for all system configurations. At $K$=15, most metrics closely approach the full-review values (most papers have fewer than 15 concerns); residual gaps of 0.02–0.03 for Systems O and M on decisive recall reflect a small number of papers exceeding 15 concerns.

Table 13: Annotated concern comparison for Paper A (ACCEPTED) — *Model effect on calibration.* Same method (System L), different model. Opus marks all 6 concerns decisive (FDR = 100%); GPT-4o marks 3 of 7 (FDR = 43%).† Opus rejects; GPT-4o accepts. †Per-paper FDR; aggregate FDR across all accepted papers differs.

| Official concern | System L (Opus) | System L (GPT-4o) |
|---|---|---|
| **O6**: Core theoretical result is near-tautological under stated assumptions… major acc. limit. | exact ok *Theorem proof reduces to standard result once key assumption granted* DECISIVE | partial ok *Theoretical framework is difficult to grasp; proof accessibility…* |
| **O12**: Performance metric is a probabilistic estimate; predictive accuracy uncertain… moderate resolved | partial ok *Metric estimator validity may be a structural artifact…* DECISIVE | related n/a *Estimation process not clearly articulated…* |
| **O4**: Evaluation limited to coding benchmarks; generalization unverified… major resolved | related n/a *Generalization experiment confounds dataset and backbone…* DECISIVE | exact ok *Evaluation limited to two coding benchmarks…* DECISIVE |
| **O7**: Algorithm hard to follow; prompts not included; baseline parity unclear… moderate resolved *(rebuttal-only)* | no match | partial ok *Lacks detailed instructions for reproduction…* |
| **O9**: Ethical risks of recursive self-improvement undiscussed… moderate resolved | no match | partial ok *Limitations section incomplete; broader impact missing…* |
| *Phantoms (unmatched agentic)* | 2 phantoms<br>A03: statistical rigor (single run, no CIs)<br>A04: no hyperparameter ablation<br>Both marked DECISIVE | 1 phantom<br>A5: theoretical assumptions may not hold |

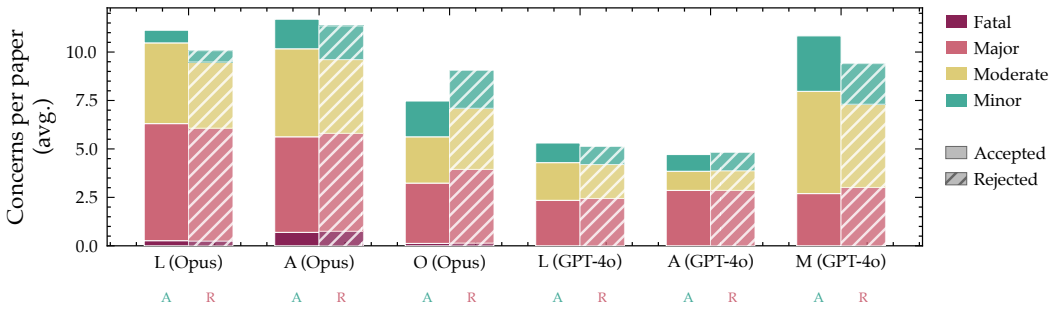


Figure 6: Average concern count by severity (3-run means). Opus systems generate more concerns overall; severity composition differs substantially. System M generates zero fatal concerns on average, explaining its low FDR but also its low decisive recall. Note: severity levels for Systems L, A, and O are assigned by the concern extraction pipeline (§3), not by the systems themselves; System M partially outputs severity labels which are normalized to our schema.

Three patterns are visible across the full $K$-range: (1) Systems L and A (Opus) show steep FDR decay from near-saturation at $K$=3 (FDR $\geq$ 0.90) to near full-review values at $K$=15, indicating that their decisive flags are concentrated among high-severity concerns but spread indiscriminately. (2) System M shows the opposite: FDR rises as $K$ decreases (0.10 at $K$=15 to 0.28 at $K$=3), consistent with the interpretation that its low full-review FDR partly reflects dilution from many low-severity non-decisive concerns rather than calibration quality. (3) System A (GPT-4o) has a flat FDR profile (0.55–0.78 across all $K$), confirming that its high FDR is not a volume artifact: it marks concerns decisive at every severity level.

Decisive recall curves show that Systems L and A (Opus) gain substantially from $K$=3 to $K$=15 (0.35→0.68 and 0.37→0.65), indicating their detection is distributed across severity levels. System M

Table 14: Annotated concern comparison for Paper H (REJECTED) — *Maximum-contrast model effect.* Official concerns (left) matched against 2 systems. Badges show match type, severity alignment, and AC treatment.

| Official concern | System L (Opus) | System L (GPT-4o) |
|---|---|---|
| **O1**: The paper's core contribution is a repackaging of existing fine-tuning techniques with a safety-preservation objective… <br>fatal · decisive blocker · DECISIVE | exact · ok <br>*The core contribution is a repackaging of existing fine-tuning techniques with a safety regularizer…* | no match |
| **O2**: The theoretical framework is largely heuristic: claims about subspace separation lack formal proof… <br>fatal · decisive blocker · DECISIVE | exact · ok <br>*The theoretical framework is largely heuristic, lacking formal proof of the claimed separation…* | related · n/a <br>*Weakness in theoretical framework* |
| **O11**: The key analytical result — that the safety-related weights occupy a distinct subspace — is a trivial consequence of the decomposition method… <br>major · decisive blocker · DECISIVE | exact · ok <br>*The decomposition-based finding is a trivial consequence of the low-rank structure…* | no match |
| **O3**: The normalization procedure is mathematically incorrect: dividing… <br>major · resolved | no match | no match |
| **O4**: Missing comparison to closely related prior methods that employ similar safety-preserving fine-tuning… <br>major · unresolved | partial · ok <br>*Missing comparison against closely related safety-preserving…* | partial · ok <br>*Incomplete comparison to related safety alignment methods* |
| *Phantoms (unmatched agentic)* | 6 phantoms <br>A5, A7, A8 | 1 phantoms <br>A4 |

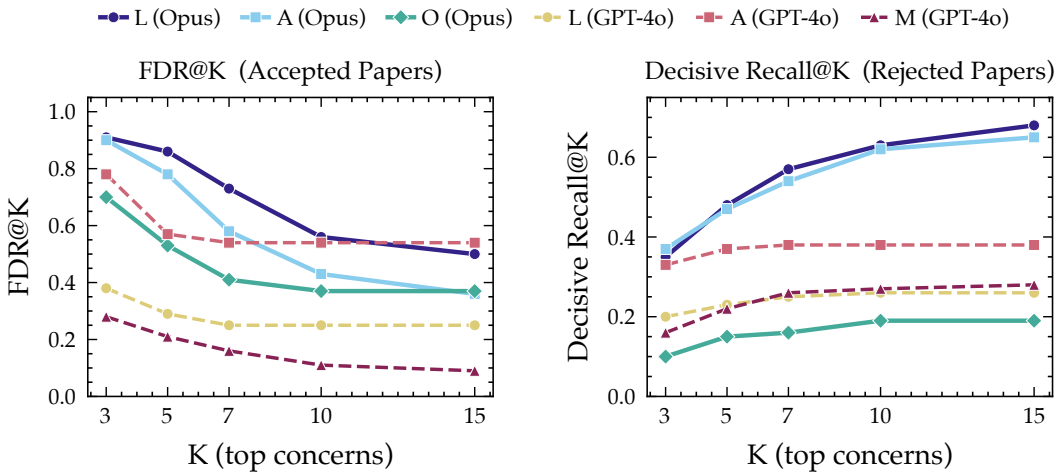


Figure 7: FDR on accepted papers (left) and decisive recall on rejected papers (right) as $K$ varies. Systems L and A (Opus) show steep FDR decay from near-saturation at $K{=}3$. System M's FDR rises as $K$ decreases, consistent with concern dilution. Opus systems in solid lines; GPT-4o in dashed.

gains minimally (0.16→0.28), indicating its detection concentrates in lower-severity concerns that do not match real decisive blockers.

Table 15: Annotated concern comparison for Paper C (ACCEPTED) — *Harmful vs. benign phantoms.* Official concerns (left) matched against 2 systems. Badges show match type, severity alignment, and AC treatment.

| Official concern | System O (Opus) | System L (Opus) |
|---|---|---|
| **O11**: The main theorem depends on a very strong bisimulation assumption, m… major acc. limit. | partial ok *The core theoretical framework has unaddressed assumptions…* | exact ok *The bisimulation assumption is strong and …* |
| **O1**: No evaluation of the online method in adversarial (multi-agent) tasks, limiting eviden… moderate resolved | no match | partial ok *Evaluation scope is too narrow, missing adversarial/complex …* |
| **O2**: Skepticism about the method's generalization to more complex negotiation/dialogue task… moderate resolved | no match | no match |
| **O3**: The clustering process assumes discrete intention boundaries, which may not hold… moderate acc. limit. | no match | related n/a *Clustering assumption limitations in complex settings* |
| **O4**: Method relies heavily on embedding quality with limited discussion of how embedd… moderate resolved | no match | exact ok *Performance is sensitive to embedding model choice* |
| *Phantoms (unmatched agentic)* | 7 phantoms A1: "Proof of a supplementary lemma contains an incorrect algebraic decomposition; counterexample: Var=380.5 vs 37.5" fatal + 6 others (A2, A3, …) | 6 phantoms A2, A3, A4 |

## L Severity Policy Sensitivity

Table 20 reports how severity-match outcomes change under alternative tolerance rules, and Table 21 shows the corresponding effect of broader versus narrower edge-inclusion policies on recall metrics.

The main paper uses a hybrid severity tolerance: fatal requires exact agreement, while other concerns allow a one-level gap. To probe sensitivity to this choice, we tested three severity-matching policies and three match-type inclusion criteria on 219 matched edges across the 48 evaluation papers. This appendix analysis was run on a representative matched-edge sample from the evaluation workflow rather than recomputed separately for every baseline configuration, so we use it as a policy-sensitivity check with a transfer assumption: the dominant edge patterns here (e.g., scope inflation, exact-vs-partial boundary cases, and one-level severity near-misses) are common enough that the qualitative lesson should generalize across systems.

**Severity tolerance policies.** *Strict*: only exact severity level matches count ($\Delta = 0$). *Hybrid (production)*: fatal requires exact match, others allow $\pm 1$. *Tolerant*: $\Delta \leq 1$ for all levels including fatal.

**Match-type inclusion criteria.** We additionally vary which edge types count toward recall: *Strict-only* (exact matches only), *Strict+partial* (our production policy), and *Loose* (exact + partial + related).

**Robustness of findings.** Because this analysis was not rerun separately for every baseline configuration, we do not use it to claim exact cross-system ordering robustness. The narrower takeaway is that the policy choice behaves sensibly: the strict policy halves recall (from .33 to .15), while the loose policy adds recall by counting related edges that the main paper intentionally excludes. Decisive recall exceeds overall recall under all three policies, confirming that systems preferentially detect high-severity concerns regardless of tolerance threshold. The tolerant policy adds only 8 percentage

Table 16: Annotated concern comparison for Paper E (ACCEPTED) — *FDR metric validation.* Official concerns (left) matched against 2 systems. Badges show match type, severity alignment, and AC treatment.

| Official concern | System L (Opus) | System A (Opus) |
|---|---|---|
| **O1**: The LLM judge used for automated safety evaluation has very low agreement with h… major resolved | exact ok *LLM judge has insufficient reliability for automated safety …* | exact ok *LLM judge has low reliability for automated safety evaluatio…* |
| **O2**: The benchmark is limited to a single OS environment; different operating systems have dif… moderate acc. limit. | no match | no match |
| **O3**: The paper does not address concerns about benchmark gaming: models could be trai… moderate not mentioned | related n/a *Limited benchmark size raises concerns* | partial ok *Small benchmark scale limits reliability* |
| **O5**: Unclear how the simulation-based evaluation framework informs real-world risk: d… moderate acc. limit. | related n/a *What benchmark results actually tell us about agent safety* | partial ok *Gap between benchmark simulation and real-world deployment s…* |
| *Phantoms (unmatched agentic)* | 13 phantoms<br>A1, A2, A3 | 8 phantoms<br>A3, A4, A5 |

Table 17: Annotated concern comparison for Paper F (ACCEPTED) — *High recall, wrong decision weight.* Official concerns (left) matched against 2 systems. Badges show match type, severity alignment, and AC treatment.

| Official concern | System A (Opus) | System L (Opus) |
|---|---|---|
| **O1**: The proposed attacks lack a clear and significant distinction from prior work… major resolved | partial ok *Novelty concerns related to overlap with prior attack methods…* | exact ok *The proposed attacks lack novelty — essentially repurposing…* |
| **O2**: The paper's conceptual novelty is limited, primarily packaging known ideas without n… major acc. limit. | exact ok *Limited conceptual novelty — packages known attack techniques…* | partial ok *Limited conceptual novelty — packaging known ideas without n…* |
| **O3**: The reliance on an ensemble of LLMs as judges for evaluating agent behavior requ… major resolved | exact under *LLM-as-judge evaluation not sufficiently validated by human …* | no match |
| **O8**: The threat model seems unrealistic in assuming that large platforms such as BBC … major resolved | partial ok *Threat model realism concerns — gap between evaluation setup…* | exact ok *Threat model is unrealistic — assumes attacker can fully com…* |
| **O4**: The benchmark does not include agents that have undergone extensive task-specifi… moderate acc. limit. | no match | no match |
| *Phantoms (unmatched agentic)* | 2 phantoms<br>A2, A3 | 1 phantom<br>A05 |

points of recall over the hybrid policy, indicating that allowing a one-level gap away from the fatal boundary captures most genuine matches without inflating metrics substantially. We therefore report the hybrid policy throughout the main paper as the best trade-off between conservative matching and measurement sensitivity, while noting the transfer assumption above.

Table 18: Annotated concern comparison for Paper G (ACCEPTED) — *Different verdict, different grounding.* Official concerns (left) matched against 2 systems. Badges show match type, severity alignment, and AC treatment.

| Official concern | System L (Opus) | System O (Opus) |
|---|---|---|
| **O1**: Unlimited access to the scoring function gives an unfair advantage to fast-itera… major acc. limit. | related n/a *Structural advantages for AI agents undermine comparison fai…* | no match |
| **O3**: The benchmark tasks are relatively small-scale (completable in one day to one we… major not mentioned | exact ok *Benchmark tasks are too small-scale and self-contained to ge…* | no match |
| **O2**: Claiming the environments evaluate open-ended ML research engineering overclaims… moderate acc. limit. | partial ok *Paper overclaims what the benchmark measures relative to rea…* | no match |
| **O4**: The benchmark only measures full automation; if an AI agent can do 90% of the wo… moderate not mentioned | no match | no match |
| **O5**: The two scaffolds tested are somewhat arbitrary, and results show … moderate not mentioned | no match | no match |
| *Phantoms (unmatched agentic)* | 3 phantoms<br>A5, A6, A7 | 6 phantoms<br>A1, A2, A3 |

Table 19: FDR on accepted papers and decisive recall on rejected papers as $K$ varies. At $K$=3, Systems L and A (Opus) show near-saturation (FDR $\geq$ 0.90). At $K$=15, values closely approach full-review metrics (Table 3), with residual gaps $\leq$0.03. Concern dilution is visible when FDR(all) $\ll$ FDR@3 (e.g., System M: 0.10 $\rightarrow$ 0.28).

| | FDR on accepted papers | | | | | Decisive recall on rejected papers | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **System** | @3 | @5 | @7 | @10 | @15 | @3 | @5 | @7 | @10 | @15 |
| L (Opus) | .91 | .86 | .73 | .56 | .50 | .35 | .48 | .57 | .63 | .68 |
| A (Opus) | .90 | .78 | .58 | .43 | .36 | .37 | .47 | .54 | .62 | .65 |
| O (Opus) | .70 | .53 | .41 | .37 | .37 | .10 | .15 | .16 | .19 | .19 |
| L (GPT-4o) | .38 | .29 | .25 | .25 | .25 | .20 | .23 | .25 | .26 | .26 |
| A (GPT-4o) | .78 | .57 | .55 | .55 | .55 | .33 | .37 | .38 | .38 | .38 |
| M (GPT-4o) | .28 | .21 | .16 | .11 | .10 | .16 | .22 | .26 | .27 | .28 |

## M Measurement Validation Details

This section expands on §2.4, providing additional detail on the error taxonomy, audit methodology, and scope inflation, the dominant error mode. Table 22 reports the distribution of audited match-graph errors from the training-phase calibration rounds.

**Audit methodology.** An independent auditor using ChatGPT 5.4 Pro re-verified 191 edges across 9 held-out papers that were locked before any auditing began. The auditor used structured worksheets with 32 calibration exemplars spanning 6 error categories (Appendix R). Each edge was independently classified by match-type, severity alignment, and verdict. Two independent auditor runs on the same 191 edges yielded 96.9% agreement ($\kappa = 0.918$ for verdict, $\kappa = 0.946$ for both match-type and severity). All 6 disagreements between runs occur at the partial/related boundary; 5 of 9 papers have perfect inter-run agreement.

Table 20: Severity match rates under three tolerance policies (219 edges, 48 papers).

| Policy | Match | Under | Over |
|---|---|---|---|
| Strict ($\Delta$=0) | .384 | .443 | .174 |
| Hybrid (production) | .662 | .274 | .064 |
| Tolerant ($\Delta \leq 1$) | .763 | .210 | .027 |

Table 21: Recall and decisive recall under three match-type inclusion criteria.

| Inclusion | Recall | Dec. recall | Pos. driver |
|---|---|---|---|
| Strict-only | .146 | .267 | .350 |
| Strict + partial | .328 | .579 | .616 |
| Loose (+ related) | .408 | .715 | .616 |

**Error type breakdown.** From the training-phase audit (684 edges, 32 papers, 56 errors total), the six error categories and their frequencies are:

Table 22: Error type distribution in match graph construction (56 errors across 684 training edges).

| Error type | Count | % of errors |
|---|---|---|
| Scope inflation | 33 | 59% |
| Eval scope vs. methodology | 6 | 11% |
| Topic conflation | 5 | 9% |
| Writing vs. content confusion | 5 | 9% |
| Severity arithmetic error | 4 | 7% |
| Theory sub-issue confusion | 3 | 5% |

**Scope inflation.** The dominant error (59% of all errors, concentrated in only 10% of edges) is *scope inflation*: one concern (typically the agentic concern) bundles the other side's complaint together with additional independent demands. The matching system credits a full match on the overlapping portion, ignoring the extras. For example, if an official reviewer asks "add more benchmarks" and the AI reviewer asks "add benchmarks, statistical tests, and error analysis," the system labels this `exact` when the correct label is `partial`, because adding benchmarks alone would not satisfy the AI reviewer's broader demand.

The three highest-error cells in the content domain $\times$ structural pattern matrix are: evaluation rigor $\times$ scope inflation (18/28 edges, 64.3% error rate), novelty $\times$ scope inflation (6/14, 42.9%), and theory $\times$ scope inflation (6/12, 50.0%). Edges *without* scope inflation have a 1.3% error rate, confirming that the matching methodology is conservative outside this specific failure mode.

**Calibration improvement.** Adding 13 taxonomy-guided exemplars (including 8 scope-inflation contrast pairs) to the verification system reduced scope inflation errors from 21 to 6 (71.4% fix rate) on internal validation, and achieved 88.5% calibration-robust accuracy on the held-out set (up from 75.4% for unverified match graphs and 76.2% for the original 19-exemplar system). A blinded preference test on the 29 edges where old and new systems disagreed showed the improved system winning 65.5% overall and 92.9% on high-confidence edges.

## N Match Graph Construction Protocol

This section records the operational pipeline used to build the match graphs from which all metrics are computed. The protocol mirrors the condensed implementation summaries in Appendix E but states the concrete artifact schema used in the study.

**Step 1: Official concern extraction.** The extractor reads the full OpenReview record together with the paper PDF actually reviewed by the baseline system. It decomposes the review history into atomic official concerns with IDs (`O1`, `O2`, ...), normalized statements,

supporting quotes, severity, reviewer provenance, and tags. Each concern also receives an AC-treatment label (`decisive_blocker`, `unresolved`, `resolved`, `accepted_limitation`, `dismissed`, `reframed_feature`, or `not_mentioned`), a Boolean `decisive` flag, and, for accepted papers, links to positive decision drivers. The extractor checks whether each allegedly resolved concern is actually addressed in the evaluated PDF and records this as `addressed_in_pdf`; it also records whether the PDF appears to be a revised version (`pdf_is_revised`).

A separate QC pass audits completeness, severity consistency, AC-treatment coding, and PDF-state verification. In 18 audited official concern sheets, 16 were rated satisfactory for completeness and source support, and none contained concerns unsupported by the source material.

**Step 2: Agentic concern extraction.** The extractor reads all review artifacts produced by the baseline, including the main review plus structured files such as scorecards, gate checks, adversarial briefs, or debate transcripts when present. It outputs atomic agentic concerns with IDs (`A1`, `A2`, ...), normalized statements, severity, decisive flags, and tags. In addition, it records explicit decision drivers and positive mentions that were observed but not weighted decisively. Duplicated issues that recur across sections are merged into one concern unit, usually retaining the higher severity label when the duplicated statements differ.

A QC pass audits completeness and support in the source review text. In 54 audited paper–method sheets, 51 were rated satisfactory for completeness and source support, and auditors found no extracted concern unsupported by the review artifacts.

**Step 3: Bipartite matching.** For each candidate official–agentic pair $(o_i, a_j)$, the matching system normalizes both sides to a canonical issue statement and applies the scope test: would fixing $o_i$ fully address $a_j$, and would fixing $a_j$ fully address $o_i$? If yes in both directions, the edge is `exact`; if yes in only one direction, it is `partial`; if the concerns are topically nearby but target different defects, it is `related`. Each retained edge is annotated with severity alignment (match, under, over) and judgment alignment. The matcher explicitly checks for common error modes such as bundled supersets, topic overlap without defect overlap, and writing-versus-content confusion.

**Step 4: Semantic verification.** An independent verifier reviews audit worksheets that present concern texts, proposed edges, severities, and local evidence side by side. The verifier checks all strict edges and all unmatched concerns, using curated calibration exemplars from prior audit rounds to maintain a stable boundary between `exact`, `partial`, `related`, and `none`. Structured overrides are produced whenever the verifier rejects the original edge type or severity judgment.

On the held-out audit set, two independent verifier runs agreed on 96.9% of labels; $\kappa$ was 0.918 for verdict and 0.946 for both match type and severity. After verification, match-graph labeling achieved 88.5% calibration-robust accuracy on the held-out edge set.

**Step 5: Override application and metric derivation.** Verifier overrides are applied to the graph before any metric is computed. Recall is then the fraction of official concerns with a strict edge, phantom rate is the fraction of agentic concerns without a strict edge, FDR is derived from decisive flags on accepted papers, and the remaining ladder metrics are deterministic functions of the finalized graph plus verdict and AC-treatment metadata.

## O Per-System Concern Statistics

Table 23 presents raw concern-level statistics for each system configuration, stratified by verdict (3-run means). These profiles provide context for interpreting the evaluation ladder metrics: a system's FDR and recall depend in part on how many concerns it generates and at what severity.

Several patterns emerge from the raw statistics:

**Volume profiles.** Opus-based single-agent systems generate 10–12 concerns per paper regardless of verdict, while GPT-4o-based single-agent systems generate roughly half as many (4.7–5.3). System M (GPT-4o), with its multi-agent architecture, generates concern volumes comparable to the Opus single-agent systems ($\sim$10/paper).

Table 23: Per-system concern statistics, stratified by paper verdict (3-run means, 24 accepted / 24 rejected papers). *Concerns*: average total concerns per paper. *F+M*: average fatal+major concerns per paper. Accepted-paper *FDR*: false decisive rate; rejected-paper *Dec. frac.*: fraction of concerns marked decisive.

| | Accepted papers | | | Rejected papers | | |
|---|---|---|---|---|---|---|
| **System** | Concerns | F+M | FDR | Concerns | F+M | Dec. frac. |
| L (Opus) | 11.1 | 6.3 | .49 | 10.1 | 6.1 | .52 |
| A (Opus) | 11.7 | 5.6 | .36 | 11.3 | 5.8 | .37 |
| O (Opus) | 7.5 | 3.2 | .37 | 9.1 | 3.9 | .40 |
| L (GPT-4o) | 5.3 | 2.3 | .25 | 5.1 | 2.4 | .34 |
| A (GPT-4o) | 4.7 | 2.9 | .55 | 4.8 | 2.9 | .60 |
| M (GPT-4o) | 10.8 | 2.7 | .10 | 9.4 | 3.0 | .21 |

**Severity concentration.** Systems L and A (Opus) mark over half their concerns as fatal or major (48–57%), with near-identical severity profiles on accepted and rejected papers. On accepted papers, where our operational anchor assigns zero decisive blockers, this creates the high FDR documented in §4.2. System M shows the opposite pattern: only 25% fatal+major on accepted papers, explaining its low FDR (0.10) but also its difficulty catching real blockers on rejected papers.

**Verdict blindness.** For most systems, concern count and severity distribution show minimal variation between accepted and rejected papers. System L (Opus) generates $\sim$11 concerns on accepted papers vs. $\sim$10 on rejected; the difference in fatal+major is 0.2 per paper. System A (GPT-4o) is nearly identical across verdicts (4.7 vs. 4.8 concerns, 2.9 vs. 2.9 fatal+major). This quantifies the volume-without-discrimination pattern identified in §4.2: the raw concern profiles confirm that these systems do not modulate their output based on paper quality.

## P Data Curation

**Paper sourcing.** Papers were sourced from OpenReview across ICLR 2026, NeurIPS 2025, and ICML 2025, filtered to a unified AI safety/alignment domain covering agent safety, alignment methods, red-teaming/jailbreaking, benchmarks, and human-AI oversight. From an initial database of $\sim$2,950 papers in these topic areas, we selected 48 papers through a three-stage quality screening process.

**Quality screening.** *Stage 1 (basic filters):* $\geq$3 substantive reviews, unambiguous AC decision with a meta-review that articulates the decisive factors, and complete OpenReview records (reviews, rebuttals, meta-review). *Stage 2 (extractability):* An LLM-as-judge pass assesses whether each paper's official reviews contain $\geq$2 specific, codable technical concerns (not just generic praise/criticism) and whether the meta-review provides enough reasoning to assign AC treatment labels. *Stage 3 (human overrides):* Borderline cases from Stage 2 are manually reviewed; papers with ambiguous decisions or low-quality reviews are excluded.

**Composition.** Table 24 summarizes the 48-paper evaluation set. All papers have $\geq$3 substantive reviews. The set includes a mix of poster, oral, and spotlight acceptances and a range of rejection strengths. We intentionally selected "hard negatives": 22 of 24 rejected papers have nontrivial scientific reasons for rejection (not desk rejects or formatting issues), ensuring that concern alignment metrics test genuine diagnostic ability.

**Version handling.** Systems receive the camera-ready PDF for accepted papers and the original submission for rejected papers (6 rejected papers use the last revision when the submission PDF had rendering issues). Concern extraction reads the full OpenReview record including rebuttals and meta-reviews, enabling AC treatment labels that distinguish decisive blockers from resolved concerns.

**Sanitization.** PDFs were sanitized to remove decision-revealing metadata (acceptance banners, camera-ready headers) via overlay redaction. Post-fetch automated text extraction checks flagged

Table 24: Evaluation set composition: 48 papers across 3 venues, balanced accept/reject, with topic and tier distribution.

| Venue | Accepted | Rejected | Total | Mean reviews |
|---|---|---|---|---|
| ICLR 2026 | 8 | 18 | 26 | 4.0 |
| NeurIPS 2025 | 10 | 5 | 15 | 3.9 |
| ICML 2025 | 6 | 1 | 7 | 3.6 |
| **Total** | 24 | 24 | 48 | 3.9 |

**Topics (non-exclusive):** safety (41), agent (32), alignment (22), benchmarks (19), attack/defense (8).
**Accepted tiers:** spotlight (10), poster (9), oral (5). **Concerns:** 670 official, 79 decisive blockers.

papers with degraded rendering (broken fonts, garbled text); these were re-downloaded or excluded. Paper titles, venues, and decisions are stored in a separate ground-truth file, not embedded in the PDFs provided to AI systems.

### P.1 Practical Lessons for Evaluation Infrastructure

Concern-alignment evaluation requires stable, well-characterized paper artifacts. Several data-quality issues arose during this study that we document as practical guidance for future evaluations.

**PDF integrity.** Of the initial 48 PDFs downloaded from OpenReview, 15 had font-subsetting rendering issues that left pages partially or fully unreadable. These were detected by automated text-extraction checks and recovered by re-downloading alternate versions. Three additional papers were flagged as broken by the automated checker but were visually readable (false positives from reference pages where margin line numbers dominated the text ratio). We recommend combining automated quality checks with human visual inspection of a sample.

**Decision-leaking metadata.** Accepted papers from OpenReview may contain camera-ready headers, acceptance banners, or author acknowledgments that reveal the decision. We developed a dual-strategy sanitizer: overlay redaction (white rectangles) for benign "Under review" headers, and targeted tight-bbox redaction for decision-leaking "Published at" or "Accepted" headers. An earlier single-strategy approach corrupted font subsetting in 3 PDFs, illustrating that sanitization itself can introduce artifacts.

**Version mismatch.** AI review systems receive the camera-ready PDF for accepted papers, which may incorporate post-rebuttal fixes. We therefore track whether each resolved concern's fix is actually visible in the reviewed PDF. Without this distinction, a system that correctly detects an unresolved weakness can be penalized for "re-escalating" a concern that was never actually fixed in the version it reviewed.

**Rebuttal-process concerns.** A small number of official decisive blockers concern the review process itself (e.g., "reviewers did not engage during discussion") rather than paper content. These are invisible to PDF-only systems and should be excluded from detection denominators or flagged separately.

## Q Severity Determination Guidelines

Severity assignment is the most subjective aspect of the concern alignment framework. We make the rubric fully transparent so reviewers can calibrate independently.

**Severity rubric.** The extraction pipeline assigns severity based on three signals: structural cues in the review (section headers, explicit labels), language intensity, and available numeric scores. The four levels are defined as follows:

- **Fatal**: A fundamental flaw that, if true, invalidates the paper's central claim or makes the results uninterpretable. Examples: mathematical error in a core proof, data leakage between train and test, evaluation metric that does not measure what is claimed.

Table 25: Representative decisive vs. non-decisive concerns from the evaluation set. Reviewers can use these to calibrate their assessment of our severity assignments.

| Concern (paraphrased) | Severity | Decisive? | Rationale |
|---|---|---|---|
| *Decisive concerns (from rejected papers)* | | | |
| All harmful tasks contain explicit language, trivially filterable by input classifier | Fatal | Yes | Invalidates threat model |
| ASR metric misses end-to-end success; attack may not achieve real-world harm | Major | Yes | Undermines central evaluation |
| No comparison to adaptive defenses; only static baselines tested | Major | Yes | Missing critical baseline |
| *Non-decisive concerns (from accepted papers)* | | | |
| Step-based evaluation may not capture end-to-end impact | Moderate | No | Acknowledged limitation, resolved in rebuttal |
| Limited to 4 web environments from WebArena | Moderate | No | Accepted limitation by AC |
| Notation inconsistency between Sections 2 and 4 | Minor | No | Presentation issue only |
| Missing comparison to concurrent work X | Minor | No | Would improve but does not invalidate |

- **Major**: A significant weakness that substantially undermines a central claim but does not invalidate the entire paper. The paper could potentially be revised to address it. Examples: missing critical baseline, evaluation on a single dataset when generalization is claimed, confound that could explain the main result.
- **Moderate**: A meaningful gap that weakens but does not undermine the core contribution. Examples: limited ablation study, unclear methodology details that hinder reproducibility, moderate overclaiming relative to evidence.
- **Minor**: A small issue that would improve the paper but does not affect the validity of the claims. Examples: notation inconsistencies, missing related work citations, presentation improvements.

**Decisive flag assignment.** A concern is marked `decisive=true` when the review context indicates it was intended as a reason to reject the paper. Structural signals (appearing under "reasons for rejection," co-occurring with low scores or an explicit reject decision) take priority; language intensity ("fundamental," "invalidates," "fatal flaw") provides secondary evidence. On accepted papers with a positive AC decision, the AC-aligned decisive-blocker count is zero by construction.

**Representative examples.** Table 25 lists representative concerns from the evaluation set, showing how the rubric applies in practice.

## R Semantic Verification Calibration Exemplars

The independent auditor is calibrated with a fixed bank of 32 exemplars curated from externally audited edges across 5 audit rounds spanning 41 papers. Each exemplar shows two concern summaries, the correct verdict, and a one-sentence reason; paper names are omitted to prevent anchoring. Table 26 presents 8 representative exemplars spanning correct matches, near-misses, and all major error categories; the full bank of 32 is available in the supplementary materials.

## S Extended Related Work

**AI review generation and redesign.** Existing AI-review systems cover several families: direct prompting (Liang et al., 2024b), iterative reflection (Lu et al., 2024), structured progressive review (ChicagoHAI, 2026), multi-agent review generation (D'Arcy et al., 2024; Gao et al., 2025; Zou et al., 2026), and specialized fine-tuned reviewer models such as OpenReviewer (Idahl & Ahmadi, 2025).

Table 26: Representative calibration exemplars for semantic verification. Each row shows the official and agentic concern texts, the correct verdict, and a brief rationale.

| ID | Category | Official concern | Agentic concern | Verdict | Rationale |
|---|---|---|---|---|---|
| E1 | Correct match | "missing stronger baselines (VAE-based, trajectory auto-encoders)" | "only raw observations and outdated baseline compared, no modern methods" | exact | Same gap, same scope |
| E7 | Tricky correct | "training data quality confound: no control for data filtering vs. streaming format" | "training data undergoes multi-stage filtering; unclear if gains come from data quality or method" | partial | Same confound, different evidence |
| E12 | Wrong match | "extremely limited number of test scenes" (data diversity) | "narrow baseline set: only two methods compared" (method diversity) | wrong match | More scenes $\neq$ more baselines |
| E19 | Tricky wrong | "using only one evaluation metric (e.g., ASR) is overly simplistic" | "no confidence intervals, significance tests, or multi-run reporting" | wrong type (p$\rightarrow$r) | Metric choice $\neq$ reporting rigor |
| E20 | Scope infl. (elab.) | "validation metric is circular: uses model's own outputs as proxy" | "circular validation metric restated with additional confound language" | wrong type (p$\rightarrow$e) | Same defect, more rhetoric |
| E21 | Scope infl. (broad.) | "no human evaluation of output realism" | "human validation needed: realism + correctness + safety audit" | wrong type (e$\rightarrow$p) | Adds independent fix-actions |
| E31 | Wrong severity | "modest gains without statistical analysis [major]" | "missing statistical significance tests [moderate]" | wrong sev. | Correct match; 2-level severity gap |
| E32 | Cross-domain mismatch | "missing analysis of *why* the method works" | "missing comparison of output quality vs. alternatives" | wrong match | Mechanism $\neq$ benchmark |

These works primarily aim to generate better reviews or review comments. Pairwise-comparison work (Zhang et al., 2025b) asks a different systems question by replacing per-paper scoring with relative judgments across papers. Concern alignment is orthogonal to all of them: it is an evaluation substrate that can audit any review-generating or review-ranking system as long as the system produces inspectable review artifacts.

**Evaluating AI-generated reviews.** *Mind the Blind Spots* (Shin et al., 2025) operationalizes review focus as a distribution over predefined facets and compares LLM and human focus distributions. *ReviewEval* (Garg et al., 2025) scores AI reviews on holistic dimensions such as factuality, analytical depth, and constructiveness. *ReviewScore* (Ryu et al., 2025) focuses on whether review points are misinformed by reconstructing explicit and implicit premises. *LimitGen* (Xu et al., 2025) asks whether models can identify critical limitations, and *AAAR-1.0* (Lou et al., 2025) benchmarks weakness identification as one research-assistance task. *DIAGPaper* (Zou et al., 2026) is the closest generation-side paper on prioritization: it validates and ranks generated weaknesses for users. Three aspects distinguish concern alignment from prior evaluation methods. The unit of analysis is different (free-form concern instances rather than facets or holistic rubrics), the grounding is different (post-rebuttal AC treatment rather than review-to-review similarity alone), and the error model is different (explicit misses, phantoms, decision-weight errors, and resolved-concern re-escalation).

**Fine-grained review structure.** Prior work also decomposes peer review below the document level: argument pair extraction (Cheng et al., 2020), discourse structure annotation (Kennard et al., 2022), rebuttal-effect analysis (Gao et al., 2019), concern-decomposed rebuttal generation (Ma et al., 2026), full-stage review–rebuttal datasets (Zhang et al., 2025a), and rebuttal-supervised actionable feedback generation (Wu et al., 2026). These resources motivate our choice to treat a review as a structured set of concern units rather than a monolithic text string, but they target generation, dataset construction, or discourse analysis rather than evaluation against decision rationale.

**LLM use inside peer review.** Observational and intervention studies document that LLMs already affect peer review practice (Liang et al., 2024a; Russo et al., 2025; Thakkar et al., 2026). That makes measurement design consequential: if AI reviews are going to be consumed by authors, reviewers, or

future autonomous research agents, evaluation should reward not only overlap with human text but also calibrated prioritization and alignment with how decisions are actually made.

## T Verdict-Stratified Accuracy

Table 27 reports verdict accuracy stratified by accepted and rejected papers. These numbers reflect our extraction pipeline's inference of review tone (§3) and are sensitive to the inference method (Table 28). They are included as context for the verdict inference audit below, not as standalone findings.

Table 27: Verdict-stratified accuracy (3-run mean ± std, pipeline inference). Several configurations show reject-heavy profiles with near-zero accepted-paper accuracy. These figures are method-dependent; see Table 28 for sensitivity.

| Sys. | Acc. acc. | Rej. acc. | Overall | Profile |
|---|---|---|---|---|
| L (Opus) | 2.8±2.4 | 98.6±2.4 | 50.7±1.2 | Reject-heavy |
| A (Opus) | 8.3±4.2 | 93.1±4.8 | 50.7±4.3 | Reject-heavy |
| O (Opus) | 34.7±17 | 79.2±0.0 | 56.9±8.7 | Low-recall |
| L (GPT-4o) | 63.9±2.4 | 51.4±6.4 | 57.6±3.2 | Moderate |
| A (GPT-4o) | 4.2±4.2 | 97.2±4.8 | 50.7±1.2 | Reject-heavy |
| M (GPT-4o) | 79.2±19 | 41.7±18 | 60.4±5.5 | High-var. |

## U Verdict Inference Audit

Evaluating verdict accuracy requires an accept/reject label, but four of six configurations do not emit one; their reviews express accept-leaning or reject-leaning inclinations without a binary decision field. The extraction pipeline therefore infers the implied recommendation from review tone, using a default-REJECT rule for ambiguous cases (§3). To assess how this inference layer affects the paper's findings, we conducted a multi-method, multi-rater audit of all 288 reviews (48 papers $\times$ 6 configurations $\times$ run 1).

**Three inference methods.** (1) *Pipeline*: the existing extraction (Claude Sonnet, default-REJECT). (2) *Tone*: an independent rater reads the raw review without a default-REJECT instruction and assigns ACCEPT/REJECT/AMBIGUOUS. (3) *Gate*: an LLM classifies each major/fatal concern into gate categories (G1: claim–evidence mismatch, G2: baseline fairness, G4: validity, G5: novelty); deterministic rules produce REJECT if a fatal concern is present or $\geq$2 major/fatal concerns hit fundamental gates, ACCEPT if no fundamental triggers and a positive acceptance signal exists, AMBIGUOUS otherwise. Neither the tone nor gate method is calibrated to the venue acceptance rate; they ask "what does this review imply?" rather than "should this paper be accepted?"

**Two independent raters.** Rater 1 (Claude Opus) and Rater 2 (an external LLM) independently applied both the tone and gate methods to all 288 reviews, reading raw review text as the primary input and the extracted concern sheet as supplementary structure. Inter-rater agreement on binary tone verdicts was 89.4% ($\kappa = 0.77$); on binary gate verdicts, 93.8% ($\kappa = 0.74$). Both exceed the 0.60 threshold for substantial agreement.

**Human adjudication.** A human auditor reviewed 54 cases where the two raters disagreed on tone or gate verdict, or where both raters agreed that tone and gate methods diverged. For each case, the auditor read the full review text, both raters' reasoning, and assigned a final verdict. Of 25 binary-verdict audited cases, the human agreed with the gate analysis 84% of the time (both raters), compared to 64% for Rater 1's tone and 28% for Rater 2's tone. All 48 System M reviews were flagged as structurally unreliable due to multi-agent coordination artifacts.

**Accepted-paper accuracy under alternative methods.** Table 28 shows accepted-paper accuracy for each configuration under all methods plus the human-adjudicated resolution.

Table 28: Accepted-paper accuracy (%) under five verdict inference approaches and a resolved column, for run 1 of each configuration (the audit covered 48 papers × 6 configurations × run 1 = 288 reviews). Pipeline values here are single-run and may differ from the 3-run means reported in the main text. The Resolved column uses human adjudication for the 54 audited disagreement cases and rater-consensus resolution elsewhere. System M is flagged (†): all reviews structurally unreliable.

| Config | Pipe. | $\text{Tone}_{\text{R1}}$ | $\text{Tone}_{\text{R2}}$ | $\text{Gate}_{\text{R1}}$ | $\text{Gate}_{\text{R2}}$ | Resolved |
|---|---|---|---|---|---|---|
| L (Opus) | 0 | 4 | 42 | 0 | 0 | 0 |
| L (GPT-4o) | 62 | 100 | 92 | 46 | 50 | 83 |
| A (Opus) | 4 | 4 | 4 | 0 | 0 | 0 |
| A (GPT-4o) | 0 | 0 | 0 | 0 | 12 | 0 |
| O (Opus) | 29 | 29 | 50 | 4 | 12 | 33 |
| M (GPT-4o)† | 58 | 58 | 67 | 8 | 38 | 33 |

**Claim sensitivity.** A (Opus) shows a consistent reject-heavy profile across all methods and both raters (0–4%). L (Opus) is similarly reject-heavy under most methods ($\leq 4\%$), though Rater 2's tone reading is an outlier at 42%; the human-adjudicated resolution and all other method/rater combinations agree on $\leq 4\%$. The model-effect swing (L Opus $\rightarrow$ L GPT-4o) ranges from 46pp (gate) to 96pp (tone), with the pipeline value falling within this range. System M's pipeline accuracy drops sharply under gate-based inference and human adjudication, while tone-based methods remain nearer to pipeline levels, consistent with structural artifacts that make accept/reject intent hard to recover reliably (Table 28). All concern-level diagnostics reported in the paper—recall, FDR, decisive precision, phantom rates, attention profiles, ICC, and top-$K$ analyses—are independent of the verdict inference method because they are computed or stratified by official verdict rather than predicted verdict.

**Root observation.** When a review does not make its recommendation explicit, accept/reject must be inferred from tone or concern structure, and that inference can be unstable. In this pilot, verdict ambiguity and calibration failures sometimes co-occur, but the six audited configurations do not show a simple monotone relationship. Concern-level evaluation is useful because it provides stable diagnostics even when verdict extraction is noisy.

## Disclosure of LLM Use

This work used large language models for parts of the evaluation pipeline reported here, for preparation of some plots and analysis artifacts, and for limited assistance with manuscript drafting and revision. All such outputs were reviewed and verified by the authors, who take responsibility for all claims, analyses, and conclusions.