LEARNING CHECKLISTS TO REDUCE LOTTERY EFFECTS IN SCIENTIFIC PEER REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Reviewer lottery error describes error due to a small set of peer reviews attending to only some of the issues with a paper that a larger pool of eligible reviewers would raise. We show this error can be reduced by surfacing a checklist of criticisms learned from a historical review corpus. Our approach generates a tree of checklists representing coarsenings of the atomic review claim space, then selects the coarsening that maximizes a theoretically-motivated measure of correlated agreement among reviewers. We realize this approach with an LLM-assisted pipeline and test it on NeurIPS 2021 and ICLR 2022 review data. In simulated review with LLM reviewers, the learned checklist reduces the mean squared error (MSE) of a three-reviewer committee by 58% on NeurIPS and 21% on ICLR, and on both venues the checklist with the highest correlated agreement has the lowest error for committees of three or more. In a human study with 20 participants, the learned NeurIPS checklist lowers the MSE of a paper ranking, measured against the NeurIPS 2021 order, from 0.48 to 0.22.

1 INTRODUCTION

Peer review systems have long played the role of a quality control mechanism for scientific production. However, human peer review is far from being a perfect system, as a result of bias, poor incentives, limited expertise, and other issues (Mann et al., 2025; Shah, 2022). One recognized problem is reviewer lottery variance: a different assignment of qualified reviewers may have resulted in a different outcome. "Luck of the draw" effects have been demonstrated across various fields (Cole et al., 1981; Mayo et al., 2006; Beygelzimer et al., 2023; Cortes & Lawrence, 2021). For example, the paired-committee 2021 NeurIPS reviewing experiment found that the two committees disagreed on 23% of papers, and that roughly half of the accepted papers would have been replaced if review assignments were rerun (Beygelzimer et al., 2023). Beyond differences in expertise and in how reviewers perceive the paper, variance can arise because any single reviewer covers only a small portion of the evaluation space (Deng et al., 2026).

More severely, as LLMs increase scientific production (Kusumegi et al., 2025), higher reviewing loads per human reviewer increase noise, instigating a "review death spiral" that further incentivizes authors to try their luck by submitting weak papers (Bergstrom & Gross, 2026; Liu & Tan, 2026).

LLMs have been proposed as part of the solution (Liang et al., 2024; Xu et al., 2025; Wei et al., 2026). For example, by systematically conducting and surfacing a broad set of checks across papers, AI reviewers can reduce lottery variance by compensating for the limited attention and idiosyncrasies of individual human reviewers. This ensures a common set of baseline checks is applied to each paper, while still allowing the individual reviewers to exercise their judgment and expertise in weighing surfaced issues and identifying other aspects of the paper to assess. However, the risk that must be balanced is that a shared representation evaluating the paper can introduce bias by causing reviewers to underweight their private information (see, e.g., Bikhchandani et al., 1992; Stasser & Titus, 1985).

We first characterize the problem of designing an AI checklist as identifying a representation of the review claim space that, when applied to a paper, reproducibly reduces lottery variance without increasing overall error via the addition of shared bias. We formalize checklist design as learning a coarsened representation of the review claim space observed in historical reviews. Because historical peer review datasets do not enable estimating reviewer decisions under checklists that reviewers didn't observe, our approach targets the coarsening that maximizes a proxy objective based in correlated

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

agreement: agreement between independent reviews of the same paper relative to the expected agreement between reviews of different papers.

We next contribute an automated pipeline for learning and implementing checklists using LLM agents. Our approach takes as input a historical review corpus, and decomposes each review into atomic claims, each labeled by the source of information that would resolve it. We require the claims to be settled by the paper or its supplementary material, such that an LLM can verify the learned checklist with high accuracy. The claims are then grouped across papers into a hierarchy of named checklist items, and our proxy objective selects the granularity at which that hierarchy is applied.

We evaluate whether checklists learned using our correlated agreement objective reduce finite-reviewer error on existing review datasets. In a simulated review experiment with LLMs on 30 papers each from NeurIPS 2021 and ICLR 2022, we find that for a typical committee of three reviewers, the learned checklist reduces the review error by 58% on NeurIPS and 21% on ICLR. On NeurIPS 2021, it also has lower error than the official NeurIPS 2021 paper checklist at every committee size. On both datasets, the checklist with the highest correlated agreement also has the lowest error for committees of three or more reviewers, so our proxy objective identifies the checklist that best balances variance reduction and bias addition.

2 Related Work

Human peer review has played the role of an imperfect quality control mechanism for scientific production for over a century (Mann et al., 2025; Shah, 2022). Reviewer lottery effects are demonstrated in grant funding by replicating review processes and finding that outcomes depend on the draw (Cole et al., 1981; Mayo et al., 2006). In machine learning, a reanalysis of the 2014 duplicated assignment NeurIPS experiment attributes half of the variation in reviewer quality scores to reviewer subjectivity (Cortes & Lawrence, 2021). The 2021 NeurIPS experiment finds substantial variation in accepted papers across two non-communicating committees who reviewed the same 882 papers (Beygelzimer et al., 2023). Computational interventions proposed to address lottery issues have tended to target assignment, calibration, and dishonest behavior rather than what reviewers attend to (Shah, 2022).

AI reviewers bring opportunities and risks. LLM-generated comments can overlap human reviews at rates comparable to a second human reviewer (Liang et al., 2024), and models can perform well on targeted, verifiable questions (Liu & Shah, 2023). State-of-the-art AI review systems increasingly rely on agentic, multi-stage pipelines that ground their comments in quoted passages or other verifiable evidence to improve interpretability and accuracy (Gao et al., 2025; Nguyen et al., 2026; Weng et al., 2026; Fang et al., 2026; Ma et al., 2026; Chicago Human+AI Lab, 2026; Reviewer3, 2026; Jiang & Ng, 2025). Given the infeasibility of large scale human peer review studies, such tools are often tested via recall of implanted defects; others propose LLM simulations as a testbed for peer review mechanisms that can recover aggregate dynamics of human peer review feedback and outcomes (Jin et al., 2024; Erturk & Durmaz, 2026). Venues have also begun to experiment with AI in review, including offering LLM reviews (Thakkar et al., 2025; Biswas & Sahu, 2026) and LLM review assistants in OpenReview (Biswas & Charlin, 2026).

However, AI reviewers can be gameable (Baumann et al., 2026), can hallucinate (Deng et al., 2026), and attend unevenly across review dimensions, over-weighting technical validity and under-weighting novelty (Shin et al., 2025). Applying consistent summary representations also contributes bias, by attempting to reduce scientific quality to a finite set of proxies that can cause reviewers to undervalue their private knowledge (Bikhchandani et al., 1992; Stasser & Titus, 1985).

3 Problem: Learning a Checklist to Reduce Lottery Variance

We present a model of peer review and the problem of reviewer lottery variance. We formulate the task of learning a checklist to reduce variance from a historical peer review corpus.

3.1 Reviewer Lottery Variance

One way to frame the goal of peer review is to uncover the latent true epistemic quality of a paper. However, even if a ground truth paper quality state could be said to exist, it would be unobservable.

Thus, a more realistic formalization of the goal of peer review is to recover the expected assessment of the paper taken over the eligible reviewer population. The problem is that we can typically only observe the assessments of a small subset of the ideal group of reviewers.

We define the population review target for a given paper i as the mean assessment under the distribution of eligible reviewers for that paper:

$$\mu_i = \mathbb{E}_{r\sim P_i}[Y_{ir}],$$

where $r \sim P_i$ represents a random reviewer draw from the population of eligible reviewers for paper i and Y_{ir} is the judgment of an unaided reviewer r on paper i. While in practice Y_{ir} could take various forms, including multi-dimensional assessments like rubric-style evaluations, for the purposes of formalizing the problem we assume that $Y_{ir} \in \mathbb{R}$ is a scalar representing a final paper score. A peer review system that targets m reviewers per paper estimates μ_i as:

$$\mathbf{R}_{im} = (r_1, \ldots, r_m) \sim P_i^{(m)}; \qquad \bar{Y}_{im} = \frac{1}{m}\sum_{j=1}^{m} Y_{ir_j};$$

where $P_i^{(m)}$ is the reviewer assignment mechanism. While $\mathbb{E}[\bar{Y}_{im}] = \mu_i$, any downstream decisions that rely on aggregated reviewer assessments are subject to sampling error, where the small values of m that are typical in human peer review can yield estimates far from μ_i.

Lottery variance may arise as a result of each reviewer's expertise being specialized relative to the range of expertise across eligible reviewers, or reviewers having limited time or cognitive resources to spend surfacing issues, or from some reviewers making errors in interpreting aspects of the paper.

Estimating Reviewer Variance and Incorrect Criticism on the NeurIPS 2021 Experiment To motivate our characterization of errors in human peer review, we contribute a novel claim-level analysis of the NeurIPS 2021 dual-committee experiment. We use the overlap between issues across the two independent committees for each paper, after merging common issues, to estimate impacts of lottery variance at the claim level. Full details are in Appendix B.

We found that independent reviewers rarely mention the same issue, and that many concrete criticisms (those that can be checked directly against the paper) are refuted by the paper materials. Specifically:

- **Low coverage**: One reviewer raises about 6% of the distinct issues that further reviewers drawn from the same pool would eventually raise (6.4%, 95% CI $[5.9, 6.9]$, Chao2 richness (Chao, 1987)).
- **Low overlap**: Only one third of reviewer claims are also raised by the other committee (34.3%, 95% CI $[32.8, 35.8]$), and 81% of issues are raised by a single reviewer (Appendix B).
- **False criticisms**: The paper itself clearly refutes 18.2% of all judged concrete claims (95% CI $[16.3, 20.2]$).[1]

These results suggest that one reviewer raises only a small share of the issues this reviewer pool would eventually surface, and that humans make mistakes in a non-trivial proportion of review claims.

3.2 Reducing Variance By Surfacing a Common LLM-Completed Checklist

We seek to reduce reviewer lottery variance by introducing a shared checklist g. A checklist consists of a set of binary claims that can be resolved against the paper materials with high accuracy and surfaced for assigned reviewers. In doing so, it stands to reduce variance arising from sources like reviewer myopia or careless mistakes. Introducing a checklist as intervention results in $Y_{ir}(g)$ at the individual reviewer level, with checklist-assisted population mean $\mu_i(g) = \mathbb{E}_{r\sim P_i}[Y_{ir}(g)]$ (where the no checklist target can similarly be written as $\mu_i(\varnothing) = \mathbb{E}_{r\sim P_i}[Y_{ir}(\varnothing)]$). Let the mean score of m reviewers ($\mathbf{R}_{im}$) be

$$\bar{Y}_{im}(g) = \frac{1}{m}\sum_{j=1}^{m} Y_{ir_j}(g).$$

[1]We judge this with an LLM (Appendix B.4). The check uses the camera-ready PDF, which works in the other direction: material added after review can refute a criticism that held for the submission.

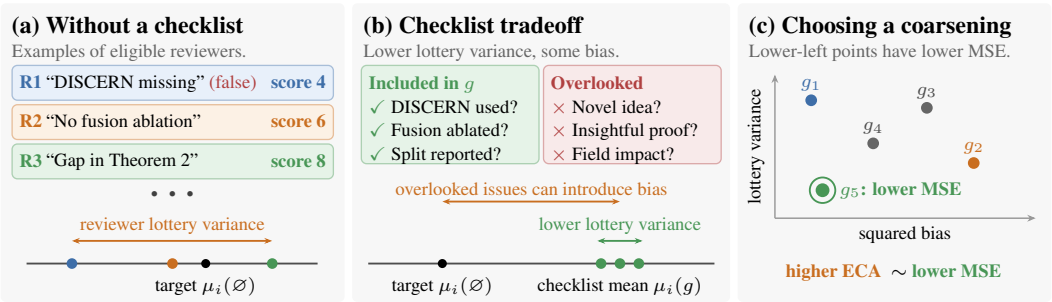


Figure 1: The problem and our selection strategy. (a) The ellipsis denotes additional reviewers from the eligible-reviewer population. Their differing scores create reviewer lottery variance around $\mu_i(\varnothing)$, the unaided population review target. (b) A shared checklist aligns attention on represented dimensions, lowering lottery variance, but omitted dimensions can shift the checklist-assisted population mean $\mu_i(g)$ and introduce bias. (c) Candidate coarsenings trade lottery variance against squared bias. Because historical reviews do not reveal counterfactual MSE, we select using expected correlated agreement (ECA), $A_{\text{same}}(g) - A_{\text{diff}}(g)$, then evaluate variance and bias separately.

Using

$$\text{MSE}_i(g, m) = \mathbb{E}_{\mathbf{R}_{im} \overset{\text{i.i.d.}}{\sim} P_i}[(\bar{Y}_{im}(g) - \mu_i(\varnothing))^2]$$

for the error of checklist-assisted review, and $\text{MSE}_i(\varnothing, m)$ for the corresponding quantity under unaided review, a good checklist satisfies:

$$\text{MSE}_i(g, m) < \text{MSE}_i(\varnothing, m).$$

That is, the total error induced by the checklist will be less than that observed under no checklist. Intuitively, a checklist can reduce variance insofar as variation across reviewer scores results from variation in which issues reviewers notice or the possibility of careless mistakes checking a verifiable issue against the paper. However, by exposing all reviewers to a shared representation, introducing a checklist may also systematically shift their assessments away from the unaided population review target by biasing what information reviewers pay attention to or how they weight it.

By the standard bias-variance decomposition, the mean squared error of a checklist-assisted reviewer assignment can be written as:

$$\text{MSE}_i(g, m) = \text{Var}(\bar{Y}_{im}(g) \mid i) + (\mu_i(g) - \mu_i(\varnothing))^2. \tag{1}$$

We want to minimize $\text{MSE}_i(g, m)$. Because error in unaided review amounts to variance only:

$$\mathbb{E}[\bar{Y}_{im}(\varnothing) \mid i] = \mu_i(\varnothing) \quad \Rightarrow \quad \text{MSE}_i(\varnothing, m) = \text{Var}(\bar{Y}_{im}(\varnothing) \mid i),$$

introducing a checklist improves finite-sample estimation only when

$$\text{Var}(\bar{Y}_{im}(\varnothing) \mid i) - \text{Var}(\bar{Y}_{im}(g) \mid i) > (\mu_i(g) - \mu_i(\varnothing))^2; \tag{2}$$

that is, where the reduction in variance from using the checklist is larger than the added squared bias. Figure 1 sketches the lottery, the intervention, and why checklist design requires choosing a coarsening that reduces variance without introducing too much bias.

3.3 Checklist Selection

Existing review corpora provide a rich source for learning what kinds of concerns human reviewers perceive as important. But how should we identify the set of relevant checkable issues from which a checklist can draw? Simply distilling the space of atomic review claims from peer review corpora into checklist items would face multiple problems. First, issues learned directly from human reviews may be value judgments rather than properties verifiable against the paper or supplement, which is where LLM checks are most reliable. We therefore want to restrict to those issues that can be directly resolved by the paper materials. Second, if atomic review claims are framed too specifically or idiosyncratically, they may not capture reproducible signal in human review, and add bias if instituted

across reviewers. Such a checklist could also be prohibitively large for human reviewers to process, so that only those that apply to a given paper are surfaced.

If human reviews contain real signal about how the community views a paper, but any individual claim may be framed too idiosyncratically to recur across reviewers as more are added, then checklist design becomes a problem of granularity: a space of issues that is too low level may yield little recurrence as reviewers are added, while one that is too high level may yield recurrence simply because the issues are framed so broadly to apply to any papers. We therefore treat checklist design as finding a coarsened representation of the observed claim space, restricted to verifiable issues, on which reviewers of the same paper agree more than reviewers of different papers do.

Let $C_{ir} \subseteq \mathcal{C}$ be the set of atomic review claims extracted from reviewer r's review of paper i, where $\mathcal{C}$ is the claim space observed across a historical review corpus. Let $\mathcal{C}_{\text{ver}} \subseteq \mathcal{C}$ be the claims in this space that are verifiable against the paper or its supplementary material. A checklist is a map $g : \mathcal{C}_{\text{ver}} \to \mathcal{S}_g$, where $\mathcal{S}_g$ is a partition of $\mathcal{C}_{\text{ver}}$ into nonempty, disjoint claim clusters. Each signal $s \in \mathcal{S}_g$ is thus a subset of claims, and $g(c) = s$ means that claim c belongs to signal s. Let $\mathcal{G}$ denote a finite family of candidate coarsenings g.

Ideally, we would select a checklist that minimizes MSE averaged over the population of papers. Let Q denote this distribution of papers, such that the population objective is $\mathbb{E}_{i\sim Q}[\mathrm{MSE}_i(g, m)]$. However, doing so would require being able to observe $Y_{ir}(g)$ (reviewer r's assessment of paper i under checklist g) for each candidate checklist g. Existing review datasets provide only the assessment under the issues the reviewer noticed ($Y_{ir}(\varnothing)$). We therefore instead use a measure of expected correlated agreement defined on the review claim space to select the coarsening of $\mathcal{C}_{\text{ver}}$ according to a property conducive to variance reduction: issues raised by independent reviewers of the same paper recur much more often than those raised by reviewers of different papers.

Given a checklist g, each review can be represented as a binary vector $Z^g_{ir} \in \{0, 1\}^{|\mathcal{S}_g|}$, standing for the activation of the checklist signal by claims in review:

$$Z^g_{ir}(s) = \mathbf{1}[\, C_{ir} \cap s \neq \varnothing \,], \qquad s \in \mathcal{S}_g, \tag{3}$$

where $\mathbf{1}[\cdot]$ is the indicator function. A signal is raised if the reviewer made at least one of the claims in its cluster. As claims are merged into fewer checklist signals, coarsening increases the chance that two reviews activate the same signal, even for unrelated papers.

Let $A(C, Z') = \sum_{c\in C\cap\mathcal{C}_{\text{ver}}} Z'(g(c))$ count the verifiable claims of a review C whose signal another review Z' also raises, and let $\bar{n}$ be the expected number of verifiable claims per review. We define

$$A_{\text{same}}(g) = \bar{n}^{-1}\, \mathbb{E}_{i\sim Q}\Big[\mathbb{E}_{r,r' \overset{\text{i.i.d.}}{\sim} P_i}[A(C_{ir}, Z^g_{ir'}) \mid i]\Big], \tag{4}$$

$$A_{\text{diff}}(g) = \bar{n}^{-1}\, \mathbb{E}_{i,j \overset{\text{i.i.d.}}{\sim} Q}\Big[\mathbb{E}_{r\sim P_i,\, r'\sim P_j}[A(C_{ir}, Z^g_{jr'}) \mid i, j]\Big]. \tag{5}$$

The two reviews share a paper in (4) and come from independent papers in (5), so both are shares of review claims whose signal the other review also raises. Denoting the **expected correlated agreement (ECA)** $\Delta(g) = A_{\text{same}}(g) - A_{\text{diff}}(g)$, we select

$$g^* \in \arg\max_{g\in\mathcal{G}} \Delta(g). \tag{6}$$

The objective rewards reproducible overlap within papers after subtracting overlap across papers.

Theoretical guarantee. We provide a result showing when our checklist is an approximation of the optimal checklist. For a fixed number of reviewers m, let $R_m(g) = \mathbb{E}_{i\sim Q}[\mathrm{MSE}_i(g, m)]$ be population MSE and $B(\mathcal{S}_g) = R_m(\varnothing) - R_m(g)$ be the MSE reduction from checklist g.

Theorem 3.1 (Informal)**.** *Suppose MSE reduction is monotone and submodular in the set of surfaced signals, signals that contribute positively to ECA have positive stand-alone MSE benefit, and the signals of g^* are not too redundant with one another. Then the maximizer g^* in* (6) *approximates the optimal checklist: $B(\mathcal{S}_{g^*}) \geq \alpha \max_{g\in\mathcal{G}} B(\mathcal{S}_g)$ with $\alpha = (1 - \kappa)/\rho$. Here $\rho \geq 1$ is the largest ratio between two signals' stand-alone MSE benefits per unit of ECA, and $\kappa \in [0, 1)$ bounds the average fraction of a signal's stand-alone benefit that is redundant given the other signals of g^*.*

The approximation ratio depends on how consistently reviewers benefit from signals. The guarantee approaches optimality as $\rho \to 1$ and $\kappa \to 0$. Appendix A gives the formal theorem and proof.

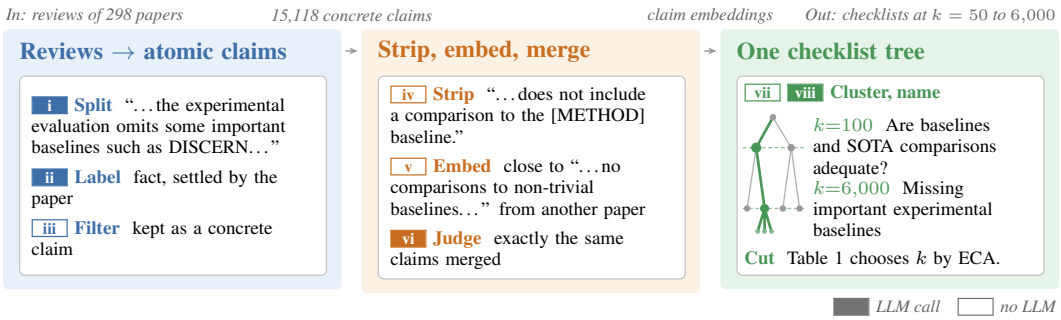


Figure 2: The checklist pipeline, shown on one NeurIPS 2021 review sentence. Steps (i)–(viii) follow the text; filled badges call an LLM. Counts are for the NeurIPS 2021 corpus.

Table 1: Concrete-claim hierarchies for NeurIPS 2021 and ICLR 2022 at eight granularities k (all values in %). Coverage is the mean share of the k items that a paper's reviewers cover. Recall is the share of 32 human-raised points (two on each of 16 papers) that an LLM finds from the paper with the checklist. ECA is $A_{\text{same}} - A_{\text{diff}}$ (Section 3.3), the sample estimate of (6), with paper-level bootstrap 95% CIs. On both corpora, A_{same} compares pairs of reviewers of the same paper and A_{diff} compares reviewers of different papers. Bold marks the highest ECA per corpus.

	NeurIPS 2021					ICLR 2022		
k	Coverage	Recall	A_{same}	A_{diff}	ECA [95% CI]	A_{same}	A_{diff}	ECA [95% CI]
50	42.0	96.9	22.1	12.4	+9.8 [8.5, 11.1]	38.9	27.2	+11.7 [10.4, 13.1]
100	26.1	93.8	16.2	6.1	**+10.2** [9.0, 11.3]	28.8	15.9	**+12.9** [11.8, 14.0]
200	14.8	84.4	12.3	3.2	+9.0 [8.2, 10.0]	21.7	9.2	+12.6 [11.7, 13.5]
500	6.7	68.8	8.7	1.4	+7.3 [6.7, 8.0]	15.8	4.2	+11.6 [10.9, 12.3]
600	5.7	65.6	8.2	1.1	+7.1 [6.4, 7.8]	14.9	3.5	+11.5 [10.8, 12.1]
1,000	3.6	56.2	6.6	0.7	+5.9 [5.4, 6.5]	13.1	2.3	+10.8 [10.2, 11.4]
2,000	1.9	34.4	4.5	0.4	+4.2 [3.8, 4.5]	10.9	1.2	+9.8 [9.2, 10.4]
6,000	0.7	21.9	2.0	0.1	+1.8 [1.6, 2.1]	8.1	0.4	+7.7 [7.2, 8.2]

4 An LLM-assisted Pipeline for Learning Checklists

We realize our approach using an LLM-assisted pipeline that learns a checklist from a historical review corpus by maximizing the expected correlated agreement measure $A_{\text{same}} - A_{\text{diff}}$ ((6)).

Given a review corpus, we grow a tree of checklists with eight steps (Figure 2). Claims from papers for the evaluation in Section 5 are held out from the start, so they do not enter the tree and cannot leak into the checklist. (i) Use an LLM (Grok 4.5) to split every review into atomic claims that refer to a self-contained, independent comment about the paper. Claims across different papers are put into a single pool. (ii) Use an LLM to label each claim with what it asserts, using the proposition types of Hua et al. (2019) (fact, evaluation, request, quote, reference, or non-argument), and what source would settle it: the paper, the literature, the world, or nothing (Appendix C.1). (iii) Filter any claims that are not labeled as facts or requests that the paper itself can settle. (iv) Strip paper-specific numbers, symbols, and rare method names to make claims applicable to other papers. (v) Embed each claim with Qwen3-Embedding-0.6B. (vi) Merge claims judged to be exactly the same by the LLM (Grok 4.5) (Appendix C.4). (vii) Grow a Ward tree over the merged claims, representing each by the mean embedding of its members. (viii) Use an LLM to relabel each cluster (checklist item) with a question that can be posed against the paper, given a sample of its members.

Given a checklist tree, cutting at any level returns a checklist at a given granularity. For example, cutting the tree near its root returns fewer, higher-level checklist items such as "Are baselines and SOTA comparisons adequate?" Cutting the tree near its leaves returns more lower-level checklist items such as "Does the paper report experiments on ATIS, Advising, or WikiSQL?"

We next calculate the expected correlated agreement (ECA) for each checklist in the tree. Because the original human claims all lie at the leaf level, we must first map each review claim into the new checklist space. For each reviewer, we label a checklist item as applying if there is at least one original claim in that reviewer's review that falls under the checklist item.

We demonstrate this pipeline on two review datasets: the NeurIPS 2021 dual-committee experiment (Beygelzimer et al., 2023) (the public OpenReview subset: 298 papers, each reviewed by two independent committees) and the ICLR 2022 review dataset (Samarth P et al., 2026) (300 selected papers: 175 rejected, 100 posters, 19 spotlights, and 6 orals). The NeurIPS dataset is the richest corpus because of its dual-committee design; the ICLR 2022 dataset is a more typical one.

Table 1 reports ECA at eight different cuts (granularities) on both datasets. We also report coverage: for each paper, the share of the k items raised by at least one of its reviewers, averaged over papers. We also report recall, the fraction of 32 points that human reviewers raised (two on each of 16 papers), that an LLM also finds when given the paper and the checklist (Appendix C.5). Coverage, recall, and A_{diff} all decrease monotonically with k, indicating that a checklist with a lower k is more general. ECA is positive at every cut. On both venues we estimate it from pairs of reviewers of the same paper, and it is highest at k=100 ($+10.2$ points on NeurIPS and $+12.9$ on ICLR), which we take as g^*.

5 Evaluation: Do the Selected Checklists Reduce Reviewer Error?

We ask two questions to evaluate our checklist learning pipeline: (Q1) when the review committee size N is small, can the learned checklist reduce review MSE? (Q2) Given a set of candidate checklists, does the proxy objective ECA choose the one that gives the minimum review MSE?

The time commitment and expertise required for peer review make a human study difficult. It could take dozens of independent reviewers to reach a stable population target $\mu_i(\varnothing)$ against which the review MSE can be computed, and reviewers would need to score each paper using multiple checklists. No existing peer review dataset satisfies either criterion, and collecting such a large review dataset is beyond the scope of this paper.

We therefore assess our method under our model assumptions with a larger LLM-simulated review study, following precedent in prior work on review tools (Jin et al., 2024; Erturk & Durmaz, 2026) and evidence that LLM simulations can often replicate aggregate-level human population dynamics (Ashokkumar et al., 2026; Cui et al., 2025). We also validate error reduction with human reviewers in a smaller experiment.

5.1 Simulation Experiment

Our simulation constructs a simulated reviewer population from which $\mu_i(\varnothing)$ can be estimated precisely. Each simulated reviewer mimics the limited coverage we observed in human reviews (Section 3.1), by focusing on a related subset of concerns and raising a small number of concrete issues before assigning a score. We then measure how the variance, bias, and total error of the committee mean change with committee size N, with and without a checklist. To compare the learned checklist with an existing one, we also run the same protocol with the official NeurIPS 2021 paper checklist (Conference on Neural Information Processing Systems, 2021a), which reviewers that year were asked to use as a reference during review.

Datasets. We evaluate checklists learned on the NeurIPS 2021 dual-committee dataset (Beygelzimer et al., 2023) and the ICLR 2022 review dataset (Samarth P et al., 2026). We evaluate on a set of 10 strong, 10 borderline, and 10 weak or rejected papers from each dataset, decided based on their review scores and decisions. Their claims are held out when the checklist tree is grown (Section 4), so the items are not fit on the papers used to measure error. Papers from each venue are reviewed with the checklist learned from that venue's review corpus, and all simulated reviewers are LLM agents running Composer 2.5.

Population target. We estimate the population target $\mu_i(\varnothing)$ by aggregating a large number of independent reviewers. To mimic human reviewers, we create multiple LLM reviewers with limited

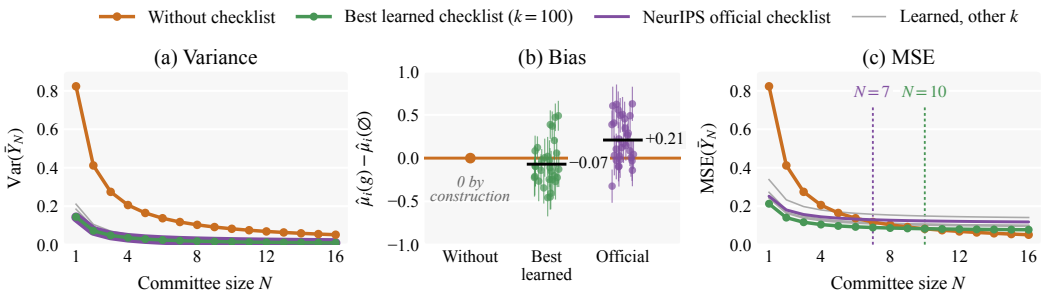


Figure 3: Variance, bias, and MSE of the committee mean on NeurIPS 2021 (30 papers). (a) Variance falls as $1/N$ in every condition but starts far lower with any checklist; the learned and official curves coincide. (b) Each point is one paper's bias, the checklist mean minus the target $\hat{\mu}_i(\varnothing)$ (± 1 SE); black bars average over papers. Without a checklist the bias is zero by construction. (c) MSE, the variance plus the squared bias. Dotted lines mark the smallest N at which the no-checklist committee is at least as accurate. Grey lines are learned checklists at other granularities, $k \in \{50, 500, 1000, 2000\}$.

coverage of issues and capacity by assigning each a different branch of the checklist tree at the k=50 level (Appendix C.6 gives the details of the assignment). The reviewers review the paper within their branch and, in a separate round, grade it according to the paper content and their own review. We draw one such reviewer for each of the F content branches (F=49 for NeurIPS and 47 for ICLR). With this number of reviewers, the aggregate score that defines the population target $\hat{\mu}_i(\varnothing)$ is stable (standard error about 0.13 points; Figure 4 in Appendix D).

Without checklist. The without-checklist setting shows how the MSE changes as the committee grows. A committee of size N consists of N reviewers drawn independently from the population described above, which reproduces the reviewer lottery variance observed in duplicated-review experiments (Cortes & Lawrence, 2021; Beygelzimer et al., 2023). Since the committees are drawn from the same population that determines the target, this condition is unbiased by construction: its error consists only of variance and decreases at a rate of $1/N$. This sets the conditions to evaluate whether a checklist reduces variance by more than the squared bias it introduces (Equation (2)).

With checklist. In the checklist setting, each LLM reviewer reviews the paper using the learned checklist at a given granularity. The reviewer goes through every item on the checklist, keeps those applicable to the paper, and writes concrete criticisms according to the checklist items, each supported by a quote from the paper. In a separate round, it grades the paper with the same scoring prompt as the without-checklist setting. We run eight reviewers per paper, who each see the checklist in a different random order. To answer Q1, we compare the MSE of checklist-assigned committees with that of without-checklist committees of the same size. To answer Q2, we repeat the condition at several granularities, $k \in \{50, 100, 500, 1000, 2000\}$ on NeurIPS and additionally k=200 on ICLR, and check whether the granularity with the highest ECA also has the lowest MSE.

Metrics. We compare the expected variance, bias, and MSE of the committee mean with and without the checklist. For each paper and condition, we treat the observed reviewer scores as an empirical reviewer population. For a committee of size N, under independent sampling the variance of the committee mean is the single-reviewer variance divided by N. The bias is the difference between the average checklist score and the target, $\hat{\mu}_i(g) - \hat{\mu}_i(\varnothing)$, and does not depend on the committee size N. The review MSE combines the two as $\mathrm{MSE} = \mathrm{Variance} + \mathrm{Bias}^2$, following (1). We average over papers with equal weight, squaring each paper's bias before averaging so that biases of opposite sign on different papers do not cancel.

5.2 Simulation Results

Q1: The learned checklist reduces MSE for small committees. On NeurIPS (Figure 3c), the best learned checklist (k=100) lowers the MSE of a single reviewer from 0.82 to 0.21 and halves it for a

committee of four (0.10 vs. 0.21). The gain comes from lower variance, at the price of a small bias (Figure 3a,b). The no-checklist committee, which is unbiased by construction, catches up with the best checklist at about $N{=}10$. On ICLR (Figure 5 in Appendix D), the $k{=}100$ checklist lowers the MSE of a single reviewer from 0.84 to 0.34, and the no-checklist committee catches up at $N{=}5$.

Comparison with the official checklist. The official NeurIPS 2021 checklist reduces variance as much as the best learned one but has higher bias. Its MSE is about 0.04 higher than that of the learned checklist at every committee size (0.25 vs. 0.21 at $N{=}1$, 0.14 vs. 0.10 at $N{=}4$).

Q2: ECA picks a low-MSE granularity. On both venues, the checklist with the highest ECA, $k{=}100$, has the lowest MSE for every committee of three or more reviewers, and the one with the lowest ECA, $k{=}2{,}000$, has the highest (at $N{=}4$, 0.10 vs. 0.18 on NeurIPS and 0.21 vs. 0.29 on ICLR), suggesting ECA is predictive of how much a checklist reduces review error. On NeurIPS, $k{=}100$ is lowest at every N. On ICLR, $k{=}200$ and $k{=}1{,}000$ are slightly lower at $N{=}1$ (0.32 and 0.33 vs. 0.34), and $k{=}200$ is level with it at $N{=}2$ (both 0.25).

5.3 Human Study

In a human validation study, 20 participants each rank two sets of three NeurIPS 2021 papers, one set with the checklist at $k{=}100$ (Table 1) already filled out for each paper by an LLM (Grok 4.5), and the other without a checklist. One set is on neural network pruning, which removes weights or channels to make a network cheaper to train or run, and the other is on adversarial robustness. Topic and checklist condition are counterbalanced, and Appendix C.7 lists the papers. A ranking task has two advantages. The order is a less contestable target than using the small-N average of human scores as $\mu_i(\varnothing)$, since the selected papers differ in mean scores (from about 5 to 7), writing quality, and citations. Ranking is also easier for participants than writing full reviews and assigning scores.

We compute variance and bias on pairs of papers, as in Shivaswamy & Chandrashekar (2021) and Cormack & Grossman (2019). Each set has three pairs. For each pair, we record 1 if a participant puts the two papers in the NeurIPS order and 0 otherwise, and apply (1) with a target of 1. If a share p of participants get the pair right, the variance is $p(1-p)$ and the squared bias is $(1-p)^2$. Their sum, averaged over pairs, is the fraction of pairs a participant gets wrong. With the checklist, a single ranking has variance 0.16 and squared bias 0.057 (MSE 0.22); without it, both are 0.24 (MSE 0.48). For a committee of three, the MSE is 0.11 with the checklist and 0.32 without.

6 Conclusion

We frame checklist design for peer review as a choice of how coarsely to represent the space of review claims: a shared checklist helps a small committee only if it removes more lottery variance than the bias it adds. We then build an LLM-assisted pipeline that turns a historical review corpus into a tree of paper-checkable checklist items and chooses the granularity that maximizes a proxy objective, the expected correlated agreement (ECA), which measures how much more two independent reviews of the same paper agree than reviews of different papers. We apply the pipeline to NeurIPS 2021 and ICLR 2022. In simulated review, the chosen checklist reduces the error of a three-reviewer committee by 58% on NeurIPS and 21% on ICLR, and on both venues the checklist with the highest ECA has the lowest error for every committee of three or more. In a human ranking study with 20 participants, the checklist lowers ranking MSE from 0.48 to 0.22.

AI Use Statement

We used AI tools to proofread and polish the writing of this paper and to organize the code submitted with it.

References

Ashwini Ashokkumar, Luke Hewitt, Isaias Ghezae, and Robb Willer. Large language models can predict the results of social science experiments. *Nature*, pp. 1–8, 2026.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation, 2026. URL https://arxiv.org/abs/2605.03202. ICML 2026 Position Paper (Oral).

Carl T Bergstrom and Kevin Gross. Screening, sorting, and the feedback cycles that imperil peer review. *PLoS Biology*, 24(2):e3003650, 2026. doi: 10.1371/journal.pbio.3003650. URL https://doi.org/10.1371/journal.pbio.3003650.

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.

Sushil Bikhchandani, David Hirshleifer, and Ivo Welch. A theory of fads, fashion, custom, and cultural change as informational cascades. *Journal of political Economy*, 100(5):992–1026, 1992.

Joydeep Biswas and Laurent Charlin. NeurIPS 2026 AI reviewing experiment, 2026. URL https://neurips.cc/Conferences/2026/ai-reviewing-experiment. Conference webpage. https://neurips.cc/Conferences/2026/ai-reviewing-experiment.

Joydeep Biswas and Gaurav Sahu. EMNLP 2026 AI reviewing experiment, 2026. URL https://2026.emnlp.org/ai-reviewing-experiment/. Conference webpage. https://2026.emnlp.org/ai-reviewing-experiment/.

Anne Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43(4):783–791, 1987. doi: 10.2307/2531532.

Anne Chao. Species estimation and applications. In Samuel Kotz, N. Balakrishnan, Campbell B. Read, and Brani Vidakovic (eds.), *Encyclopedia of Statistical Sciences*, volume 12, pp. 7907–7916. Wiley, New York, 2005. doi: 10.1002/0471667196.ess5051.

Chicago Human+AI Lab. OpenAIReview, 2026. URL https://github.com/ChicagoHAI/OpenAIReview. GitHub repository.

Stephen Cole, Jonathan R Cole, and Gary A Simon. Chance and consensus in peer review. *Science*, 214(4523):881–886, 1981. doi: 10.1126/science.7302566.

Conference on Neural Information Processing Systems. NeurIPS 2021 paper checklist guidelines, 2021a. URL https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist. Conference webpage. https://neurips.cc/Conferences/2021/PaperInformation/PaperChecklist.

Conference on Neural Information Processing Systems. NeurIPS 2021 paper FAQ, 2021b. URL https://neurips.cc/Conferences/2021/PaperInformation/NeurIPS-FAQ. Conference webpage. https://neurips.cc/Conferences/2021/PaperInformation/NeurIPS-FAQ.

Gordon V. Cormack and Maura R. Grossman. Quantifying bias and variance of system rankings. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1089–1092, Paris, France, 2019. Association for Computing Machinery. doi: 10.1145/3331184.3331356.

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

Ziyan Cui, Ning Li, and Huaikang Zhou. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 5(8): 627–634, 2025.

Hexuan Deng, Xiaopeng Ke, Yichen Li, Ruina Hu, Dehao Huang, Derek F Wong, Yue Wang, Xuebo Liu, and Min Zhang. Cocoreviewbench: A completeness- and correctness-oriented benchmark for ai reviewers. *arXiv preprint arXiv:2605.07905*, 2026. URL https://arxiv.org/abs/2605.07905. ICML 2026.

MD Sukru Mehmet Erturk and MD Mustafa Durmaz. Large language models as peer reviewers: prompt sensitivity and model-dependent reproducibility. *Academic Radiology*, 33(9):3642–3650, 2026.

Haishuo Fang, Yue Feng, and Iryna Gurevych. From passive generation to investigation: A proactive scientific peer review agent, 2026. URL https://arxiv.org/abs/2606.13349.

Xian Gao, Jiacheng Ruan, Zongyun Zhang, Jingsheng Gao, Ting Liu, and Yuzhuo Fu. Reviewagents: Bridging the gap between human and AI-generated paper reviews, 2025. URL https://arxiv.org/abs/2503.08506.

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

Yixing Jiang and Andrew Ng. Tech overview: Stanford agentic reviewer, 2025. URL https://paperreview.ai/tech-overview. Project webpage. https://paperreview.ai/tech-overview.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with llm agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1208–1226, 2024.

Keigo Kusumegi, Xinyu Yang, Paul Ginsparg, Mathijs de Vaan, Toby Stuart, and Yian Yin. Scientific production in the era of large language models. *Science*, 390(6779):1240–1243, 2025. doi: 10.1126/science.adw3000.

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

Haokun Liu and Chenhao Tan. Ai-assisted reviewing is necessary for avoiding the review death spiral. Preprint. https://openaireview.org/assets/review-death-spiral.pdf, 2026.

Ryan Liu and Nihar B. Shah. ReviewerGPT? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023. URL https://arxiv.org/abs/2306.00622.

Yiling Ma, Yilun Zhao, Sihong Wu, Ziyu Chen, Manasi Patwardhan, and Arman Cohan. Actreview: Rebuttal-guided training data and rubric rewards for actionable peer review generation, 2026. URL https://arxiv.org/abs/2609.09076.

Sebastian Porsdam Mann, Mateo Aboy, Joel Jiehao Seah, Zhicheng Lin, Xufei Luo, Daniel Rodger, Hazem Zohny, Timo Minssen, Julian Savulescu, and Brian D Earp. Ai and the future of academic peer review. *arXiv preprint arXiv:2509.14189*, 2025.

Nancy E Mayo, James Brophy, Mark S Goldberg, Marina B Klein, Sydney Miller, Robert W Platt, and Judith Ritchie. Peering at peer review revealed high degree of chance associated with funding of grant applications. *Journal of Clinical Epidemiology*, 59(8):842–848, 2006. doi: 10.1016/j.jclinepi.2005.12.007.

Dang Nguyen, Wanqing Hao, Yanai Elazar, and Chenhao Tan. Benchmarking agentic review systems, 2026. URL https://arxiv.org/abs/2606.19749.

Reviewer3. Reviewer3: AI peer review for research papers, 2026. URL https://reviewer3.com/. Accessed September 2026. https://reviewer3.com/.

Samarth P, Sai Sathvik Motamarri, Vyoman Jain, and Shiva Golugula. ReviewArena: A large-scale cross-conference dataset and benchmark for LLM peer review, 2026. URL https://openreview.net/forum?id=yugEO52gkR. ICML 2026 Workshop on AI for Science, Spotlight. OpenReview.

Nihar B. Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022. doi: 10.1145/3528086.

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/.

Pannaga Shivaswamy and Ashok Chandrashekar. Bias-variance decomposition for ranking. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 472–480. Association for Computing Machinery, 2021. doi: 10.1145/3437963.3441772.

Garold Stasser and William Titus. Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of personality and social psychology*, 48(6):1467, 1985.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can LLM feedback enhance review quality? a randomized study of 20K reviews at ICLR 2025, 2025. URL https://arxiv.org/abs/2504.09737.

Qiyao Wei, Samuel Holt, Jing Yang, Markus Wulfmeier, and Mihaela van der Schaar. Position: The ML community must build an AI-augmented peer-review ecosystem, 2026. URL https://arxiv.org/abs/2506.08134. ICML 2026 Position Paper Track (Oral).

Yixuan Weng, Minjun Zhu, Qiujie Xie, Zhiyuan Ning, Shichen Li, Panzhong Lu, Zhen Lin, Enhao Gu, Qiyao Sun, and Yue Zhang. Deepreviewer 2.0: A traceable agentic system for auditable scientific peer review, 2026. URL https://arxiv.org/abs/2604.09590.

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.

A Full assumptions and proof for Theorem 3.1

Let Q be the target distribution of papers and P_i the eligible-reviewer distribution for paper i. Fix the panel size $m \geq 1$. Let $\mathcal{C}_{\text{ver}}$ be a finite, nonempty space of verifiable atomic claims, and let $\mathcal{G}$ be a fixed, finite, nonempty family of candidate checklist mappings. Each $g \in \mathcal{G}$ maps a claim to its cluster in a partition $\mathcal{S}_g$ of $\mathcal{C}_{\text{ver}}$: $g(c) = s$ when $c \in s \in \mathcal{S}_g$. Each cluster s is one checklist item, also called a signal.

$$\mathcal{E} = \bigcup_{g \in \mathcal{G}} \mathcal{S}_g, \qquad \mathcal{D} = \{T : T \subseteq \mathcal{S}_g \text{ for some } g \in \mathcal{G}\}.$$

Here $\mathcal{E}$ contains all candidate items, and $\mathcal{D}$ contains all candidate checklists and their partial versions. For $T \in \mathcal{D}$, define

$$\mathcal{C}_T = \bigcup_{s \in T} s, \qquad g_T : \mathcal{C}_T \to T, \qquad g_T(c) = s \quad \text{for the unique } s \in T \text{ containing } c.$$

The partial mapping g_T surfaces only signals in T under the same rule as the original g. In particular, $g_{\mathcal{S}_g} = g$ and $g_\varnothing = \varnothing$ is the empty mapping representing unaided review.

For a mapping h, let $Y_{ir}(h) \in \mathbb{R}$ be reviewer r's actual score on paper i. Draw $r_1, \ldots, r_m$ independently from P_i and define

$$\bar{Y}_{im}(h) = \frac{1}{m}\sum_{j=1}^{m} Y_{ir_j}(h), \qquad \mu_i(\varnothing) = \mathbb{E}_{r\sim P_i}[Y_{ir}(\varnothing) \mid i].$$

The target $\mu_i(\varnothing)$ is the unaided population mean. Define

$$\begin{aligned} \mathrm{MSE}_i(h,m) &= \mathbb{E}[(\bar{Y}_{im}(h) - \mu_i(\varnothing))^2 \mid i], \\ R_m(h) &= \mathbb{E}_{i\sim Q}[\mathrm{MSE}_i(h,m)]. \end{aligned} \tag{7}$$

For fixed m, define the benefit and marginal benefit as set functions:

$$B(T) = R_m(\varnothing) - R_m(g_T), \qquad B(s \mid T) = B(T\cup\{s\}) - B(T). \tag{8}$$

The marginal is defined for $s \notin T$ whenever $T \cup \{s\} \in \mathcal{D}$. Thus $B(\varnothing) = 0$, and $B(\mathcal{S}_g) = R_m(\varnothing) - R_m(g)$. Note that the MSE and R_m take mappings such as h and g as arguments, while B takes signal sets.

Signals and agreement in historical (unaided) reviews. Let C_{ir} be the atomic claims in reviewer r's unaided review of paper i, and let $\bar{n} = \mathbb{E}_{i\sim Q}\mathbb{E}_{r\sim P_i}|C_{ir}\cap\mathcal{C}_{\mathrm{ver}}| \in (0,\infty)$ be the expected number of verifiable claims per review. Every candidate partitions the same space $\mathcal{C}_{\mathrm{ver}}$, so $\bar{n}$ does not depend on g. For each signal $s \in \mathcal{E}$, put

$$\begin{aligned} Z_{ir}(s) &= \mathbf{1}[C_{ir}\cap s \neq \varnothing], \qquad p_s(i) = \Pr(Z_{ir}(s) = 1 \mid i), \\ \nu_s(i) &= \frac{\mathbb{E}[|C_{ir}\cap s| \mid i]}{\bar{n}}, \qquad d_s = \mathrm{Cov}_{i\sim Q}\big(\nu_s(i), p_s(i)\big). \end{aligned}$$

Here $\mathbf{1}[\cdot]$ is the indicator function, $p_s(i)$ is the mention probability under P_i, and $\nu_s(i)$ is the expected number of claims a review of paper i makes in s, in units of the average review length. The quantity d_s measures whether the papers on which reviewers write more about s are also the papers on which other reviewers raise s. For a candidate checklist g, restrict these indicators to $s \in \mathcal{S}_g$ to obtain the review vector Z^g_{ir}. As in the main text, $A_{\mathrm{same}}(g)$ is the share of review claims whose signal an independent review of the same paper also raises, and $A_{\mathrm{diff}}(g)$ is that share when the second review is of an independently drawn paper. The selection objective is $\Delta(g) = A_{\mathrm{same}}(g) - A_{\mathrm{diff}}(g)$.

Assumptions. The theorem uses the following four conditions.

A1. Representative sampling. We assume historical reviews represent the target paper and reviewer populations, review pairs are sampled independently conditional on the paper. Explicitly,

$$A_{\mathrm{same}}(g) = \frac{1}{\bar{n}}\,\mathbb{E}_{i\sim Q}\mathbb{E}_{r,r'\overset{\text{i.i.d.}}{\sim} P_i}\left[\sum_{c\in C_{ir}\cap\mathcal{C}_{\mathrm{ver}}} Z_{ir'}(g(c))\,\middle|\, i\right],$$

$$A_{\mathrm{diff}}(g) = \frac{1}{\bar{n}}\,\mathbb{E}_{i,j\overset{\text{i.i.d.}}{\sim} Q}\mathbb{E}_{r\sim P_i,\, r'\sim P_j}\left[\sum_{c\in C_{ir}\cap\mathcal{C}_{\mathrm{ver}}} Z_{jr'}(g(c))\,\middle|\, i,j\right].$$

Under this condition, ECA decomposes over signals, $\Delta(g) = A_{\mathrm{same}}(g) - A_{\mathrm{diff}}(g) = \sum_{s\in\mathcal{S}_g} d_s$, as established in the proof.

A2. Positivity. We assume every signal has $d_s \geq 0$: papers on which reviewers write more about s are not papers on which other reviewers are less likely to raise it. This holds, for example, if a review that raises s makes the same expected number λ_s of claims in s on every paper, since then $\nu_s(i) = \lambda_s p_s(i)/\bar{n}$ and $d_s = (\lambda_s/\bar{n})\,\mathrm{Var}_{i\sim Q}(p_s(i))$. We further assume signals whose contribution to ECA is positive ($d_s > 0$) provide positive stand-alone MSE benefits, while signals with $d_s = 0$ provide zero stand-alone benefit. Every signal with $d_s > 0$ has $B(\{s\}) > 0$, and every signal with $d_s = 0$ has $B(\{s\}) = 0$. Let $\mathcal{E}_+ = \{s \in \mathcal{E} : d_s > 0\}$. Define

$$\beta_s = \frac{B(\{s\})}{d_s} \quad (s\in\mathcal{E}_+), \qquad \rho = \max_{s,t\in\mathcal{E}_+}\frac{\beta_s}{\beta_t} = \frac{\max_{s\in\mathcal{E}_+}\beta_s}{\min_{s\in\mathcal{E}_+}\beta_s}. \tag{9}$$

The quantity β_s is the MSE benefit from signal s alone per unit of ECA. The ratio $\rho \geq 1$ measures the largest relative variation in this quantity across signals.

A3. Monotonicity and submodularity. We assume adding a signal cannot increase MSE (monotonicity), and its marginal MSE reduction cannot grow as more signals are provided (submodularity). For $T \subseteq U$, $s \notin U$, and $U \cup \{s\} \in \mathcal{D}$,

$$B(s \mid T) \geq B(s \mid U) \geq 0. \tag{10}$$

This condition compares compatible subsets of candidate checklists. In particular, $0 \leq B(s \mid T) \leq B(\{s\})$, so signals that have zero stand-alone benefit have zero marginal benefit in every compatible context.

A4. Limited redundancy. We assume signals in the selected checklist have a positive fraction of their stand-alone benefit on average when each is added after all the others. For a candidate item set S and $s \in S$, define its leave-one-out contribution

$$\delta_s(S) = B(S) - B(S \setminus \{s\}) = B(s \mid S \setminus \{s\}).$$

For $B(\{s\}) > 0$, the fractional loss $1 - \delta_s(S)/B(\{s\}) \in [0, 1]$ measures how much of signal s's stand-alone benefit is redundant when all other signals in S are present. For $W(S) = \sum_{s \in S} B(\{s\}) > 0$, define the weighted average loss

$$\begin{aligned} \bar{\kappa}(S) &= \sum_{s \in S \cap \mathcal{E}_+} \frac{B(\{s\})}{W(S)} \left(1 - \frac{\delta_s(S)}{B(\{s\})}\right) \\ &= 1 - \frac{\sum_{s \in S} \delta_s(S)}{W(S)} \in [0, 1]. \end{aligned} \tag{11}$$

Zero-benefit signals carry zero weight and have $\delta_s(S) = 0$ by A3. For the signal set $S^* = \mathcal{S}_{g^*}$ of the selected checklist g^* in (6), we require

$$\bar{\kappa}(S^*) \leq \kappa, \qquad \kappa \in [0, 1). \tag{12}$$

These are conditions on actual reviewer behavior. Factors (including ρ and κ) that give a useful bound in Theorem 3.1 must be supported by assisted-review outcomes.

The following restates the main-text result using the definitions above.

Theorem A.1. *Assume A1–A4. Let*

$$g^* \in \arg\max_{g \in \mathcal{G}} \Delta(g), \qquad g_{\text{opt}} \in \arg\min_{g \in \mathcal{G}} R_m(g).$$

Then

$$B(\mathcal{S}_{g^*}) \geq \alpha B(\mathcal{S}_{g_{\text{opt}}}), \qquad \alpha = \frac{1 - \kappa}{\rho} \in (0, 1]. \tag{13}$$

Equivalently,

$$R_m(g^*) \leq (1 - \alpha) R_m(\varnothing) + \alpha R_m(g_{\text{opt}}). \tag{14}$$

Proof. Because $\mathcal{S}_g$ partitions $\mathcal{C}_{\text{ver}}$, the claim sum splits by signal: $\sum_{c \in C_{ir} \cap \mathcal{C}_{\text{ver}}} Z(g(c)) = \sum_{s \in \mathcal{S}_g} |C_{ir} \cap s|\, Z(s)$ for any review vector Z. By A1, the two reviews are independent given their papers, so the same-paper term for signal s conditional on paper i is $\mathbb{E}[|C_{ir} \cap s| \mid i]\, p_s(i) = \bar{n}\, \nu_s(i) p_s(i)$. When the second review is of an independently drawn paper, it is $\bar{n}\, \mathbb{E}_{i \sim Q}[\nu_s(i)]\, \mathbb{E}_{i \sim Q}[p_s(i)]$. Hence

$$\Delta(g) = \sum_{s \in \mathcal{S}_g} \{\mathbb{E}_{i \sim Q}[\nu_s(i) p_s(i)] - \mathbb{E}_{i \sim Q}[\nu_s(i)]\, \mathbb{E}_{i \sim Q}[p_s(i)]\} = \sum_{s \in \mathcal{S}_g} d_s. \tag{15}$$

For a signal set $S = \mathcal{S}_g$, A3 implies, for every $s \in S$ and $T \subseteq S \setminus \{s\}$,

$$\delta_s(S) \leq B(s \mid T) \leq B(\{s\}).$$

Telescoping along any ordering of S gives

$$\sum_{s \in S} \delta_s(S) \leq B(S) \leq W(S). \tag{16}$$

Thus A4 gives the lower bound needed only for the selected checklist:

$$B(S^*) \geq (1 - \bar{\kappa}(S^*)) W(S^*) \geq (1 - \kappa) W(S^*).$$

For the comparison across signals, let $\beta_{\min} = \min_{s \in \mathcal{E}_+} \beta_s > 0$. This is a derived scale, not an additional assumption. By the definition of ρ, every signal with $d_s > 0$ has $B(\{s\}) = \beta_s d_s$ with $\beta_s \in [\beta_{\min}, \rho\beta_{\min}]$. By A2, no signal has $d_s < 0$, and signals with $d_s = 0$ have zero singleton benefit. Therefore

$$\beta_{\min}\Delta(g) \le W(\mathcal{S}_g) \le \rho\beta_{\min}\Delta(g).$$

Combining this with (16) and the optimality of g^* for ECA yields

$$\begin{aligned} B(\mathcal{S}_{g^*}) &\ge (1-\kappa)\beta_{\min}\Delta(g^*) \\ &\ge (1-\kappa)\beta_{\min}\Delta(g_{\text{opt}}) \\ &\ge \frac{1-\kappa}{\rho} B(\mathcal{S}_{g_{\text{opt}}}). \end{aligned}$$

The scale $\beta_{\min}$ cancels, leaving only the relative ratio ρ. Substituting $B(\mathcal{S}_g) = R_m(\varnothing) - R_m(g)$ proves the MSE formulation. □

If $\mathcal{E}_+ = \varnothing$, A2 gives zero singleton benefits and A3 forces every feasible marginal benefit to be zero. Every candidate then has $B(\mathcal{S}_g) = 0$, and the result is trivial; one may set $\rho = 1$ by convention.

Interpreting the approximation factor. The two factors in $\alpha = (1-\kappa)/\rho$ describe different possible losses. The factor $1/\rho$ accounts for how much a unit of ECA can differ in stand-alone MSE value across signals. The factor $1-\kappa$ lower-bounds the fraction of total stand-alone benefit retained by the selected checklist. For example, if these values differ by at most a factor 1.25 ($\rho = 1.25$) and the average diminishing value of signals is at most 20% ($\kappa = 0.2$), the selected checklist achieves at least $0.8/1.25 = 0.64$ of the best attainable MSE reduction. A4 constrains only the average loss, so individual signals in the selected checklist may still have small marginal contributions.

When $\rho = 1$, every signal has the same benefit per unit of ECA. When $\kappa = 0$, the selected checklist retains all its stand-alone benefit. If both hold, maximizing ECA is MSE-optimal within $\mathcal{G}$. More generally, the factor stays bounded away from zero when ρ is uniformly bounded and the selected checklist's average redundancy is uniformly below one.

A sharper finite-size bound. Let $k = \max_{g \in \mathcal{G}} |\mathcal{S}_g|$. Under the same assumptions, the theorem also holds with

$$\alpha_k = \frac{1 - \kappa + \kappa/k}{\rho} \ge \alpha. \tag{17}$$

To see this, let $S = S^*$ and $n = |S| \le k$. For each $s \in S$, telescope along an ordering that puts s first. By A3, every later signal t adds at least $\delta_t(S)$, so $B(S) \ge B(\{s\}) + \sum_{t \in S \setminus \{s\}} \delta_t(S)$. Averaging over the n choices of s gives

$$B(S) \ge \frac{W(S)}{n} + \left(1 - \frac{1}{n}\right) \sum_{t \in S} \delta_t(S) = \left(1 - \bar{\kappa}(S) + \frac{\bar{\kappa}(S)}{n}\right) W(S) \ge \left(1 - \kappa + \frac{\kappa}{k}\right) W(S).$$

Using this bound in place of $B(S^*) \ge (1-\kappa)W(S^*)$ in the proof gives α_k. This improves $\alpha = (1-\kappa)/\rho$ used by the main-text theorem and its restatement above.

Empirical selection. Let $\widehat{\Delta}$ estimate population agreement with uniform error bound

$$\sup_{g \in \mathcal{G}} |\widehat{\Delta}(g) - \Delta(g)| \le \varepsilon, \qquad \varepsilon \ge 0.$$

If $\widehat{g}$ maximizes $\widehat{\Delta}$, then $\Delta(\widehat{g}) \ge \Delta(g_{\text{opt}}) - 2\varepsilon$. If A4 also holds with $\mathcal{S}_{\widehat{g}}$ in place of S^*, the same proof gives

$$B(\mathcal{S}_{\widehat{g}}) \ge \alpha B(\mathcal{S}_{g_{\text{opt}}}) - 2(1-\kappa)\beta_{\min}\varepsilon.$$

Here $\beta_{\min}$, defined in the proof, converts agreement to squared-score units.

Remarks. The result concerns the actual benefit B, including reviewer biases. Submodularity, positivity, and limited redundancy may fail under different reasoning, attention overload, or systematic misinterpretation. Our theorem does not derive these conditions from the rationality of reviewers.

B CLAIM-LEVEL ANALYSIS OF THE NEURIPS 2021 REVIEWS

This appendix supports Section 3 by answering four questions about the NeurIPS 2021 dual-committee corpus: what share of a paper's issues a single reviewer raises (Appendix B.2), how much two committees agree (Appendix B.3), whether their criticisms are correct (Appendix B.4), and what the area chair keeps (Appendix B.5). All intervals are 95% bootstrap confidence intervals over papers.

B.1 DATA AND DEFINITIONS

Corpus. We use the public OpenReview subset of the NeurIPS 2021 consistency experiment (Beygelzimer et al., 2023): 298 papers, each reviewed by two committees that did not communicate, with 2,319 reviews and 596 area-chair (AC) decision notes in total. Papers have six to eleven reviewers, typically seven or eight.

Claims and layers. An LLM splits every review and AC note into atomic *claims*. An atomic claim is an independent, self-contained comment about one problem or merit of the paper. An LLM then rewrites each claim as a statement that can be checked against the paper (Appendix C.1). This gives 25,868 claims: 23,920 from reviewers and 1,948 from ACs. Each claim gets three labels: what kind of statement it is (fact, evaluation, request, quote, reference, or non-argument, following Hua et al. (2019)), what would settle it (the paper, the literature, the world, or nothing), and whether it assumes the paper is missing something. The first two labels sort claims into three *layers*: *Concrete* claims are facts or requests that the paper itself can settle. *Conceptual* claims are evaluations. *Out-of-scope* claims can only be checked against other work or the world. For example, "the experimental evaluation omits some important baselines such as DISCERN" is concrete, and "the techniques proposed in this paper look incremental" is conceptual. Quotes, references, and non-arguments are dropped, because they do not assess the paper.

Matches and issues. An LLM labels pairs of claims in two settings: strict and loose (Appendix C.2); 38,038 pairs were labelled. A *strict* match is an exact restatement, for example "it would be helpful to have a separate discussion on the limitations of the model" and "the authors don't discuss the limitations of the model in any detail". A *loose* match also allows the same point at a different level of detail, for example "only evaluated over one training dataset and one task" and "the paper would be strengthened by including other tasks, such as defect detection".

Within each reviewer and layer, strictly matching claims are merged, so a point the reviewer repeats counts once. Between reviewers of the same paper, and between the two committees, claims are compared across layers, and a loose match counts as the same point; a concrete claim can therefore match a conceptual one. Loosely matching claims form an *issue*, a problem that has been described from different angles in reviews. Loose matches can chain several claims together, so when a group links more than two claims, an LLM examines the whole group and splits it if its claims do not raise a single issue (Appendix C.2).

B.2 WHAT SHARE OF A PAPER'S ISSUES DOES A SINGLE REVIEWER RAISE?

For each paper, we merge the claims of all its reviewers into issues. Their union is the paper's observed issue set. We then estimate the richness of the pool those reviews are drawn from: how many distinct issues would appear if further reviewers were drawn in the same way. The estimator is bias-corrected Chao2 (Chao, 1987; 2005). It treats issues like species and reviewers like sampling occasions. We only record whether each reviewer raised an issue, so these are incidence data, the setting Chao2 is built for: if f_1 issues were raised by exactly one of the paper's m reviewers and f_2 by exactly two, the estimated richness is $S_{\text{obs}} + \frac{m-1}{m} f_1(f_1-1)/\big(2(f_2+1)\big)$. A large singleton count means many issues in this pool are still unseen. The total contains the issues this sampling process would surface. Points that reviewers in this pool would not raise are outside it.

Table 2 gives the shares. A paper's reviewers raise 27 distinct issues between them, about five per reviewer. One reviewer covers a fifth of the observed set, and a committee of three or four covers a little over half. Each added reviewer keeps finding new issues: n random reviewers cover 20, 35, 48, 60, 72, and 83% of the set for $n = 1$ to 6, and 81% of issues are raised by only one reviewer. The Chao2 estimate puts this pool at about 90 issues for a typical paper (the median; the mean of 124 is

pulled up by a few papers). Of that pool, one reviewer raises about 6%, a committee 18%, and all of a paper's reviewers together 31%.

Table 2: Share of a paper's review issues raised by one reader. Observed issues are the union over the paper's reviewers. The estimated total is the Chao2 richness of issues that further draws from the same reviewer pool would raise. Means over 298 papers.

Reader	Of observed issues (%)	Of Chao2 richness (%)
One reviewer	19.7 [19.3, 20.2]	6.4 [5.9, 6.9]
One committee (3 to 4 reviewers)	57.3 [56.8, 57.7]	18.1 [16.9, 19.2]
All of a paper's reviewers	100 by construction	30.9 [29.2, 32.6]

B.3 Do two committees raise the same issues?

A claim *overlaps* if the other committee has a claim that loosely matches it, in any layer. Table 3 gives the overlap rates. One reviewer claim in eight is restated exactly by the other committee, and one in three is raised at some level of detail. Judgements overlap more than specifics: the loose overlap is 40.4% for conceptual claims and 30.9% for concrete ones. AC claims overlap with the other committee's reviews at the same rate as reviewer claims (35.3% versus 34.3%). This is the lottery term of (1) at the level of claims: two committees of qualified reviewers mostly do not raise the same points.

Table 3: Cross-committee overlap: share of claims with a matching claim in the other committee. The denominator is all claims; a match the LLM misses counts as no overlap.

Claims compared	n	Strict (%)	Loose (%)
Reviewer claims vs. other committee's reviews	23,920	11.9 [11.1, 12.7]	34.3 [32.8, 35.8]
AC claims vs. other committee's reviews	1,948	11.7 [10.1, 13.3]	35.3 [32.4, 38.2]

B.4 Are the criticisms right?

For 101 papers, Grok 4.5 checked every checkable claim against the manuscript (Appendix C.3). The check asks whether the paper itself bears a claim out, not whether the claim is important or fair. A claim is *refuted* only when the checker finds evidence against it in the paper: it is asked to quote at least eight consecutive words from the paper, and we check automatically that the quote appears there, downgrading the verdict if it does not. Any claim can be refuted this way. A reviewer may say the paper omits something that it in fact contains, or may misstate what the paper reports or shows; in both cases the verdict rests on a passage of the paper. A typical refutation: a reviewer wrote "I was surprised to see no mention of de Haan et al., Causal confusion in imitation learning", and the paper says "de Haan et al. (2019) explore the case when expert and imitator can observe the same contexts, but the causal diagram is not available." A claim is *partial* when the paper has related material that does not fully settle it, *unsettleable* when the paper's text cannot settle it, and *unclear* when the checker cannot tell. Claims that only the literature or the world can settle could also be checked against web sources; 37 refutations, none of them concrete, rest on such a source.

The opposite verdict is narrower. A claim is *upheld* when it says the paper lacks something and the checker searched for it under at least three terms without finding it (`criticism_valid` in Appendix C.3). A search can confirm that something is absent, but not that a table entry is wrong or that a proof step does not follow, so a correct criticism of that kind ends up among the partial or unclear verdicts. The upheld rate therefore undercounts correct criticisms, while every claim can be refuted. We report refuted claims as a share of all judged claims. This rate is conservative: a claim counts as wrong only when the checker produces a quote against it, and partial, unsettleable, and unclear verdicts are not counted as wrong.

The manuscript in this check is the camera-ready PDF linked from the paper's public OpenReview forum, not the anonymized submission the reviewers read. NeurIPS 2021 did not accept a revised

PDF during review, but authors of an accepted paper could revise the manuscript for the camera-ready deadline (Conference on Neural Information Processing Systems, 2021b). Our corpus is the public subset of accepted duplicated papers, and the submitted PDFs are not in that release (Beygelzimer et al., 2023). A criticism that the submission omitted a citation, baseline, or experiment can therefore be marked refuted because the authors added the missing material afterward. Of the 1,164 refuted claims, 89.1% assume the paper is missing something, the kind of claim most exposed to this. The refuted rates below can be higher than the rate at which the submission itself contradicted the reviewer. This works against the conservative counting rule above, and without the submitted PDFs we cannot tell which effect is larger. The de Haan sentence above is quoted from the camera-ready PDF, and we cannot tell whether it was already in the submission.

Table 4 gives the verdicts by layer. Concrete claims are the ones the checker can really decide, and the paper refutes 18.2% of all judged concrete claims. Counting only clear verdicts (upheld or refuted), reviewers and ACs are refuted at the same rate: 34.2% and 34.3% across all layers. Agreement does not help: claims that the other committee also raised are upheld and refuted at the same rates as claims it did not raise, within 0.8 points and with overlapping intervals.

To check the checker, we took 25 sentences from a paper A, turned each into a criticism that A omits it, and checked each criticism against A, where it should be refuted, and against an unrelated paper B, where it should be upheld. The checker matched the answer key on 48 of 50 verdicts (96%, Wilson 95% CI [86.5, 98.9]).

Table 4: Manuscript check of criticisms (101 papers). Upheld and refuted are shares of the judged claims in each layer; the rest are partial, unsettleable, or unclear.

Layer	n	Upheld (%)	Refuted (%)
Concrete	5,527	34.1 [31.5, 36.8]	18.2 [16.3, 20.2]
Conceptual	1,776	13.9 [11.4, 16.9]	6.8 [5.3, 8.4]
Out of scope	444	24.3 [19.2, 29.6]	8.8 [5.5, 12.2]
All judged	7,747	28.9 [26.6, 31.4]	15.0 [13.5, 16.6]

B.5 What does the area chair keep?

Each AC read only its own committee's reviews and then wrote the decision note, so we can see what survives from the reviews to the note. We look in both directions. Forward, an AC claim is *traced* if a review the AC read contains the same issue. Backward, a reviewer issue is *kept* if the AC note contains it. Among the 482 committees whose note contains at least one claim, a note has four claims on average and the reviewers raise 18 issues.

Where AC claims come from. Table 5 gives the forward direction. Just over half of AC claims can be traced to a review (54.6% loose, 22.9% strict). A traced claim usually restates a single reviewer: 59% match one reviewer, 29% two, and 12% three or more. Among the AC claims the manuscript check could judge, 16% of traced claims are refuted, against 23% of the AC's own additions, with overlapping intervals.

What reaches the note. Whether an issue reaches the note depends on how many reviewers raised it: 5.7% [5.1, 6.4] of the 7,607 issues raised by one reviewer reach the note, 27.1% [24.0, 30.6] of the 783 raised by two, and 56.5% [51.2, 61.7] of the 416 raised by three or more. Whether the issue is correct makes no difference. Among single-reviewer issues, upheld, partial, and refuted issues are each kept about 5% of the time; among multi-reviewer issues, 30 to 41% each, with overlapping intervals. The AC filters by consensus, not by truth.

C Pipeline details

This appendix gives the prompts and settings behind the pipeline in Section 4 and the scoring study in Section 5; further results of the scoring study are in Appendix D. All LLM steps use Grok 4.5 through Cursor, except the simulated reviewers in Section 5, which use Composer 2.5. All random seeds are 42.

Table 5: Where AC claims come from: share of AC claims with a matching claim in the reviews of the AC's own committee.

AC claims	n	Traced, strict (%)	Traced, loose (%)
All	1,948	22.9 [20.3, 25.5]	54.6 [51.1, 57.7]
Concrete	1,108	21.9	54.9 [50.4, 59.1]
Conceptual	550	31.8	58.2 [53.9, 62.8]
Out of scope	290	9.7	46.6 [40.4, 53.8]

C.1 Claim extraction

Extraction has three passes. The first cuts a review into spans, the second rewrites each span as a statement about the paper that can be checked, and the third labels each span.

Pass 1: segmentation. We split each review field at line breaks and give the model one block at a time. The model must copy each span word for word, and any span that does not appear in the block is thrown away.

```
Split this block of a peer review into propositions. Do not judge them.

A proposition is the smallest span that makes one point about one topic.

- Copy each span verbatim. Do not fix typos or paraphrase.
- Keep spans in reading order, without overlap.
- Cover the whole block except greetings and section headers.
- Do not filter or cap the number of spans. Keep praise, questions, and
  typo reports too.
- Return an empty list if the block has no propositions.
```

Pass 2: propositionization. The second pass rewrites each span as a statement that a reader could mark true or false against the paper, which makes claims easier to understand and to verify. For example, "The authors should add an ablation of the KD loss" becomes "the paper reports an ablation of the knowledge-distillation loss". A pure judgement such as "the discussion is too vague" has no such statement and becomes null.

```
Rewrite each proposition as one statement about the paper that a reader could
mark true or false, or null. Point it at the paper, not at the authors.

  "The authors should add an ablation of the KD loss"
  -> "The paper reports an ablation of the knowledge-distillation loss."
  "How did you choose these learning rates?"
  -> "The paper explains how the learning rates were chosen."
  "The discussion is too vague"
  -> null

Almost every request or question has a checkable core ("the paper contains /
reports / explains X"). Use null only for a pure judgement.
```

Pass 3: typing. The third pass gives each span three labels: what kind of statement it is (a fact, an evaluation, a request, or a quote, reference, or non-argument), what would settle it (the paper, other literature, the world, or nothing), and whether it assumes the paper is missing something. The first two labels decide the layer used in Appendix B.1. Evaluations are conceptual. Facts and requests that the paper can settle are concrete. Facts and requests that only other literature or the world can settle are out of scope. Quotes, references, and non-arguments are dropped.

```
proposition_type: fact | evaluation | request | quote | reference | non_arg.
  fact: a checkable statement about the paper or the world.
  evaluation: a judgement ("not novel enough", "reads well").
  request: asks for an action or asks a question.
  If a span is both fact and evaluation, choose fact.
```

```
presupposes_absence: true if the span assumes the paper lacks something.
  "Please also evaluate with AutoAttack"      -> true
  "Why is SAM a second-order method?"         -> false
  "Table 4 shows accuracy below the baseline" -> false

truthmaker: what settles it. paper | literature | world | none.
  "The paper does not cite Gulcehre et al. 2021" is "paper": the citation
  list is in the paper.

Label what the words say, not what the reviewer might have meant.
```

C.2 SAME-ISSUE PAIR LABELLING

Labels are `exactly_the_same`, `a_more_specific`, `b_more_specific`, or `partially_related`. Loose overlap is any of the first three. Pairs the model does not write are unrelated.

The claim-level analysis in Appendix B uses these labels in two ways. Within each reviewer and layer, only `exactly_the_same` pairs are merged. Between reviewers, and between the two committees, any loose overlap counts as the same point, including across layers.

```
Two jobs. Write only pairs that overlap. Judge each pair on its own.

Labels: exactly_the_same | a_more_specific | b_more_specific |
        partially_related.
Loose overlap is any of the first three.

1. Within one reviewer and one layer, judge that reviewer's pairs.
   Merge only exactly_the_same.
2. Between reviewers, and between the two committees, judge every pair
   on the paper, including across layers. Loose overlap is the same point.

Sharing a topic is not enough. Different defects of the same kind (both
about writing) are at most partially_related.
```

When loose matches link more than two claims from different reviewers into one group, an LLM judges the whole group with the prompt below and splits it when its claims do not raise a single issue.

```
You see a group of review claims about one paper. Loose matches between
pairs of claims linked them into this group. Decide whether every claim
in the group raises the same issue about the paper.

A claim may state the issue broadly or in more detail. That is fine.
Sharing a topic is not enough: two different defects of the same kind
(both about baselines, or both about writing) are different issues.

Output:
- same_issue: true if all claims raise one issue, false otherwise.
- If false, split the claims into subgroups that each raise one issue.
  Every claim goes into exactly one subgroup. If a broad claim fits more
  than one subgroup, put it with the subgroup it fits best.
```

C.3 MANUSCRIPT CHECK

The checker receives three things: the reviewer's original sentence, its Pass 2 statement, and the full manuscript. The manuscript is the camera-ready PDF on the paper's OpenReview forum. Appendix B.4 explains why checking that version can inflate the refuted rate. It may call a criticism refuted only if it quotes at least eight consecutive words from the paper as evidence. We check the quote automatically and downgrade the verdict if the quote is not in the paper. It may call a criticism valid only if it lists at least three terms it searched for.

```
Judge the criticism. Use the statement to know what to look for: "you never
explain how T is chosen" becomes "The paper explains how T is chosen".

Verdicts:
```

```
- criticism_refuted: the paper contains what the reviewer said was missing,
  or contradicts the reviewer.
- criticism_partial: the paper has something related but not a full answer.
- criticism_valid: you searched and the paper does not contain it.
- out_of_scope: the paper's text cannot settle this.
- unclear: you could not tell.

Rules:
1. refuted requires a verbatim quote of at least 8 words from the paper.
   The quote is checked automatically.
2. valid requires at least three search terms (synonyms, abbreviations,
   section names).
3. When unsure, say unclear. Never guess refuted.
4. A paper that explains something badly still explains it: partial,
   not refuted.
5. Ignore whether the criticism is important or fair. Only whether the paper
   bears it out.
```

The checker also records whether the span is a criticism, a description of the paper, or a question. A description that turns out to be wrong is not counted as a wrong criticism. `criticism_valid` only records that a claimed absence is real. A positive error that the paper does contain, such as a wrong table entry or a proof step that does not follow, receives no upheld label (Appendix B.4).

C.4 Building the hierarchy

Fingerprint stripping. From here on, each claim's text is its Pass 2 statement, or the original span when Pass 2 returned null. Reviewers of different papers describe the same issue with different line numbers and method names, so before embedding we replace these with placeholders: numbered references become `[LINE]`, `[SEC]`, `[FIG]`, `[TAB]`, `[EQ]`, `[ALG]`, or `[PAGE]`, and rare method names become `[METHOD]`.

Embedding. We embed each text with Qwen3-Embedding-0.6B using the instruction below, then reduce it to 128 dimensions with a random projection and normalize it.

```
Instruct: Represent the underlying issue raised by this scientific peer-review
claim. Claims about the same issue should be close even across different papers.
Preserve distinctions between a broad issue and a more specific version of it,
and ignore paper-specific names, numbering, and reviewer phrasing.
Query:
```

Tree. Claims from the papers used in Section 5 are held out from the start of the pipeline, so those papers do not enter the tree. This leaves 15,118 concrete claims. Step (vi) of Section 4 pools all of them and merges claims that an LLM judges to be exactly the same. Candidate pairs link each claim to its three nearest neighbours by cosine similarity, computed on the full Qwen3 embeddings before the 128-dimensional projection, and the LLM judges each candidate pair with the prompt below. This step is separate from the claim-level analysis in Appendix B. Each connected component of merged pairs becomes one point, whose embedding is the mean of its members' embeddings. Step (vii) grows a Ward tree on the embeddings of these points.

```
You see pairs of concrete review claims. The two claims in a pair can come
from different papers. Paper-specific numbers and method names are already
replaced by placeholders such as [METHOD] and [TAB].

For each pair, decide whether the two claims are exactly the same.
- exactly_the_same: the same issue at the same level of detail, so either
  claim could stand in for the other as a checklist item.
- not_same: anything else, including when one claim is broader than the
  other, or when both raise different defects of the same kind (both
  about baselines, or both about writing).

Judge each pair on its own. Sharing a topic is not enough.
```

Cuts. Every checklist ($k \in \{50, 100, 200, 500, 600, 1000, 2000, 6000\}$) is a cut of this one tree, so each coarse item is the union of the finer items under it.

Naming. At each cut, an LLM writes a short question for each cluster, given the cluster's size, the number of papers it draws on, its parent, and a sample of its claims. These questions are the checklist items.

C.5 Recall against verified issues

The recall column in Table 1 asks whether an LLM given the checklist at one cut finds the issues that human reviewers found. The targets are 32 human criticisms, two from each of 16 papers, that the manuscript check upheld. Each cut is a separate run. The LLM receives the paper and that cut's checklist, the k questions, and judges each question on its own. A target is found when the LLM flags a gap on the item that contains it. Recall at cut k is the share of the 32 targets found.

```
Judge every question on this checklist on its own.

For each question, output gap_present:
- true: the paper is missing, weak, wrong, or unclear on this issue.
- false: the paper addresses it, or it does not apply.
Partial coverage is a gap if a substantive part is missing.

Do not read the reviews. Do not use the examples as evidence about this paper.
```

C.6 The scoring study

Each simulated reviewer works in two calls: it first writes its criticisms, then reads them back and gives a score. Splitting the two keeps the model from picking a score first and writing a review to match.

Without checklist. Each reviewer is assigned one branch of the $k{=}50$ cut and ten random $k{=}6{,}000$ items from that branch, and may add up to two issues of its own within the branch. Because the branches do not overlap, neither do the items that different reviewers receive, so each reviewer has its own main focus. The simulation models the first round of review, so branches about rebuttal and revision are skipped (one on NeurIPS, three on ICLR), which leaves one reviewer for each of the $F{=}49$ remaining branches on NeurIPS and $F{=}47$ on ICLR. The average score over all F reviewers is the population target $\hat{\mu}_i(\varnothing)$, and Appendix D.1 shows how closely committees of different sizes approach it.

```
You are a reviewer assigned one k=50 field and 10 random k=6000 claims under it.
Judge only those claims. You may add up to 2 more concrete issues from the
same field.

- Read the whole paper. Stay inside your field.
- Do not read other reviewers or any labels.
- Do not turn a claim into a generic question ("is the paper novel?"). If a
  claim names a method or dataset this paper never uses, answer na.
- verdict: gap (missing, weak, or wrong), ok (handled), or na (does not apply).
- Quote at least one full sentence from the paper for every gap or ok.
```

With checklist. Eight reviewers per paper receive the same checklist, each in a different random order, and may open the finer items under any item. The prompt below is the $k{=}100$ version; the other granularities use the same prompt with their own items.

```
Read the whole paper and the k=100 checklist. Use the finer items under each
item as a guide to what it means.

1. Go through every item. Keep only the ones plausible for this paper.
2. The finer items and other-paper examples explain the item; they are not
   evidence about this paper.
3. Search the whole paper before calling an item missing.
4. Write at most eight concrete criticisms, fewer if fewer are well supported:
   omissions, inconsistencies, unsupported claims, or inadequate experiments
   that the paper text can settle.
5. Quote at least one full sentence from the paper for each criticism.
   Items you skipped are not defects.
```

Estimand. For each paper and condition, we compute variance, bias, and MSE exactly from the observed scores (8 reviewers per checklist condition and F unaided reviewers) and average them over the 30 papers of each venue with equal weight, squaring each paper's bias before averaging.

Scoring. Both conditions use the same scoring prompt, apart from the phrase that names the condition. The checklist version:

```
You already wrote criticisms for this paper. Now give your NeurIPS score (1-10).
Do not write a new review.

- Read the paper and your own review. Do not read other reviewers.
- Score the paper as you would after reading it with that focus. Gaps you
  flagged pull the score down; items you skipped are not defects.
- Give a score only, not an accept/reject decision.

Scale: 10 top 5 percent; 8 accept; 6 marginally above; 5 marginally below;
3 clear reject; 1 trivial or wrong. 7 and 9 are allowed.
```

C.7 HUMAN STUDY PAPERS

Both sets in Section 5.3 come from NeurIPS 2021. Within each set, papers are listed in order of their NeurIPS mean rating. The pruning papers are GraNet, "Sparse Training via Boosting Pruning Plasticity with Neuroregeneration" (7.0); CHIP, "CHIP: CHannel Independence-based Pruning for Compact Neural Networks" (6.0); and "Efficient Neural Network Training via Forward and Backward Propagation Sparsification" (5.4). The adversarial robustness papers are NuAT, "Towards Efficient and Effective Adversarial Training" (7.0); "Data Augmentation Can Improve Robustness" (6.0); and "Shift Invariance Can Reduce Adversarial Robustness" (5.25). Like the simulation papers, all six are held out when the checklist tree is grown (Section 4).

D ADDITIONAL RESULTS OF THE SCORING STUDY

This appendix gives further and more detailed results of the scoring study in Section 5: how stable the population target is (Appendix D.1), the ICLR counterpart of Figure 3 (Appendix D.2), and how the spread of simulated scores compares with that of human scores (Appendix D.3). The prompts and settings of the study are in Appendix C.6.

D.1 STABILITY OF THE POPULATION TARGET

The population target $\hat{\mu}_i(\varnothing)$ is the average score over all F unaided simulated reviewers of a paper (Appendix C.6). Figure 4 shows how closely committees of different sizes approach it.

D.2 ICLR RESULTS

Figure 5 repeats Figure 3 for ICLR 2022. As on NeurIPS, the $k{=}100$ checklist cuts single-reviewer variance from 0.84 to 0.18 and has a small mean bias (-0.06). Its per-paper biases spread more widely than on NeurIPS, so its mean squared bias is 0.16 against 0.07 there, and the no-checklist committee catches up sooner, at $N{=}5$ rather than $N{=}10$. At $N{=}4$ the checklist is still marginally ahead (0.205 against 0.209).

D.3 SIMULATED AND HUMAN REVIEWER VARIANCE

The unaided simulated reviewers stand in for the reviewer lottery, so their scores for a paper should spread about as much as human scores for that paper do. Table 6 and Figure 6 compare the within-paper spread of single-reviewer scores on the 30 papers of each venue. Humans number three to nine per paper, so every variance here uses the unbiased estimator, with $n - 1$ in the denominator.

On NeurIPS, humans and simulated reviewers use the same 1-to-10 scale. The simulated variance, 0.84, is close to the variance among the reviewers of one human committee, 0.93, and so is the share of reviewer pairs two or more points apart (23% and 21%). Pooling the two committees raises the

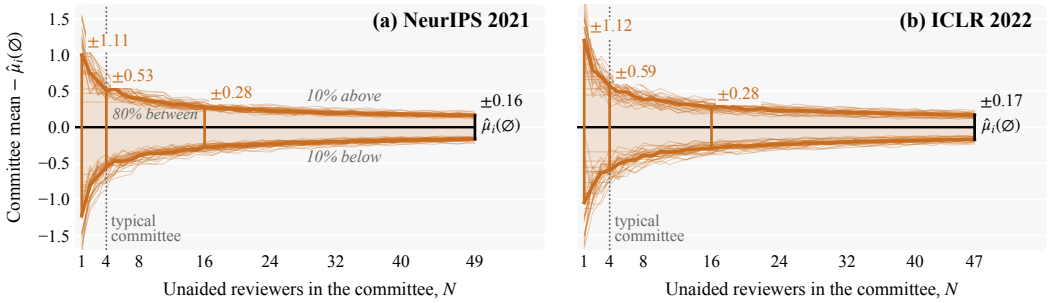


Figure 4: Variance in the average review score as more reviewers are added. For each paper, we draw committees of N unaided reviewers at random (with replacement) from its F simulated reviewers (F=49 for NeurIPS, 47 for ICLR), average their scores, and subtract the paper's target $\hat{\mu}_i(\varnothing)$, the average over all F. Zero means the committee matches the target. Each paper has two thin lines: a randomly drawn committee of size N lands above the upper line 10% of the time, below the lower line 10% of the time, and between them the remaining 80%. The thick lines and shading show the typical paper, taking the median of these lines over the 30 papers. A single reviewer is typically within about ±1.1 points of the target, a committee of four within about ±0.55, and all F reviewers within about ±0.17, which is the remaining uncertainty of $\hat{\mu}_i(\varnothing)$ itself.

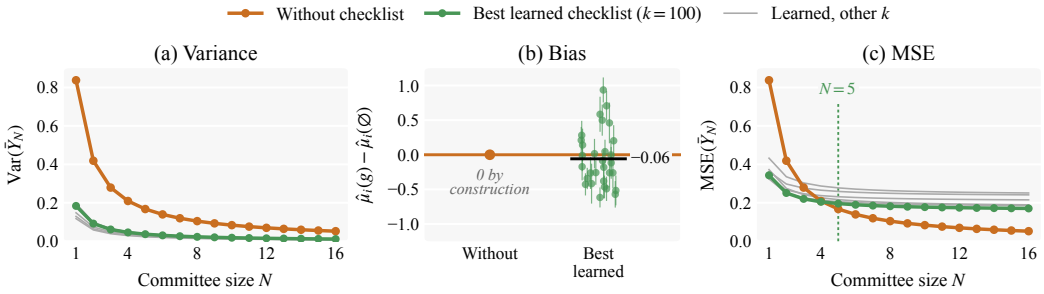


Figure 5: Variance, bias, and MSE of the committee mean on ICLR 2022 (30 papers), drawn as in Figure 3. ICLR has no official checklist, so only learned checklists appear. (b) Each point is one paper's bias, the checklist mean minus the target $\hat{\mu}_i(\varnothing)$ (±1 SE); the black bar averages over papers. (c) The dotted line marks the smallest N at which the no-checklist committee is at least as accurate as the k=100 checklist. Grey lines are learned checklists at other granularities, $k \in \{50, 200, 500, 1000, 2000\}$.

human variance to 1.50, because reviewers agree more with their own committee than with the other one. The same gap appears over all 298 NeurIPS pairs (1.07 within a committee, 1.33 pooled). The simulation draws reviewers independently and has no committees, so it matches the spread within a human committee but not the extra spread between committees. In Figure 6b, the simulated curve stays inside the one-committee band at every N, and the gap between the two is small next to the drop to the k=100 checklist curve.

On ICLR, each paper has one committee of about four reviewers, who score on the ICLR 2022 scale of 1, 3, 5, 6, 8, and 10. On that scale any two different scores other than 5 and 6 are at least two points apart, while the simulated reviewers use the 1-to-10 scale of the scoring prompt in Appendix C.6. As a result, 62% of human pairs give the same score and only 5% differ by one point, against 32% and 44% of simulated pairs. The human variance, 1.44, is higher than the simulated 0.86, but with four reviewers per paper its interval is wide, and the ratio of the two, 1.69 [0.87, 2.70], includes one. In Figure 6c, the simulated curve stays inside the human band at every N.

The simulation does not reproduce which papers split their reviewers: per-paper human and simulated variances are uncorrelated (Spearman −0.27 on NeurIPS, 0.05 on ICLR). Human committees do

Table 6: Within-paper spread of single-reviewer scores on the 30 papers of each venue, for simulated unaided reviewers and for the human reviewers of the same papers. n is the mean number of reviewers per paper, or per committee in the within-committee row. The last column is the share of same-paper reviewer pairs whose scores differ by at least two points. Brackets are paper-level bootstrap 95% CIs.

Venue	Reviewers	n	Variance	Pairs ≥ 2 apart (%)
NeurIPS 2021	Simulated, without checklist	49	0.84 [0.77, 0.92]	23.2 [21.0, 25.4]
	Human, within one committee	3.9	0.93 [0.63, 1.33]	20.7 [14.6, 27.2]
	Human, both committees pooled	7.7	1.50 [1.18, 1.84]	37.4 [30.9, 43.9]
ICLR 2022	Simulated, without checklist	47	0.86 [0.76, 0.95]	23.7 [20.9, 26.6]
	Human, one committee	3.8	1.44 [0.74, 2.28]	32.6 [21.2, 43.3]

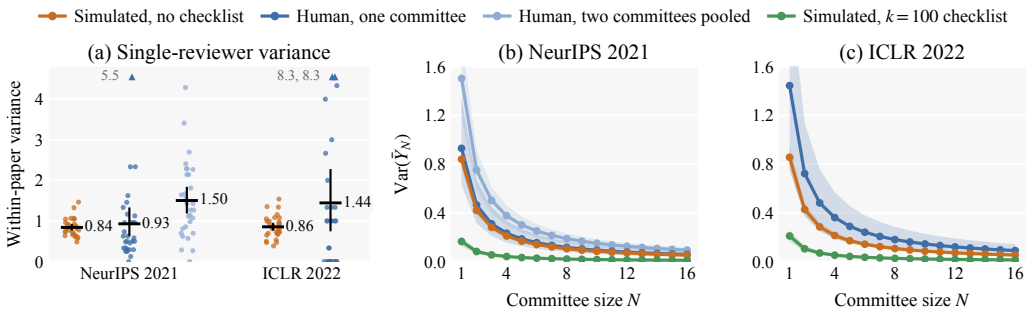


Figure 6: Score spread of simulated and human reviewers on the 30 papers of each venue. (a) Each point is one paper's within-paper variance of single-reviewer scores (unbiased estimator). Black bars are means with paper-level bootstrap 95% CIs; triangles mark papers above the axis, labelled with their values. For NeurIPS, the one-committee value averages the variances of the paper's two committees. (b, c) Variance of the committee mean, the mean variances in (a) divided by N, drawn as in Figure 3a with human committees added and the k=100 checklist for scale. Shading is the 95% CI; the ICLR human band reaches 2.28 at N=1, above the axis.

not agree on this either. Over all 298 NeurIPS pairs, the per-paper variances of the two committees correlate at 0.09. Paper means agree more closely: across the 30 papers, which run from strong to weak, the simulated target $\hat{\mu}_i(\varnothing)$ and the mean human rating have Spearman correlation 0.79 on NeurIPS and 0.77 on ICLR.

For reference only, we also score the human ratings against $\hat{\mu}_i(\varnothing)$ as in Figure 3; these results are not reported in the main text. A single human rating has MSE 2.123 on NeurIPS and 2.874 on ICLR, and human ratings sit on average 0.38 points below $\hat{\mu}_i(\varnothing)$ on NeurIPS and 0.33 points above it on ICLR. Because $\hat{\mu}_i(\varnothing)$ is defined by LLM reviewers, this gap says nothing about human bias.