BETTER REVIEWS, BUSIER AREA CHAIRS: WHY A LITTLE HUMAN JUDGMENT MATTERS IN AI-ASSISTED REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Growing submission volumes have intensified interest in AI-assisted peer review, but its benefits depend on the expert effort required to turn reviews into reliable decisions. We study the downstream workload of Area Chairs (ACs), whose capacity to integrate reports and determine what requires further assessment can become a bottleneck. We develop a Bayesian framework linking AI-assisted reports to the AC's downstream workload, defined as the minimum additional assessment cost required to estimate paper quality at a fixed reliability target. We establish a paradox: AI assistance can improve both individual report accuracy and the panel's aggregate assessment while increasing AC workload, even when the AC chooses assessments adaptively and optimally. We further show that a small amount of independent reviewer judgment may be sufficient to eliminate this increase and substantially reduce AC workload. Experiments using expert-annotated real reviews from PEERREVIEW BENCH are consistent with these theoretical findings. Our results highlight the value of preserving independent human judgment when designing AI-assisted review processes.

1 INTRODUCTION

AI assistance has the potential to ease the burden of peer review by helping reviewers understand papers and prepare their reports. Whether it also reduces the human effort required to reach a reliable publication decision remains less clear. An Area Chair (AC) must combine several reports and determine what issues still require independent investigation before making a recommendation. More accurate reports would naturally be expected to reduce this downstream burden. We show that AI assistance can instead increase the verification needed to reach the same level of reliability, and characterize when retaining a small amount of reviewers' independent judgment within AI-assisted reports is sufficient to avoid this increase.

The practical stakes lie in the limited capacity to turn reviews into reliable decisions as submission volumes continue to grow. While initial assessments are distributed across multiple reviewers, each paper's reports must ultimately be reconciled by an AC exercising expert judgment. NeurIPS 2025, for example, handled 21,575 submissions with 1,663 ACs and 199 SACs, and its organizers identified recruiting experienced committee members at scale as a challenge (NeurIPS Communications Chairs, 2025). If AI assistance saves reviewer time while increasing downstream verification, it shifts work from the many to the few. When AC capacity becomes the bottleneck, time saved in preparing reviewer reports may not increase the number of reliable decisions the process can support. We therefore analyze AC verification as a distinct constraint on the benefits of AI-assisted reviewing.

The mechanism begins with correlation among LLM responses. AI-generated reviews often raise overlapping concerns, even when produced by different AI models (Kim et al., 2026a). When reviewers rely heavily on such drafts, their reports can become more homogeneous, emphasizing the same aspects of a paper while leaving similar questions insufficiently examined. These shared limitations reduce the complementary information available across reports, so agreement need not reflect independent confirmation. The verification burden therefore depends on both overall accuracy and where uncertainty remains. Assistance can improve the panel's aggregate assessment while leaving

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

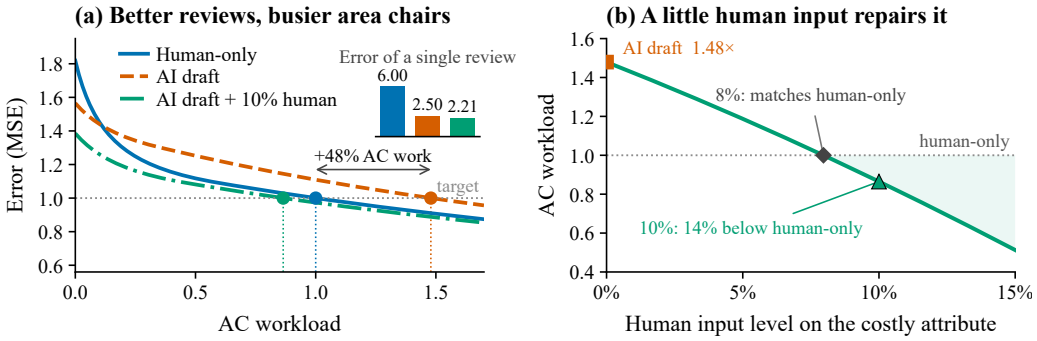

Figure 1: **Better reviews, busier area chairs; a little independent human judgment repairs it.** (a) Error of the AC's quality estimate against AC workload; (b) AC workload as independent human judgment is added on the costly attribute. See Section 3 for setting details and Section 4 for the repair.

uncertainty concentrated in aspects that are costly for the AC to investigate. The effort required to resolve this remaining uncertainty can outweigh the gains from improved overall accuracy, thereby increasing the total verification required to reach the same reliability target.

To the best of our knowledge, we are the first to identify and formally characterize this AC workload paradox in AI-assisted peer review. We develop a Bayesian framework that connects the information in assisted reports to the minimum additional verification an AC must perform to meet a fixed reliability target. We show that assistance can improve both individual report accuracy and the panel's aggregate assessment while increasing this downstream burden, even when the AC accounts for dependence among reports and chooses verification optimally. This identifies a distinct limitation of AI-assisted reviewing: improvements in report accuracy need not translate into less expert effort at the decision stage. This raises a practical question:

How can AI assistance ease reviewers' workload without increasing the AC's verification burden?

Our theoretical framework points to a possible remedy: retaining a small amount of each reviewer's independent judgment within the AI-assisted report. When reports rely heavily on AI drafts, even limited human input can supply valuable complementary information and substantially reduce downstream verification. We characterize conditions under which this input is sufficient to recover the human-only AC workload benchmark while preserving the benefits of AI assistance. Figure 1 illustrates both results in a two-attribute instance. AI drafts make every report more accurate yet require 48% more AC work to reach the same reliability target, and independent judgment on a single attribute, at a weight of only 10%, brings the workload 14% below the human-only benchmark.

We examine both the paradox and the proposed remedy through semi-synthetic experiments using expert-annotated human and AI reviews from PEERREVIEW BENCH (Kim et al., 2026a). The results show a pattern consistent with the paradox: AI reports exhibit higher individual quality but greater simulated AC inspection burden, accompanied by significantly more repetition across reports. Incorporating limited human input into AI drafts significantly improves coverage of distinct, expert-validated concerns and reduces inspection burden. Using only a quarter of the items from the paired human review brings the workload point estimate close to the human-only baseline.

1.1 RELATED WORK

AI-assisted peer review. Growing submission volumes have motivated increasing interest in AI-assisted peer review. Prior work highlights the potential of LLMs to improve review quality and efficiency, including providing useful scientific feedback and helping reviewers refine their reports (Liang et al., 2024; Thakkar et al., 2026; Chen et al., 2025; Bhargava et al., 2026). Recent studies examine the quality and overlap of AI-generated reviews (Kim et al., 2026a), as well as how LLM-use policies and tools affect reviewer and AC experiences (Kim et al., 2026b). Closest to ours is

the concurrent and independent work of Gans (2026), which studies how providing reviewers with a common AI-generated report changes what they investigate. We complement these studies by examining how reviewer assistance changes the minimum additional assessment cost an AC must incur to estimate paper quality at a fixed reliability target.

Correlated errors and collective outcomes. Research on algorithmic monoculture shows that adopting a more accurate common algorithm can reduce collective welfare (Kleinberg & Raghavan, 2021). Similar findings arise in other domains, including content creation (Doshi & Hauser, 2024). More recently, Kim et al. (2025) document substantial correlated errors among LLMs, even across distinct architectures and providers. We connect these error correlations to the AC's downstream assessment workload and characterize conditions under which limited independent human input suffices to eliminate the resulting workload increase.

Human–AI collaboration. More broadly, our work relates to the growing literature on human–AI collaboration, which investigates how human and machine capabilities complement one another and how their interaction shapes joint performance (Bansal et al., 2021; Vaccaro et al., 2024; Yin & Xin, 2026). We study this question in a setting where several assisted reviewers provide information to an AC, focusing on how independent reviewer judgment affects the AC's subsequent assessment burden.

2 Model Formulation

Typically, authors submit papers to journals or conferences under a self-selected contribution category (e.g., theoretical contribution, methodological innovation, or empirical study). Under each category, the venue evaluates the paper's quality against a rubric. We model the quality of a paper by

$$q = \theta^\top x, \qquad x \sim \mathcal{N}(m_0, \Lambda), \qquad \Lambda \succ 0,$$

where $x \in \mathbb{R}^d$ is a latent feature vector of the submitted paper, representing, for instance, the correctness of the main technical claim, the validity of the experiments, and novelty relative to prior work. The public rubric $\theta \in \mathbb{R}^d$ specifies how the venue weighs these attributes in its assessment of the paper's overall quality. The AC estimates quality by assessing these attributes and combining them according to the rubric.

Human-only reports and AI-assisted reports. The AC decides on the paper's acceptance, helped by K reviewers who estimate latent feature x. Traditionally, reviewers review the paper without a generative model drafting the report or summarizing the paper. We call these *human-only* reports. The reviewer may use an AI model for polishing, but the assessment itself is first drafted by the reviewer. We model a human-only report by

$$Y_j^H = x + \epsilon_j, \qquad \epsilon_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma), \qquad \Sigma \succ 0. \tag{1}$$

As LLM use spreads, reviewers increasingly use AI models to help understand a paper, examine its technical claims, identify strengths and weaknesses, and even draft their reviews. Reviewers may then verify, revise, or supplement the AI-generated assessment with their own understanding and judgment. We call the resulting reports *assisted* reports. To model them, we let the AI model first provide an initial estimate

$$D_j = x + b + \zeta_j, \qquad b \sim \mathcal{N}(0, T_D), \qquad \zeta_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_D), \qquad \Sigma_D \succ 0,$$

where $T_D \succeq 0$ and $\Sigma_D \succ 0$. Here, b captures a paper-specific error shared across AI-generated assessments. This shared component is motivated by documented preferences for linguistic presentation in LLM evaluation (Wu & Aji, 2025) and substantial error correlations across models, including those with different LLMs (Kim et al., 2025). The term ζ_j captures reviewer-specific variation associated with model choice, prompting strategy, and review workflow. We assume that x, b, $(\epsilon_j)_{j=1}^K$, and $(\zeta_j)_{j=1}^K$ are mutually independent.

We model an assisted report as a weighted combination of the AI-generated assessment D_j and the assessment Y_j^H that the reviewer would produce under independent human-only reviewing:

$$Y_j^A = (I - E)D_j + EY_j^H, \qquad E = \operatorname{diag}(e_1, \ldots, e_d), \qquad 0 \le e_i \le 1.$$

The coefficient e_i represents reviewer input on attribute i, measured by the contribution of human judgment to the final assessment. The remaining weight $1 - e_i$ represents reliance on AI. These weights may differ across attributes but, for simplicity, we assume they are fixed and common across reviewers. The endpoints $E = 0$ and $E = I$ recover the AI-generated assessment and the human-only report, respectively.

Writing $\beta = (I - E)b$ and $\widetilde{\epsilon}_j = (I - E)\zeta_j + E\epsilon_j$, we obtain

$$Y_j^A = x + \beta + \widetilde{\epsilon}_j, \quad \beta \sim \mathcal{N}(0, T), \quad \widetilde{\epsilon}_j \overset{\text{iid}}{\sim} \mathcal{N}(0, \Sigma_A)$$

Here, $T = (I - E)T_D(I - E)^\top, \Sigma_A = (I - E)\Sigma_D(I - E)^\top + E\Sigma E^\top$ and β is independent of $\widetilde{\epsilon}_j$. Compared to Equation (1), assisted reports contain a shared error component inherited from AI, with covariance T, and change the reviewer-specific error covariance from Σ to Σ_A. Reviewer input level E determines both components and therefore the accuracy of the resulting report. An individual assisted report has error covariance $T + \Sigma_A$. To examine whether better individual reports also reduce the AC's workload, we make the following assumption.

Assumption 2.1 (Individual improvement)**.** An individual assisted report has a strictly smaller error covariance than a human-only report: $T + \Sigma_A \prec \Sigma$.

Under Assumption 2.1, $u^\top Y_j^A$ has lower mean-squared error than $u^\top Y_j^H$ as an estimate of $u^\top x$, for every $u \neq 0$. This includes the paper's quality $q = \theta^\top x$.

2.1 Information available to the AC

We first consider the information the AC can obtain from the K reviews without independently checking the paper. For $g \in \{H, A\}$, let $\bar{Y}^g = K^{-1}\sum_{j=1}^K Y_j^g$. The panel means have error covariances

$$\Omega_H = \frac{\Sigma}{K}, \qquad \Omega_A = T + \frac{\Sigma_A}{K}. \tag{2}$$

Reviewer-specific errors average out across reviewers, whereas the shared error remains. Further, since $Y_j^A - \bar{Y}^A$ is independent of β, the sample covariance or disagreement level across reports estimates does not capture information about the shared error.

To characterize the best estimate available from the reports only, we consider an AC who knows E and all model parameters but observes only the submitted reports. This benchmark allows the AC to account correctly for their dependence. The panel mean $\bar{Y}^g$ is a sufficient statistic for x: it carries the same information about x as $Y_1^g, \ldots, Y_K^g$. Applying Bayes' rule, we know

$$x \mid Y_1^g, \ldots, Y_K^g \sim \mathcal{N}(m_g, V_g), \quad V_g = (\Lambda^{-1} + \Omega_g^{-1})^{-1}, \quad m_g = V_g\big(\Lambda^{-1}m_0 + \Omega_g^{-1}\bar{Y}^g\big). \tag{3}$$

See Appendix A.1 for details. Under squared-error loss, the Bayes estimator of q is $\theta^\top m_g$, with minimum posterior mean-squared error $\theta^\top V_g \theta$. This is the uncertainty that remains after optimally combining the submitted reports.

2.2 AC workload and reliability

In practice, the AC may devote additional effort to examining the paper, checking reviewers' assessments, and identifying strengths or weaknesses that the reports have overlooked. We model this additional work through L assessment tasks, with directions $a_\ell \in \mathbb{R}^d$, $\ell = 1, \ldots, L$. Allocating effort level $t_\ell > 0$ to task ℓ produces a measurement

$$Z_\ell = a_\ell^\top x + \eta_\ell, \qquad \eta_\ell \sim \mathcal{N}(0, 1/t_\ell),$$

at cost $c_\ell t_\ell$, where $c_\ell > 0$. A zero allocation omits the task. For a fixed allocation, the measurement errors are mutually independent and independent of x and the original report errors. The task-specific costs c_ℓ capture differences in the work required to obtain the same measurement precision.

The AC combines the submitted reports with these independent measurements. For a fixed allocation t, the posterior covariance and minimum posterior mean-squared error are

$$V_g(t) = \left(V_g^{-1} + \sum_{\ell=1}^L t_\ell a_\ell a_\ell^\top\right)^{-1}, \qquad \mathcal{R}_g(t) = \theta^\top V_g(t)\theta. \tag{4}$$

See Appendix A.1 for details. The minimum risk is attained by the posterior mean of q. For a target mean-squared error $\varepsilon > 0$, define the minimum additional AC workload by

$$W_g(\varepsilon) = \min_{t \in \mathbb{R}_+^L} \left\{ \sum_{\ell=1}^{L} c_\ell t_\ell : \mathcal{R}_g(t) \le \varepsilon \right\}.$$

The quantity $W_g(\varepsilon)$ measures the minimum cost required for the AC to reach the target accuracy. We set $W_g(\varepsilon) = +\infty$ if the constraint is infeasible. If $\theta^\top V_g \theta \le \varepsilon$, the reports alone already meet the target, and no additional assessment is needed. We next show $W_g(\varepsilon)$ is the minimum cost for all possible policy.

Proposition 2.2. *Allow the AC to choose each assessment task, its effort, and when to stop based on the reports and previous measurements. Among policies that stop almost surely, the minimum expected assessment cost of producing a final estimate $\widehat{q}$ with $\mathbb{E}[(\widehat{q} - q)^2] \le \varepsilon$ is $W_g(\varepsilon)$.*

Thus, $W_g(\varepsilon)$ characterizes the minimum expected workload required to meet the reliability target, even when the AC chooses assessments adaptively.

3 Better Reviews Can Increase AC Workload

Assumption 2.1 concerns individual reports, whereas the AC's information is governed by the panel covariances in Equation (2). Individual improvement, $T + \Sigma_A \prec \Sigma$, does not imply panel information dominance, $\Omega_A \preceq \Omega_H$.

Condition 3.1 (Dominance)**.** $\Omega_A \preceq \Omega_H$.

Averaging removes reviewer-specific error but not the shared error, so more accurate reports can leave the panel less informative in some directions. Whether such a loss costs the AC work depends on which directions the AC can verify and at what price.

Proposition 3.2. *For fixed $\theta \neq 0$, Condition 3.1 holds if and only if $W_A(\varepsilon) \le W_H(\varepsilon)$ for every assessment menu, positive cost vector, and target $\varepsilon > 0$.*

Read in the contrapositive, Proposition 3.2 says that whenever Condition 3.1 fails, for every rubric, there are an assessment menu, positive costs, and a target for which assisted reviewing requires strictly more AC workload; the proof in Appendix B.2 exhibits one with finite workloads, i.e., $W_H(\varepsilon) < W_A(\varepsilon) < \infty$. Since dominance can fail under Assumption 2.1, i.e., the individual improvement assumption, this is already the paradox in qualitative form. In the following, Theorem 3.3 quantifies the loss for a given AC.

The AC's work strategy design. Consider first an AC who sets the reports aside. Write $J(t) := \sum_{\ell=1}^{L} t_\ell a_\ell a_\ell^\top$ for the information generated by an effort allocation t. The minimum cost of learning q to unit precision from assessments alone is

$$W_0 := \inf \Big\{ \sum_{\ell=1}^{L} c_\ell t_\ell : \; t \in \mathbb{R}_+^L, \; J(t) \succ 0, \; \theta^\top J(t)^{-1} \theta \le 1 \Big\}. \tag{5}$$

We assume the infimum is attained at some $\tilde{t}$ with $J(\tilde{t}) \succ 0$. Let $w^* := J(\tilde{t})^{-1}\theta$.

Theorem 3.3. *Fix a menu, positive costs, and a rubric $\theta \neq 0$ as above. There is $\varepsilon_* \in (0, \infty)$, given in Appendix B, such that for every $0 < \varepsilon < \varepsilon_*$,*

$$W_g(\varepsilon) = \frac{W_0}{\varepsilon} - W_0 \, w^{*\top} \big(\Lambda^{-1} + \Omega_g^{-1}\big) w^*, \qquad g \in \{H, A\}, \tag{6}$$

and therefore

$$W_A(\varepsilon) - W_H(\varepsilon) = W_0 \, w^{*\top} (\Omega_H^{-1} - \Omega_A^{-1}) \, w^*. \tag{7}$$

Only $\Omega_H^{-1} - \Omega_A^{-1}$ on the right of Equation (7) depends on the reports. W_0 and w^* are fixed by the menu, costs and rubric. Assistance therefore increases the AC's work exactly when the assisted panel is less precise along w^*, $w^{*\top}\Omega_A^{-1} w^* < w^{*\top}\Omega_H^{-1} w^*$, however much more precise it is elsewhere;

the direction w^* determines which differences in initial precision affect the workload gap. Below ε_*, under both human-only and assisted setting, the AC ends at the same posterior covariance εw^* between x and q, so the gap is the cost of reaching this common endpoint from different starting precisions.

For attribute-specific assessments ($L = d$ and task i measures attribute i) with diagonal Λ, Σ, T, and Σ_A and $\theta_i \neq 0$, the design problem separates across attributes, $w_i^* \propto \operatorname{sgn}(\theta_i)\sqrt{c_i}$, and Equation (7) reads

$$W_A(\varepsilon) - W_H(\varepsilon) = \sum_{i=1}^{d} c_i \Big(\frac{1}{\omega_{H,i}} - \frac{1}{\omega_{A,i}} \Big), \qquad \omega_{g,i} = (\Omega_g)_{ii}, \tag{8}$$

with $\varepsilon < \varepsilon_*$. This isolates the source of the reversal – precision losses on expensive attributes can outweigh precision gains on inexpensive ones, and the rubric weights drop out.

A two-attribute instance. Consider $K = 3$ reviewers and two attributes, with $\theta = (1, 1)^\top$, $\Lambda = 10I$, $\Sigma = 3I$, $T_D = \operatorname{diag}(1/3, 1)$, and $\Sigma_D = \operatorname{diag}(1/2, 2/3)$. We use unmodified AI drafts, $E = 0$, and attribute-specific assessments with costs $c = (1, 25)^\top$. The second attribute may represent the correctness of a technical argument that the AC must reconstruct, making it substantially more costly to assess.

Figure 1(a) illustrates the reversal. Assisted reports improve both individual accuracy and the report-only estimate of overall quality, yet increase the AC's minimum workload. The panel gains information about the inexpensive attribute but loses it on the expensive one, and the resulting saving on the first attribute is outweighed by the additional cost of assessing the second. As a result, reaching the same target requires 48% more AC work than human-only reviewing. Figure 1(b) previews the remedy. Independent judgment with a weight of only 10%, on the expensive attribute alone, brings the AC's workload below the human-only benchmark. In the next section, we characterize how much independent judgment is needed, and where.

4 Limited Reviewer Input Can Repair the Panel

In this section, we show that a limited contribution of independent judgment suffices to remove the additional burden. We now vary the input matrix E while keeping the primitive covariances T_D, Σ_D, and Σ fixed, and write $\Omega_A(E)$ and $W_A(\varepsilon; E)$ to make this dependence explicit. Consider first the uniform path $E = eI$, $0 \leq e \leq 1$. With $\Omega_D := T_D + \Sigma_D/K$ the covariance of the mean AI draft,

$$\Omega_A(E) = (I - E)\,\Omega_D\,(I - E)^\top + E\,\Omega_H\,E^\top = (1 - e)^2\,\Omega_D + e^2\,\Omega_H, \tag{9}$$

whose derivative at $e = 0$ is $-2\Omega_D \prec 0$. Therefore a small independent contribution reduces draft error to first order, while human error enters only at second order. How much input judgment is enough depends on how much noisier the drafts' panel can be than the human-only panel in any direction, $r := \max_{u \neq 0} \frac{u^\top \Omega_D\, u}{u^\top \Omega_H\, u} = \lambda_{\max}\big(\Omega_H^{-1/2} \Omega_D \Omega_H^{-1/2}\big)$. We compare it with two individual-level ratios,

$$\kappa := \lambda_{\max}\big(\Sigma^{-1/2}(T_D + \Sigma_D)\Sigma^{-1/2}\big), \qquad \rho := \lambda_{\max}\big(\Sigma^{-1/2} T_D \Sigma^{-1/2}\big),$$

the error of a single draft, and of its shared component, relative to a human-only report.

Theorem 4.1. *Let $e_H := (r-1)_+/(r+1)$. For $e \in [0, 1]$, the following are equivalent: (a) $e \geq e_H$; (b) $W_A(\varepsilon; eI) \leq W_H(\varepsilon)$ for every rubric, assessment menu, positive cost vector, and target $\varepsilon > 0$. Moreover,*

$$\frac{(K\rho - 1)_+}{K\rho + 1} \;\leq\; e_H \;\leq\; \frac{\big(\kappa + (K-1)\rho - 1\big)_+}{\kappa + (K-1)\rho + 1}. \tag{10}$$

Thus e_H is exactly the smallest uniform input level under which assisted reviewing requires no more AC work than human-only reviewing for every rubric, assessment menu, cost vector, and target; below it, some assessment problem requires more. Equation (10) reads the threshold off individual-level quantities. Aggregation removes independent error but not shared error, so $K - 1$ additional drafts add a penalty $(K - 1)\rho$ that averaging cannot reduce, against which stands the drafts' individual accuracy margin $1 - \kappa$. No input is needed when the margin covers the penalty,

$(K-1)\rho \le 1-\kappa$. Conversely, once $\rho > 1/K$, some input level is necessary no matter how accurate the drafts are otherwise.

If the unmodified drafts improve on human-only reports, $T_D + \Sigma_D \prec \Sigma$, then $\rho \le \kappa < 1$, so the upper bound in Equation (10) is at most $(K-1)\rho/(2+(K-1)\rho)$; for $K=3$ this is $\rho/(1+\rho) \le \min\{1/2, \rho\}$. The required uniform input level is therefore small whenever the shared error is a small share of human-report error (ρ is small) at $K=3$. In the instance of Section 3, $\kappa = 5/9$, $\rho = 1/3$, and $e_H = 1/10$ attains the upper bound.

Theorem 4.1 restores precision in every direction, which a guarantee over all assessment problems requires. For a given rubric, assessment menu, and costs, however, Theorem 3.3 shows that for targets below ε_* only the precision along the direction w^* enters the workload gap. Applied with $\Omega_A(E)$ in place of Ω_A, and writing $\varepsilon_*(E)$ for the corresponding threshold, it gives, for $0 < \varepsilon < \varepsilon_*(E)$,

$$W_A(\varepsilon; E) - W_H(\varepsilon) = W_0\, w^{*\top}\big(\Omega_H^{-1} - \Omega_A(E)^{-1}\big)w^*, \tag{11}$$

because W_0 and w^* do not depend on the panel. Assisted reviewing with input level E therefore needs no more AC work than human-only reviewing at such targets if and only if $w^{*\top}\Omega_A(E)^{-1}w^* \ge w^{*\top}\Omega_H^{-1}w^*$: only the precision along w^* has to be restored. Along the uniform path, let e_W be the least input level beyond which this assessment problem's excess workload stays removed,

$$e_W := \min\big\{e \in [0,1] :\ w^{*\top}\Omega_A(e'I)^{-1}w^* \ge w^{*\top}\Omega_H^{-1}w^* \text{ for all } e' \in [e,1]\big\}. \tag{12}$$

Proposition 4.2. *Under the hypotheses of Theorem 3.3, every uniform input level $e \ge e_W$ gives $W_A(\varepsilon; eI) \le W_H(\varepsilon)$ for all targets below $\varepsilon_*(eI)$, and $e_W \le e_H$.*

Restoring precision along one direction can cost less than restoring it in every direction. In the instance of Section 3, $e_W \approx 0.07$ on the uniform path, against $e_H = 1/10$. The proof is in Appendix C.4.

Where should reviewers add input? Attribute-specific input can leave attributes on which the panel is already informative untouched. Suppose the primitive covariances are diagonal with entries $\sigma_i = \Sigma_{ii}$, $\tau_i = (T_D)_{ii}$, and $\gamma_i = (\Sigma_D)_{ii}$, so that $\omega_{D,i} = \tau_i + \gamma_i/K$ and $\omega_{A,i}(E) = (1-e_i)^2\omega_{D,i} + e_i^2\omega_{H,i}$. For attribute-specific assessments ($L = d$ and task i measures attribute i), Equation (11) becomes

$$W_A(\varepsilon; E) - W_H(\varepsilon) = \sum_{i=1}^{d} c_i\Big(\frac{1}{\omega_{H,i}} - \frac{1}{(1-e_i)^2\omega_{D,i} + e_i^2\omega_{H,i}}\Big). \tag{13}$$

Equation (13) is linear in the costs, so the excess workload is removed for every cost vector if and only if each term is nonpositive, that is, $(1-e_i)^2\omega_{D,i} + e_i^2\omega_{H,i} \le \omega_{H,i}$ for every i, which solves to

$$e_i \ \ge\ e_{H,i} := \frac{(r_i - 1)_+}{r_i + 1}, \qquad r_i := \frac{\omega_{D,i}}{\omega_{H,i}} = \frac{K\tau_i + \gamma_i}{\sigma_i}.$$

Let $E_H := \mathrm{diag}(e_{H,1}, \ldots, e_{H,d})$. It is the smallest repair, adds input only on attributes with $r_i > 1$. The uniform threshold is its largest entry, $e_H = \max_i e_{H,i}$.

In the instance of Section 3, $E_H = \mathrm{diag}(0, 1/10)$. A weight of one tenth on the costly attribute alone restores dominance, removes the excess workload (Figure 1(b)), and for this specific assessment problem about 0.08 already suffices. A little independent judgment, placed where the shared error concentrates and verification is costly, is enough to reduce the AC's workload, here to below the human-only benchmark.

5 Experiments

5.1 Setup

Dataset and data processing. We use the expert-annotated dataset PeerReview Bench (Kim et al., 2026a).[1] The dataset contains human and AI reviews of Nature-family papers, with individual

[1]Dataset release: `https://huggingface.co/datasets/prometheus-eval/peerreview-bench`.

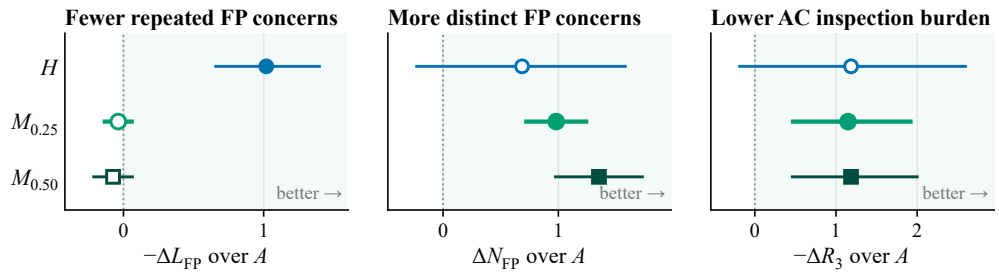

Figure 2: **Improvement over the AI-draft panel A.** Each row compares a three-review panel with A; Δ denotes the row's panel minus A, with signs chosen so that positive values are improvements. Filled markers indicate 95% confidence interval excludes zero.

review items annotated by domain experts for correctness, significance, and sufficiency of evidence. We focus on the authors' *fully positive* (FP) criterion: an item is fully positive if it is rated correct, significant, and supported by sufficient evidence. Our analysis uses a subset of 54 evaluation papers, each with three human reviews and three released AI reviews, one each from GPT-5.2, Claude Opus 4.5, and Gemini 3.0 Pro.

Because the AI reviewers were instructed to produce at most five significant review items, we apply the same cap to human reports, using `gpt-6-sol` to select five items from longer reports. We denote the resulting capped human-authored reports by H and the released AI reports by A, using the latter as a proxy for the draft-dominated endpoint of AI-assisted reviewing. On average, H and A contain 4.36 and 4.24 items, respectively. See Appendix D for details.

Human-informed revision. We denote a constructed human-informed revision by M_α. A `gpt-6-sol` editor receives one AI draft and a uniformly sampled subset containing a fraction α of a paired human review's items. We consider exposure fractions $\alpha \in \{0.25, 0.50\}$. Starting from the AI draft, the editor selects at most five original items from the draft and the sampled human items, aiming to retain substantive concerns. It can retain, replace, or append items, but cannot rewrite them or introduce new claims. The resulting review therefore preserves the original expert annotations of each selected item. The exposure fraction controls available human input and simulates a human verifier. It can be understood as a human input proxy.

Simulated AC works and workload measure. For each paper, we group items from the three human reports and three AI reports according to whether they express the same specific concern. Each panel includes three reports, and exposes only its own items within these groups. We measure the FP coverage by N_{FP}, the number of distinct FP items in the panel. Let $N_{\mathrm{FP},j}$ denote the corresponding count within report j. We define cross-report repetition loss as $L_{\mathrm{FP}} = \sum_{j=1}^{3} N_{\mathrm{FP},j} - N_{\mathrm{FP}}$.

We use `gpt-6-sol` to rank the visible groups. We then simulate an AC inspecting these representative items in the resulting order. We assume the AC at least needs to find three FP reviews. We report FP_5, the number of fully positive items among the AC's first five selections. To characterize inspection burden, let pos_3 denote the first ranking position at which three FP items have been encountered, with $\mathrm{pos}_3 = \infty$ if the target is unattainable. Since each review contains at most 5 items, we consider $R_3 = \min\{\mathrm{pos}_3, 15\}$. We report the mean of R_3 across papers as a measure of restricted inspection burden, together with its confidence interval.

5.2 Results

Better reports, while more AC workload. AI reports have a numerically higher individual FP fraction than human reports, but their FP concerns are significantly more repeated across reviewers. As shown in Table 1 and the top row of Figure 2, panel-level point estimates favor human reports. Human panels provide broader distinct FP coverage N_{FP}, yield more FP items among the AC's first five selections FP_5, and incur lower restricted inspection burden R_3. This pattern is consis-

Table 1: **Individual review quality, panel coverage, and AC workload.** Each cell reports an estimate with its pointwise 95% paper-bootstrap confidence interval below. Bold estimates have pointwise confidence intervals excluding zero.

	Individual report	Three-review panel			
Report / panel	FP rate (%) ↑	N_{FP} ↑	L_{FP} ↓	FP_5 ↑	R_3 ↓
H	54.10 $[47.62, 60.39]$	6.13 $[5.30, 6.94]$	0.24 $[0.13, 0.35]$	3.11 $[2.74, 3.48]$	6.28 $[5.20, 7.48]$
A	60.30 $[51.95, 68.47]$	5.44 $[4.54, 6.37]$	1.26 $[0.91, 1.65]$	2.74 $[2.28, 3.20]$	7.46 $[6.13, 8.83]$
$H - A$	-6.20 $[-14.82, 2.38]$	$+0.69$ $[-0.24, 1.59]$	$\mathbf{-1.02}$ $[-1.41, -0.65]$	$+0.37$ $[-0.15, 0.89]$	-1.19 $[-2.61, 0.20]$

Table 2: **Human-informed revisions: individual quality and panel outcomes.** Entries are paired with 95% bootstrap confidence intervals below. Bold estimates have pointwise confidence intervals excluding zero.

		Individual report	Three-review panel			
α	Report / panel	Δ FP rate (%)	ΔN_{FP}	ΔL_{FP}	ΔFP_5	ΔR_3
0.25	$M_\alpha - A$	$+1.88$ $[-0.25, 3.93]$	$\mathbf{+0.98}$ $[0.70, 1.26]$	$+0.04$ $[-0.07, 0.15]$	$+0.15$ $[0.00, 0.30]$	$\mathbf{-1.15}$ $[-1.94, -0.44]$
	$H - M_\alpha$	$\mathbf{-8.08}$ $[-15.43, -0.98]$	-0.30 $[-1.11, 0.50]$	$\mathbf{-1.06}$ $[-1.41, -0.70]$	$+0.22$ $[-0.24, 0.69]$	-0.04 $[-1.09, 1.00]$
0.50	$M_\alpha - A$	$+2.65$ $[-0.41, 5.77]$	$\mathbf{+1.35}$ $[0.96, 1.74]$	$+0.07$ $[-0.07, 0.22]$	$+0.20$ $[-0.02, 0.44]$	$\mathbf{-1.19}$ $[-2.02, -0.44]$
	$H - M_\alpha$	$\mathbf{-8.85}$ $[-15.44, -2.49]$	-0.67 $[-1.43, 0.09]$	$\mathbf{-1.09}$ $[-1.50, -0.72]$	$+0.17$ $[-0.28, 0.61]$	0.00 $[-0.98, 1.02]$

tent with the phenomenon highlighted by the theory – individually useful reports need not supply complementary information that makes the AC's task easier.

Limited human input improves downstream review outcomes significantly. From Figure 2 and Table 2, we know human-informed revisions increase distinct FP coverage and reduce restricted inspection burden relative to AI-only panels at both exposure levels. At 25% exposure, the revisions add approximately one distinct FP concern per panel and reduce mean R_3 by 1.15 positions. The same qualitative pattern holds at 50% exposure. Workload differences between human-informed and human-only panels are small in point estimates.

Taken together, our theoretical and empirical findings highlight the importance of designing mechanisms that encourage reviewers to examine AI-generated assessments and contribute independent observations, as even limited human input can substantially reduce downstream AC workload. The ICLR 2027 policy already takes a related approach. It permits AI assistance while requiring reviewers using AI to disclose their original assessments and AI interactions (ICLR, 2026).

6 Conclusion

In this paper, we develop a Bayesian framework showing that individually more accurate AI-assisted reports can increase AC workload at a fixed reliability target, while limited independent reviewer judgment can prevent this increase. Semi-synthetic experiments with expert-annotated reviews provide evidence consistent with these findings. Our limitation includes the Gaussian errors structure and simulated AC assessment in the experiment. Future work could incorporate heterogeneous expertise and mechanism design for reviewer incentives, and evaluate workload in real review processes.

AI USE STATEMENT

We used generative AI tools to assist with writing code, word polishing, literature search and comprehension, and the presentation of mathematical expressions. We also used generative AI tools to construct semi-synthetic experimental conditions and perform model-based review-item selection, grouping, and ranking. The models and prompts used in these experiments are documented in Appendix D. We take full responsibility for the final content of this work, including all AI-assisted text, claims, code, and experimental artifacts.

ETHICS STATEMENT

This study uses existing, publicly released peer-review data and does not involve recruiting new human participants. Our work aims to provide insights for the design of responsible AI-assisted peer-review processes.

REPRODUCIBILITY STATEMENT

The experimental code is provided in the supplementary material. Appendix D documents the data processing procedures, model configurations, and prompts used to construct and evaluate the semi-synthetic review panels.

REFERENCES

Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–16, 2021.

Hemant K Bhargava, Sarah H Bana, Zhe Zhang, Laura Brandimarte, Vidyanand Choudhary, J Frank Li, Pantelis Loupos, and Daniel Zantedeschi. Fighting fire with fire: Infusing artificial intelligence into peer review to sustain quality scholarship. *Management Science*, 2026.

Shiping Chen, Duncan Brumby, and Anna Cox. Envisioning the future of peer review: Investigating llm-assisted reviewing using chatgpt as a case study. In *Proceedings of the 4th Annual Symposium on Human-Computer Interaction for Work*, pp. 1–18, 2025.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

Joshua S Gans. Designing ai-augmented peer review. Technical report, National Bureau of Economic Research, 2026.

ICLR. Policy on llm use for writing reviews or meta reviews, 2026. URL `https://iclr.cc/Conferences/2027/AIPolicyForReviewers`. Accessed September 5, 2026.

Elliot Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. *arXiv preprint arXiv:2506.07962*, 2025.

Seungone Kim, Dongkeun Yoon, Kiril Gashteovski, Juyoung Suk, Jinheon Baek, Pranjal Aggarwal, Ian Wu, Viktor Zaverkin, Spase Petkoski, Daniel R Schrider, et al. On the limits and opportunities of ai reviewers: Reviewing the reviews of nature-family papers with 45 expert scientists. *arXiv preprint arXiv:2605.20668*, 2026a.

Sunnie SY Kim, Wesley Hanwen Deng, Jennifer Wortman Vaughan, Buxin Su, Weijie Su, Alekh Agarwal, Sharon Li, Martin Jaggi, Daniel G Goldstein, Nihar B Shah, et al. Use and effects of llms in peer review: A randomized experiment and survey at icml 2026. *arXiv preprint arXiv:2609.19420*, 2026b.

Jon Kleinberg and Manish Raghavan. Algorithmic monoculture and social welfare. *Proceedings of the National Academy of Sciences*, 118(22):e2018340118, 2021.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024.

NeurIPS Communications Chairs. Reflections on the 2025 review process from the program committee chairs, 2025. URL https://blog.neurips.cc/2025/09/30/reflections-on-the-2025-review-process-from-the-program-committee-chairs/. Accessed September 5, 2026.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. A large-scale randomized study of large language model feedback in peer review. *Nature Machine Intelligence*, 8(3):326–336, 2026.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303, 2024.

Lieven Vandenberghe and Stephen Boyd. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 297–312, 2025.

Qichuan Yin and Linwei Xin. Synthetic but not infinite: How much llm-generated data to use in market research. *Available at SSRN 6078686*, 2026.

A Proofs for Section 2

A.1 Derivation of the Posterior Updates

In this subsection, we derive the posterior distributions before and after the AC's additional assessments.

Information in the panel mean. Under the assisted reports setting,

$$\bar{Y}^A = x + \beta + \frac{1}{K}\sum_{j=1}^{K} \widetilde{\epsilon}_j.$$

Independence of the shared and reviewer-specific errors gives $\bar{Y}^A \mid x \sim \mathcal{N}(x, T + \Sigma_A/K)$. Similarly, $\bar{Y}^H \mid x \sim \mathcal{N}(x, \Sigma/K)$. This establishes Equation (2).

The centered assisted reports satisfy

$$Y_j^A - \bar{Y}^A = \widetilde{\epsilon}_j - \frac{1}{K}\sum_{r=1}^{K} \widetilde{\epsilon}_r.$$

For each j, this difference has zero covariance with x, and

$$\mathrm{Cov}(Y_j^A - \bar{Y}^A, \bar{Y}^A) = \frac{\Sigma_A}{K} - \frac{\Sigma_A}{K} = 0.$$

All variables are jointly Gaussian. Consequently, the entire collection $(Y_j^A - \bar{Y}^A)_{j=1}^K$ is independent of $(x, \bar{Y}^A)$. The same argument applies to human-only reports, with Σ_A replaced by Σ.

Observing the full panel is equivalent to observing its mean and the centered reports. Because the centered reports are independent of $(x, \bar{Y}^g)$, they provide no additional information about x conditional on $\bar{Y}^g$. Thus $\bar{Y}^g$ is a sufficient statistic for x, and, for $g \in \{H, A\}$,

$$p(x \mid Y_1^g, \ldots, Y_K^g) = p\big(x \mid \bar{Y}^g, (Y_j^g - \bar{Y}^g)_{j=1}^K\big) = p(x \mid \bar{Y}^g).$$

Combining $\bar{Y}^g \mid x \sim \mathcal{N}(x, \Omega_g)$ with the prior $x \sim \mathcal{N}(m_0, \Lambda)$ gives

$$\log p(x \mid \bar{Y}^g) = -\frac{1}{2}(x - m_0)^\top \Lambda^{-1}(x - m_0) - \frac{1}{2}(\bar{Y}^g - x)^\top \Omega_g^{-1}(\bar{Y}^g - x) + C,$$

where C does not depend on x. Completing the square yields

$$x \mid Y_1^g, \ldots, Y_K^g \sim \mathcal{N}(m_g, V_g), \quad V_g = (\Lambda^{-1} + \Omega_g^{-1})^{-1}, \quad m_g = V_g\big(\Lambda^{-1} m_0 + \Omega_g^{-1} \bar{Y}^g\big),$$

as stated in Equation (3). In particular, the posterior mean and variance of $q = \theta^\top x$ are $\theta^\top m_g$ and $\theta^\top V_g \theta$, respectively.

Posterior after additional assessments. Fix an allocation $t \in \mathbb{R}_+^L$. Let $\mathcal{F}_g(t)$ denote the information generated by the submitted reports and the measurements Z_ℓ for which $t_\ell > 0$. Tasks with zero allocation contribute no observation. By independence of the measurement errors,

$$\log p(x \mid \mathcal{F}_g(t)) = -\frac{1}{2}(x - m_g)^\top V_g^{-1}(x - m_g) - \frac{1}{2}\sum_{\ell : t_\ell > 0} t_\ell (Z_\ell - a_\ell^\top x)^2 + C_t,$$

where C_t does not depend on x. Completing the square again gives

$$x \mid \mathcal{F}_g(t) \sim \mathcal{N}(m_g(t), V_g(t)), V_g(t) = \left(V_g^{-1} + \sum_{\ell=1}^{L} t_\ell a_\ell a_\ell^\top\right)^{-1},$$

$$m_g(t) = V_g(t)\left(V_g^{-1} m_g + \sum_{\ell : t_\ell > 0} t_\ell a_\ell Z_\ell\right).$$

For any square-integrable estimator $\widehat{q}$ based on $\mathcal{F}_g(t)$, the conditional squared-error decomposition is

$$\mathbb{E}\big[(\widehat{q} - q)^2 \mid \mathcal{F}_g(t)\big] = \big(\widehat{q} - \theta^\top m_g(t)\big)^2 + \theta^\top V_g(t)\theta.$$

The posterior mean $\theta^\top m_g(t)$ therefore minimizes posterior mean-squared error, and its risk is $\theta^\top V_g(t)\theta$, establishing Equation (4). For fixed t, the covariance $V_g(t)$ does not depend on the observed reports or measurement values, so the unconditional mean-squared error of this estimator has the same value. Taking $t = 0$ also establishes optimality of the reports-only estimator $\theta^\top m_g$.

Since $\Lambda \succ 0$ and $\Omega_g \succ 0$, we have $V_g \succ 0$. Consequently, $V_g(t)$ is well defined for every nonnegative allocation.

A.2 Proof of Proposition 2.2

Proof. Fix $g \in \{H, A\}$ and an attainable target $\varepsilon > 0$. Consider any feasible sequential policy with finite expected cost; policies with infinite expected cost cannot improve the benchmark. Let $N < \infty$ almost surely be its number of measurements, and let ℓ_n and $s_n > 0$ be the task and effort selected at step n. Each new measurement follows the stated model using a fresh independent standard Gaussian noise draw scaled by $s_n^{-1/2}$. Define the cumulative effort and total cost by

$$\tau_\ell = \sum_{n=1}^{N} s_n \mathbf{1}\{\ell_n = \ell\}, \qquad \tau = (\tau_1, \ldots, \tau_L), \qquad \mathcal{C} = \sum_{\ell=1}^{L} c_\ell \tau_\ell.$$

Since every $c_\ell > 0$, finite expected cost implies $\mathbb{E}[\tau_\ell] < \infty$ for every task.

Let $\mathcal{F}_N$ denote the complete observed history at stopping. Because task selection, effort allocation, and stopping depend only on observed information and independent randomization, their likelihood factors do not depend on x once this history is fixed. The posterior calculation in Appendix A.1 therefore applies to the stopped history, with precision contributions accumulated across measurements:

$$\mathrm{Cov}(x \mid \mathcal{F}_N) = \left(V_g^{-1} + \sum_{\ell=1}^{L} \tau_\ell a_\ell a_\ell^\top\right)^{-1} = V_g(\tau).$$

By the posterior-mean optimality established in the same appendix, any feasible final estimator satisfies

$$\mathbb{E}[\mathcal{R}_g(\tau)] \le \mathbb{E}[(\widehat{q} - q)^2] \le \varepsilon.$$

The risk $\mathcal{R}_g(t)$ is convex in t. Indeed,

$$\mathcal{R}_g(t) = \sup_{v \in \mathbb{R}^d} \left\{ 2\theta^\top v - v^\top V_g^{-1} v - \sum_{\ell=1}^{L} t_\ell (a_\ell^\top v)^2 \right\}$$

expresses it as a supremum of affine functions of t. Jensen's inequality then gives

$$\mathcal{R}_g(\mathbb{E}[\tau]) \le \mathbb{E}[\mathcal{R}_g(\tau)] \le \varepsilon.$$

Thus the deterministic allocation $\bar{t} = \mathbb{E}[\tau]$ is feasible and has the same cost as the policy's expected cost:

$$W_g(\varepsilon) \le \sum_{\ell=1}^{L} c_\ell \bar{t}_\ell = \mathbb{E}[\mathcal{C}].$$

Conversely, an optimal deterministic allocation $t^\star$ exists: the feasible set is nonempty and closed, and positive costs make its bounded-cost sublevel sets compact. Implementing $t^\star$ requires at most one measurement per task and hence stops after at most L measurements. By Appendix A.1, the resulting posterior mean of q has unconditional mean-squared error $\mathcal{R}_g(t^\star) \le \varepsilon$. Its cost is $W_g(\varepsilon)$, so the lower bound is attained. □

B Proofs for Section 3

B.1 Preliminaries

Throughout, $J(t) = \sum_{\ell=1}^{L} t_\ell a_\ell a_\ell^\top$ for $t \in \mathbb{R}_+^L$, and

$$P_g := \Lambda^{-1} + \Omega_g^{-1} = V_g^{-1}, \quad \Gamma := \Omega_H^{-1} - \Omega_A^{-1} = P_H - P_A,$$

$$w_g(t) := (P_g + J(t))^{-1}\theta = V_g(t)\theta, \quad \mathcal{R}_g(t) = \theta^\top w_g(t),$$

where the last two identities are Equation (4).

Positive definiteness. $\Omega_H = \Sigma/K \succ 0$. For $\Omega_A = T + \Sigma_A/K$, note that $\Sigma_A = (I-E)\Sigma_D(I-E) + E\Sigma E \succ 0$: for $v \neq 0$,

$$v^\top \Sigma_A v = \big((I-E)v\big)^\top \Sigma_D \big((I-E)v\big) + (Ev)^\top \Sigma (Ev) > 0,$$

because E is diagonal with entries in $[0,1]$, so $(1-e_i)v_i$ and $e_i v_i$ cannot both vanish at a coordinate with $v_i \neq 0$. Hence $\Omega_A \succ 0$ and $P_g \succ 0$ for $g \in \{H, A\}$. Since inversion reverses the Loewner order on positive definite matrices,

$$\Omega_A \preceq \Omega_H \quad \Longleftrightarrow \quad \Omega_A^{-1} \succeq \Omega_H^{-1} \quad \Longleftrightarrow \quad \Gamma \preceq 0. \tag{14}$$

Three facts about the risk. First, $\mathcal{R}_g$ is convex on $\mathbb{R}_+^L$, being the supremum $\sup_{v \in \mathbb{R}^d}\{2\theta^\top v - v^\top P_g v - \sum_\ell t_\ell (a_\ell^\top v)^2\}$ of affine functions of t (Appendix A.2). Second, $\mathcal{R}_g$ is nonincreasing in every coordinate, because $J(t)$ is nondecreasing in the Loewner order and inversion reverses it. Third, differentiating the inverse,

$$\frac{\partial \mathcal{R}_g}{\partial t_\ell}(t) = -\theta^\top (P_g + J(t))^{-1} a_\ell a_\ell^\top (P_g + J(t))^{-1}\theta = -\big(a_\ell^\top w_g(t)\big)^2. \tag{15}$$

Finally, if $J(t) \succ 0$ for some $t \ge 0$, then $(P_g + sJ(t))^{-1} \preceq (sJ(t))^{-1} \to 0$ as $s \to \infty$, so $\mathcal{R}_g(st) \to 0$ and $W_g(\varepsilon) < \infty$ for every $\varepsilon > 0$; the minimum in the definition of $W_g(\varepsilon)$ is then attained, as shown in Appendix A.2.

B.2 Proof of Proposition 3.2

Lemma B.1 (Exhausting all but one direction)**.** *Let $\Psi \succ 0$, let $\nu \in \mathbb{R}^d$ with $\|\nu\| = 1$, and let $B_\perp \in \mathbb{R}^{(d-1)\times d}$ have orthonormal rows orthogonal to ν. Then for every θ, the map $s \mapsto \theta^\top(\Psi + sB_\perp^\top B_\perp)^{-1}\theta$ is nonincreasing on $[0,\infty)$ and converges to $(\theta^\top\nu)^2/(\nu^\top\Psi\nu)$ as $s \to \infty$.*

Proof. Monotonicity holds because $\Psi + sB_\perp^\top B_\perp$ is nondecreasing in s. For the limit, let $Q = [B_\perp^\top \ \nu]$, an orthogonal matrix, and write $\Psi_{11} = B_\perp \Psi B_\perp^\top$, $\psi_{12} = B_\perp \Psi \nu$, $\psi_{22} = \nu^\top \Psi \nu > 0$. Then

$$Q^\top(\Psi + sB_\perp^\top B_\perp)Q = \begin{pmatrix} \Psi_{11} + sI & \psi_{12} \\ \psi_{12}^\top & \psi_{22} \end{pmatrix}.$$

By the block inverse formula, the $(2,2)$ entry of the inverse is $\left(\psi_{22} - \psi_{12}^\top(\Psi_{11} + sI)^{-1}\psi_{12}\right)^{-1} \to \psi_{22}^{-1}$, the $(1,2)$ block is $-(\Psi_{11} + sI)^{-1}\psi_{12}$ times that entry and tends to 0, and the $(1,1)$ block $\left(\Psi_{11} + sI - \psi_{12}\psi_{22}^{-1}\psi_{12}^\top\right)^{-1}$ tends to 0. Hence $(\Psi + sB_\perp^\top B_\perp)^{-1} \to \nu\nu^\top/(\nu^\top\Psi\nu)$. □

Proof of Proposition 3.2. Fix $\theta \neq 0$.

Dominance implies $W_A \leq W_H$. If $\Omega_A \preceq \Omega_H$, then $\Omega_A^{-1} \succeq \Omega_H^{-1}$ by Equation (14), so for every allocation $t \geq 0$,

$$P_A + J(t) \succeq P_H + J(t), \qquad \text{hence} \qquad V_A(t) \preceq V_H(t) \quad \text{and} \quad \mathcal{R}_A(t) \leq \mathcal{R}_H(t).$$

Every allocation that is feasible under the human-only protocol is therefore feasible under the assisted protocol, at the same cost, for every menu, cost vector, and target. Minimizing the cost gives $W_A(\varepsilon) \leq W_H(\varepsilon)$.

Failure of dominance is visible to some AC, with finite workloads. Suppose $\Omega_A \not\preceq \Omega_H$. By Equation (14), $\Gamma = P_H - P_A$ has a positive eigenvalue, so $U := \{\nu : \nu^\top\Gamma\nu > 0\}$ is a nonempty open cone. An open set is not contained in the hyperplane $\{\nu : \theta^\top\nu = 0\}$, so we may pick $\nu \in U$ with $\|\nu\| = 1$ and $\theta^\top\nu \neq 0$. Membership in U means $\nu^\top P_A\nu < \nu^\top P_H\nu$. Put

$$\bar{\varepsilon}_g := \frac{(\theta^\top\nu)^2}{\nu^\top P_g \nu}, \qquad \text{so that} \qquad \bar{\varepsilon}_A > \bar{\varepsilon}_H > 0,$$

and fix a target $\varepsilon \in (\bar{\varepsilon}_H, \bar{\varepsilon}_A)$. Let $B_\perp$ be as in Lemma B.1 and consider the menu consisting of the $d-1$ rows of $B_\perp$, each at cost 1, together with $a_d = \nu$ at cost $c_0 > 0$ (if $d = 1$, the menu is the single task $a_1 = \nu$ at cost c_0 and the limits below are trivial). Its directions span $\mathbb{R}^d$, so both workloads are finite by Appendix B.1.

Under H, Lemma B.1 with $\Psi = P_H$ shows that $\mathcal{R}_H(s,\dots,s,0) \downarrow \bar{\varepsilon}_H < \varepsilon$ as $s \to \infty$, so some finite s_0 is feasible and $W_H(\varepsilon) \leq (d-1)s_0 =: \bar{W}$, a bound that does not involve c_0. Under A, let $t \geq 0$ be any feasible allocation. Monotonicity of $\mathcal{R}_A$ in the first $d-1$ coordinates and Lemma B.1 with $\Psi = P_A + t_d\nu\nu^\top$, for which $\nu^\top\Psi\nu = \nu^\top P_A\nu + t_d$, give

$$\varepsilon \ \geq \ \mathcal{R}_A(t) \ \geq \ \lim_{s\to\infty} \mathcal{R}_A(s,\dots,s,t_d) = \frac{(\theta^\top\nu)^2}{\nu^\top P_A\nu + t_d}.$$

Hence $t_d \geq t_0 := (\theta^\top\nu)^2/\varepsilon - \nu^\top P_A\nu$, and $t_0 > 0$ because $\varepsilon < \bar{\varepsilon}_A$. Therefore $W_A(\varepsilon) \geq c_0 t_0$, and choosing $c_0 > \bar{W}/t_0$ yields $W_H(\varepsilon) \leq \bar{W} < c_0 t_0 \leq W_A(\varepsilon) < \infty$. This contradicts $W_A(\varepsilon) \leq W_H(\varepsilon)$ for every menu, cost vector, and target, which proves the converse of the proposition and exhibits the finite-workload instance claimed in the main text. □

B.3 The design problem

Let $\mathcal{D} := \{t \in \mathbb{R}^L : J(t) \succ 0\}$, an open convex set, and $\phi(t) := \theta^\top J(t)^{-1}\theta$ on $\mathcal{D}$. As in Appendix B.1, ϕ is convex and differentiable with $\partial_\ell\phi(t) = -(a_\ell^\top J(t)^{-1}\theta)^2$. The design problem in Equation (5) minimizes the linear function $c^\top t$ over $\mathcal{D} \cap \mathbb{R}_+^L$ subject to $\phi(t) \leq 1$. Throughout this subsection and the next, $\tilde{t}$ is a minimizer with $J(\tilde{t}) \succ 0$, as assumed in the main text, and

$$w^* = J(\tilde{t})^{-1}\theta, \qquad W_0 = c^\top\tilde{t}, \qquad \mathcal{S} := \{\ell : \tilde{t}_\ell > 0\}.$$

Lemma B.2 (Optimality conditions for the design). *(i)* $\theta^\top J(\tilde{t})^{-1}\theta = 1$*; equivalently* $\theta^\top w^* = 1$.
(ii) $(a_\ell^\top w^*)^2 = c_\ell/W_0$ *for* $\ell \in \mathcal{S}$ *and* $(a_\ell^\top w^*)^2 \le c_\ell/W_0$ *for every* ℓ*. In particular* $a_\ell^\top w^* \neq 0$ *on* $\mathcal{S}$*, and* $\{a_\ell\}_{\ell\in\mathcal{S}}$ *spans* $\mathbb{R}^d$.
(iii) w^* *is the same for every minimizer* $\tilde{t}$ *with* $J(\tilde{t}) \succ 0$.

Proof. The point $2\tilde{t}$ lies in $\mathcal{D}\cap\mathbb{R}^L_+$ with $\phi(2\tilde{t}) = \phi(\tilde{t})/2 \le \frac{1}{2} < 1$, because $J(2\tilde{t}) = 2J(\tilde{t})$ and $\phi(\tilde{t}) \le 1$ by feasibility; so Slater's condition holds and the Karush–Kuhn–Tucker conditions are necessary at the minimizer $\tilde{t}$ (Vandenberghe & Boyd, 2004, Section 5.5.3): there exist $\lambda \ge 0$ and $\eta \in \mathbb{R}^L_+$ with

$$c_\ell - \lambda\,(a_\ell^\top w^*)^2 - \eta_\ell = 0, \qquad \eta_\ell \tilde{t}_\ell = 0, \qquad \lambda\,(\phi(\tilde{t}) - 1) = 0, \qquad \ell = 1,\dots,L.$$

If $\lambda = 0$ then $\eta_\ell = c_\ell > 0$ for every ℓ, forcing $\tilde{t} = 0$ and $J(\tilde{t}) = 0$, a contradiction; so $\lambda > 0$ and $\phi(\tilde{t}) = 1$, which is (i). Multiplying the stationarity condition by $\tilde{t}_\ell$ and summing,

$$W_0 = c^\top \tilde{t} = \lambda \sum_\ell \tilde{t}_\ell (a_\ell^\top w^*)^2 = \lambda\, w^{*\top} J(\tilde{t}) w^* = \lambda\, \theta^\top J(\tilde{t})^{-1}\theta = \lambda.$$

Substituting $\lambda = W_0$ into the stationarity condition gives (ii); the spanning claim holds because $J(\tilde{t}) \succ 0$ is supported on $\mathcal{S}$.

For (iii), let t^1, t^2 be two minimizers in $\mathcal{D}\cap\mathbb{R}^L_+$, and let $t^s = (1-s)t^1 + st^2$ for $s \in [0,1]$. Each t^s is feasible with cost W_0, hence a minimizer, so $\psi(s) := \phi(t^s) = 1$ for every s by (i). With $\delta = t^2 - t^1$ and $w^s = J(t^s)^{-1}\theta$,

$$\psi''(s) = 2\left(J(\delta)w^s\right)^\top J(t^s)^{-1}\left(J(\delta)w^s\right) \ge 0,$$

so $\psi \equiv 1$ forces $J(\delta)w^s = 0$ for every s. At $s = 0$ this reads $J(t^2)w^1 = J(t^1)w^1 = \theta$, whence $w^1 = J(t^2)^{-1}\theta = w^2$. □

B.4 Full statement and proof of Theorem 3.3

For $g \in \{H, A\}$ let $\xi_g \in \mathbb{R}^L$ be supported on $\mathcal{S}$ and satisfy

$$\sum_{\ell\in\mathcal{S}} \xi_{g,\ell}\,(a_\ell^\top w^*)\,a_\ell = P_g\, w^*, \tag{16}$$

the precision the reports already contribute to each verification task. The linear map $\xi \mapsto \sum_{\ell\in\mathcal{S}} \xi_\ell (a_\ell^\top w^*) a_\ell$ from $\mathbb{R}^{\mathcal{S}}$ to $\mathbb{R}^d$ is onto, because $a_\ell^\top w^* \neq 0$ on $\mathcal{S}$ and $\{a_\ell\}_{\ell\in\mathcal{S}}$ spans $\mathbb{R}^d$ by Lemma B.2(ii); so such a vector exists, and it is unique when $|\mathcal{S}| = d$ (if $|\mathcal{S}| > d$, fix any such choice; the theorem below holds for each). Multiplying Equation (16) on the left by $w^{*\top}$ gives $\sum_{\ell\in\mathcal{S}} \xi_{g,\ell}(a_\ell^\top w^*)^2 = w^{*\top} P_g w^* > 0$, and $(a_\ell^\top w^*)^2 > 0$ on $\mathcal{S}$, so some $\xi_{g,\ell}$ is positive. The threshold in Theorem 3.3 is therefore finite:

$$\varepsilon_* := \min\left\{\tilde{t}_\ell/\xi_{g,\ell} :\ g \in \{H,A\},\ \ell \in \mathcal{S},\ \xi_{g,\ell} > 0\right\} \in (0,\infty), \tag{17}$$

the largest target below which every task of the report-free design receives strictly positive effort under both protocols.

Theorem B.3 (Theorem 3.3, full version). *Fix a menu, positive costs, and a rubric* $\theta \neq 0$ *for which the infimum in Equation (5) is attained at some* $\tilde{t}$ *with* $J(\tilde{t}) \succ 0$*. For every* $0 < \varepsilon < \varepsilon_*$ *and* $g \in \{H, A\}$*:*

(i) the allocation $t_g^* := \tilde{t}/\varepsilon - \xi_g$ *is optimal, and it uses exactly the tasks in* $\mathcal{S}$*;*
(ii) $W_g(\varepsilon) = W_0/\varepsilon - W_0\, w^{*\top} P_g\, w^*$*, which is Equation (6);*
(iii) after optimal verification, the posterior covariance between x *and* q *is* $V_g(t_g^*)\theta = \varepsilon\, w^*$*, the same under both protocols;*
(iv) $W_A(\varepsilon) - W_H(\varepsilon) = W_0\, w^{*\top}\Gamma\, w^*$*, which is Equation (7).*

Proof. Fix $g \in \{H, A\}$ and $0 < \varepsilon < \varepsilon_*$.

Step 1: feasibility and the verification direction. By the definition of ε_*, $t^*_{g,\ell} = \tilde{t}_\ell/\varepsilon - \xi_{g,\ell} > 0$ for $\ell \in \mathcal{S}$, while $t^*_{g,\ell} = 0$ otherwise. Using $J(\tilde{t})w^* = \theta$ and Equation (16),

$$\big(P_g + J(t^*_g)\big)\,\varepsilon w^* = \varepsilon P_g w^* + \sum_{\ell\in\mathcal{S}} \big(\tilde{t}_\ell - \varepsilon\xi_{g,\ell}\big)(a_\ell^\top w^*)\,a_\ell = \varepsilon P_g w^* + \theta - \varepsilon P_g w^* = \theta.$$

Hence $w_g(t^*_g) = \varepsilon w^*$, which is (iii), and $\mathcal{R}_g(t^*_g) = \theta^\top w_g(t^*_g) = \varepsilon\,\theta^\top w^* = \varepsilon$ by Lemma B.2(i). In particular t^*_g is feasible.

Step 2: optimality. The workload problem $\min\{c^\top t : t \in \mathbb{R}^L_+,\ \mathcal{R}_g(t) \le \varepsilon\}$ has a linear objective and a convex differentiable constraint, so the Karush–Kuhn–Tucker conditions are sufficient for optimality (Vandenberghe & Boyd, 2004, Section 5.5.3). Take the multiplier $\mu = W_0/\varepsilon^2$ for the risk constraint. By Equation (15) and Step 1,

$$\mu\,\big(a_\ell^\top w_g(t^*_g)\big)^2 = W_0\,(a_\ell^\top w^*)^2,$$

which by Lemma B.2(ii) equals c_ℓ for $\ell \in \mathcal{S}$, where $t^*_{g,\ell} > 0$, and is at most c_ℓ for $\ell \notin \mathcal{S}$, where $t^*_{g,\ell} = 0$. With $\mathcal{R}_g(t^*_g) = \varepsilon$, these are the KKT conditions, the multipliers of the sign constraints being $c_\ell - \mu(a_\ell^\top w_g(t^*_g))^2 \ge 0$. Thus t^*_g is optimal, which is (i).

Step 3: value. By Lemma B.2(ii) and Equation (16),

$$\sum_{\ell\in\mathcal{S}} c_\ell\xi_{g,\ell} = W_0 \sum_{\ell\in\mathcal{S}} (a_\ell^\top w^*)^2\xi_{g,\ell} = W_0\, w^{*\top} \sum_{\ell\in\mathcal{S}} \xi_{g,\ell}(a_\ell^\top w^*)a_\ell = W_0\, w^{*\top} P_g w^*,$$

so $W_g(\varepsilon) = c^\top t^*_g = W_0/\varepsilon - W_0\, w^{*\top} P_g w^*$, which is (ii). Subtracting the two protocols, the common terms cancel, including the prior precision Λ^{-1} inside P_g, and $P_H - P_A = \Gamma = \Omega_H^{-1} - \Omega_A^{-1}$ gives (iv). □

B.5 Attribute-specific assessments

The following corollary is the statement behind Equation (8). Write $\lambda_i = \Lambda_{ii}$, $\omega_{g,i} = (\Omega_g)_{ii}$, and $\pi_{g,i} := \lambda_i^{-1} + \omega_{g,i}^{-1}$, the precision on attribute i available before any assessment.

Corollary B.4 (Attribute-specific assessments)**.** *Let $L = d$, let task i measure attribute i, so that a_i is the ith coordinate unit vector, suppose $\theta_i \neq 0$ for every i, and let Λ, Σ, T, and Σ_A be diagonal. Put $S := \sum_{k=1}^d |\theta_k|\sqrt{c_k} > 0$. Then:*

(a) the infimum in Equation (5) is attained at the unique point $\tilde{t}_i = |\theta_i|S/\sqrt{c_i}$, with $W_0 = S^2$ and $w_i^* = \operatorname{sgn}(\theta_i)\sqrt{c_i}/S$;
(b) $\xi_{g,i} = \pi_{g,i}$ and $\varepsilon_* = \min_{g\in\{H,A\},\,i} |\theta_i|S/(\sqrt{c_i}\,\pi_{g,i})$;
(c) for $0 < \varepsilon < \varepsilon_*$ the optimal allocation under protocol g is unique, equals $t^*_{g,i} = |\theta_i|S/(\varepsilon\sqrt{c_i}) - \pi_{g,i} > 0$, and

$$W_g(\varepsilon) = \frac{S^2}{\varepsilon} - \sum_{i=1}^d c_i\pi_{g,i}, \qquad W_A(\varepsilon) - W_H(\varepsilon) = \sum_{i=1}^d c_i\Big(\frac{1}{\omega_{H,i}} - \frac{1}{\omega_{A,i}}\Big).$$

Proof. (a) With a_i the unit vectors, $J(t) = \operatorname{diag}(t)$, so $J(t) \succ 0$ if and only if every $t_i > 0$, and then $\theta^\top J(t)^{-1}\theta = \sum_i \theta_i^2/t_i$. For every feasible t, the Cauchy–Schwarz inequality gives

$$S^2 = \Big(\sum_{i=1}^d \sqrt{c_i t_i}\,\frac{|\theta_i|}{\sqrt{t_i}}\Big)^2 \le \Big(\sum_{i=1}^d c_i t_i\Big)\Big(\sum_{i=1}^d \frac{\theta_i^2}{t_i}\Big) \le \sum_{i=1}^d c_i t_i,$$

with equality throughout if and only if $t_i \propto |\theta_i|/\sqrt{c_i}$ and $\sum_i \theta_i^2/t_i = 1$, that is, $t_i = |\theta_i|S/\sqrt{c_i}$. This point is feasible and attains the bound, so it is the unique minimizer, $W_0 = S^2$, and $w_i^* = \theta_i/\tilde{t}_i = \operatorname{sgn}(\theta_i)\sqrt{c_i}/S$.

(b) With Λ, Σ, T, and Σ_A diagonal, Ω_g and P_g are diagonal with $(P_g)_{ii} = \pi_{g,i}$, and $\mathcal{S} = \{1,\dots,d\}$. Equation (16) reads $\xi_{g,i}w_i^* = \pi_{g,i}w_i^*$ coordinatewise, and $w_i^* \neq 0$, so $\xi_{g,i} = \pi_{g,i}$; Equation (17) then gives the stated ε_*.

(c) Here $P_g + J(t) = \mathrm{diag}(\pi_g + t)$, so the workload problem is

$$\min_{t \geq 0} \sum_{i=1}^{d} c_i t_i \quad \text{subject to} \quad \sum_{i=1}^{d} \frac{\theta_i^2}{\pi_{g,i} + t_i} \leq \varepsilon.$$

Let $s_i = \pi_{g,i} + t_i$. For every feasible t, Cauchy–Schwarz as in (a) gives $S^2 \leq (\sum_i c_i s_i)(\sum_i \theta_i^2/s_i) \leq \varepsilon \sum_i c_i s_i$, hence

$$\sum_{i=1}^{d} c_i t_i \;\geq\; \frac{S^2}{\varepsilon} - \sum_{i=1}^{d} c_i \pi_{g,i},$$

with equality if and only if $s_i = |\theta_i| S/(\varepsilon\sqrt{c_i})$ for every i and the constraint binds. The corresponding $t^*_{g,i} = |\theta_i| S/(\varepsilon\sqrt{c_i}) - \pi_{g,i}$ is nonnegative for every i exactly when $\varepsilon \leq |\theta_i| S/(\sqrt{c_i}\pi_{g,i})$ for every i, and strictly positive for $\varepsilon < \varepsilon_*$. It is then feasible and attains the lower bound, so it is the unique optimal allocation and $W_g(\varepsilon) = S^2/\varepsilon - \sum_i c_i \pi_{g,i}$. Subtracting the two protocols cancels S^2/ε and the prior precisions λ_i^{-1}, leaving $\sum_i c_i(\pi_{H,i} - \pi_{A,i}) = \sum_i c_i(\omega_{H,i}^{-1} - \omega_{A,i}^{-1})$. This agrees with Theorem B.3: here $t^*_g = \tilde{t}/\varepsilon - \xi_g$ and $W_0\, {w^*}^\top P_g w^* = \sum_i c_i \pi_{g,i}$. □

Remark B.5. Diagonality of Λ is used only for the explicit ε_* and t^*_g. Since w^* depends only on the menu, the costs, and the rubric, Theorem B.3(iv) gives $W_A(\varepsilon) - W_H(\varepsilon) = W_0\, {w^*}^\top \Gamma w^* = \sum_i c_i(\omega_{H,i}^{-1} - \omega_{A,i}^{-1})$ for every $\Lambda \succ 0$ whenever Σ, T, and Σ_A are diagonal and $0 < \varepsilon < \varepsilon_*$, with ε_* as in Equation (17).

C Proofs for Section 4

C.1 Preliminaries

Panel covariance under input level E. With input level E, the assisted reports are $Y_j^A = x + (I-E)b + (I-E)\zeta_j + E\epsilon_j$, so the panel mean has error $(I-E)b + (I-E)\bar{\zeta} + E\bar{\epsilon}$ with the three terms independent, and

$$\Omega_A(E) = (I-E)T_D(I-E)^\top + \frac{(I-E)\Sigma_D(I-E)^\top + E\Sigma E^\top}{K} = (I-E)\,\Omega_D\,(I-E)^\top + E\,\Omega_H\,E^\top,$$

using $\Omega_D = T_D + \Sigma_D/K$ and $\Omega_H = \Sigma/K$. For $E = eI$ this is $(1-e)^2\Omega_D + e^2\Omega_H$, whose derivative in e is $-2(1-e)\Omega_D + 2e\Omega_H$, equal to $-2\Omega_D$ at $e = 0$. This is Equation (9). Since $(1-e)^2 + e^2 > 0$, $\Omega_A(eI) \succ 0$ for every $e \in [0,1]$.

The ratio r. Substituting $u = \Omega_H^{-1/2} v$,

$$\max_{u \neq 0} \frac{u^\top \Omega_D u}{u^\top \Omega_H u} = \max_{v \neq 0} \frac{v^\top \Omega_H^{-1/2} \Omega_D \Omega_H^{-1/2} v}{v^\top v} = \lambda_{\max}(M), \qquad M := \Omega_H^{-1/2} \Omega_D \Omega_H^{-1/2} \succ 0.$$

Write $C := \Sigma^{-1/2}(T_D + \Sigma_D)\Sigma^{-1/2}$ and $B := \Sigma^{-1/2} T_D \Sigma^{-1/2}$, so that $\kappa = \lambda_{\max}(C)$ and $\rho = \lambda_{\max}(B)$. Since $\Omega_H^{-1/2} = \sqrt{K}\,\Sigma^{-1/2}$ and $\Omega_D = (KT_D + \Sigma_D)/K$,

$$M = \Sigma^{-1/2}(KT_D + \Sigma_D)\Sigma^{-1/2} = C + (K-1)B. \tag{18}$$

C.2 Proof of Theorem 4.1

Equivalence. Fix $e \in [0,1]$. Applied, for each rubric $\theta \neq 0$, to the assisted panel with covariance $\Omega_A(eI)$, Proposition 3.2 says that $W_A(\varepsilon; eI) \leq W_H(\varepsilon)$ for every assessment menu, positive cost vector, and target if and only if $\Omega_A(eI) \preceq \Omega_H$. This condition does not involve θ, so statement (b) is equivalent to $\Omega_A(eI) \preceq \Omega_H$. It remains to show that $\Omega_A(eI) \preceq \Omega_H$ if and only if $e \geq e_H$. Conjugating by $\Omega_H^{-1/2}$ and using Equation (9),

$$\Omega_A(eI) \preceq \Omega_H \iff (1-e)^2 M + e^2 I \preceq I \iff (1-e)^2 M \preceq (1-e^2)\, I.$$

At $e = 1$ both sides vanish, so the condition holds, and $1 \geq e_H$ because $e_H < 1$. For $e \in [0, 1)$, dividing by $1-e > 0$ gives $(1-e)M \preceq (1+e)I$, that is, $(1-e)\, r \leq 1+e$, that is, $e \geq (r-1)/(r+1)$; together with $e \geq 0$ this is $e \geq e_H$. This proves (a)⇔(b). In particular, for $e < e_H$ the panel covariance is not dominated, and Proposition 3.2 provides, for any rubric, an assessment menu, positive costs, and a target with $W_A(\varepsilon; eI) > W_H(\varepsilon)$, as stated after the theorem.

Bounds. By Equation (18) and $\Sigma_D \succeq 0$, $M - KB = \Sigma^{-1/2}\Sigma_D\Sigma^{-1/2} \succeq 0$, so $r \geq \lambda_{\max}(KB) = K\rho$. By subadditivity of the largest eigenvalue, $r = \lambda_{\max}(C + (K-1)B) \leq \kappa + (K-1)\rho$. The function $f(x) := (x-1)_+/(x+1)$ is nondecreasing on $[0, \infty)$ and $e_H = f(r)$, hence $f(K\rho) \leq e_H \leq f(\kappa + (K-1)\rho)$, which is Equation (10). □

C.3 Consequences of Equation (10)

Consequences. If $(K-1)\rho \leq 1-\kappa$, the numerator of the upper bound is $(\kappa + (K-1)\rho - 1)_+ = 0$, so $e_H = 0$. If $\rho > 1/K$, the lower bound is $(K\rho - 1)/(K\rho + 1) > 0$, so $e_H > 0$ whatever Σ_D is. Because $B \preceq C$ when $\Sigma_D \succeq 0$, always $\rho \leq \kappa$; if $T_D + \Sigma_D \prec \Sigma$, then $C \prec I$ and $\kappa < 1$, so f nondecreasing gives

$$e_H \leq f\big(\kappa + (K-1)\rho\big) \leq f\big(1 + (K-1)\rho\big) = \frac{(K-1)\rho}{2 + (K-1)\rho},$$

which for $K = 3$ equals $\rho/(1+\rho)$; this is at most ρ since $1 + \rho \geq 1$, and below $1/2$ because $\rho < 1$.

C.4 Proof of Proposition 4.2

Equation (11). The quantities W_0 and w^* in Theorem 3.3 are determined by the assessment menu, the costs, and the rubric alone, while the panel enters only through Ω_A. Applying the theorem to the assisted panel with covariance $\Omega_A(E)$, with $\varepsilon_*(E)$ its threshold, and subtracting the human-only expression gives Equation (11) for $0 < \varepsilon < \varepsilon_*(E)$. Since $W_0 > 0$, the right-hand side is nonpositive if and only if $w^{*\top}\Omega_A(E)^{-1}w^* \geq w^{*\top}\Omega_H^{-1}w^*$.

e_W is well defined. Let $g(e) := w^{*\top}\Omega_A(eI)^{-1}w^*$ for $e \in [0, 1]$; it is continuous because $\Omega_A(eI) \succ 0$ depends continuously on e, and $g(1) = w^{*\top}\Omega_H^{-1}w^*$ because $\Omega_A(I) = \Omega_H$. The set in Equation (12) is $\mathcal{I} := \{e \in [0, 1] : \min_{e' \in [e,1]} g(e') \geq g(1)\}$. It contains 1, and $e \mapsto \min_{[e,1]} g$ is continuous and nondecreasing, so $\mathcal{I}$ is a nonempty closed interval containing 1; its minimum e_W exists, and every $e \geq e_W$ belongs to $\mathcal{I}$.

First claim. For $e \geq e_W$ we have $e \in \mathcal{I}$, hence $g(e) \geq g(1)$, and Equation (11) with $E = eI$ gives $W_A(\varepsilon; eI) - W_H(\varepsilon) = W_0\,(g(1) - g(e)) \leq 0$ for every $0 < \varepsilon < \varepsilon_*(eI)$.

Second claim. By the proof of Theorem 4.1, $\Omega_A(e'I) \preceq \Omega_H$ for every $e' \geq e_H$; inverting, $\Omega_A(e'I)^{-1} \succeq \Omega_H^{-1}$ and $g(e') \geq g(1)$ for every $e' \in [e_H, 1]$. Thus $e_H \in \mathcal{I}$ and $e_W \leq e_H$. □

C.5 Attribute-specific Input

Suppose the primitive covariances are diagonal and $E = \mathrm{diag}(e_1, \ldots, e_d)$. Then $T(E)$, $\Sigma_A(E)$, and, by the first display of this appendix, $\Omega_A(E)$ are diagonal, with $(\Omega_A(E))_{ii} = (1 - e_i)^2\omega_{D,i} + e_i^2\omega_{H,i}$, where $\omega_{D,i} = \tau_i + \gamma_i/K$ and $\omega_{H,i} = \sigma_i/K$. For attribute-specific assessments and diagonal Λ, the reduction of Section 3 therefore applies with $\omega_{A,i} = \omega_{A,i}(E)$, and Equation (8) becomes Equation (13) for $0 < \varepsilon < \varepsilon_*(E)$.

The right-hand side of Equation (13) is linear in c. If every term in parentheses is nonpositive, it is nonpositive for every positive cost vector; if the ith term is positive, taking c_i large enough makes it positive. Hence the excess workload is removed for every cost vector if and only if $(1 - e_i)^2\omega_{D,i} + e_i^2\omega_{H,i} \leq \omega_{H,i}$ for every i. Dividing by $\omega_{H,i}$, this is the scalar case of the computation in the proof of Theorem 4.1, with M replaced by $r_i = \omega_{D,i}/\omega_{H,i} = (K\tau_i + \gamma_i)/\sigma_i$; it holds if and only if $e_i \geq e_{H,i} = (r_i - 1)_+/(r_i + 1)$. The conditions are coordinatewise, so the set of such E is the product of the intervals $[e_{H,i}, 1]$, whose smallest element is E_H; and $e_{H,i} > 0$ if and only if $r_i > 1$. Finally, with diagonal primitives $M = \mathrm{diag}(r_1, \ldots, r_d)$, so $r = \max_i r_i$ and, as f is nondecreasing, $e_H = f(r) = \max_i f(r_i) = \max_i e_{H,i}$.

C.6 THE INSTANCE OF SECTION 3

With $K = 3$, $\Sigma = 3I$, $T_D = \mathrm{diag}(1/3, 1)$, and $\Sigma_D = \mathrm{diag}(1/2, 2/3)$: $C = \mathrm{diag}(5/18, 5/9)$ and $B = \mathrm{diag}(1/9, 1/3)$, so $\kappa = 5/9$, $\rho = 1/3$, and $M = C + 2B = \mathrm{diag}(1/2, 11/9)$ gives $r = 11/9 = \kappa + 2\rho$; hence $e_H = (2/9)/(20/9) = 1/10$ attains the upper bound in Equation (10). The attribute thresholds are $e_{H,1} = 0$ and $e_{H,2} = 1/10$, so $E_H = \mathrm{diag}(0, 1/10)$ and $\Omega_A(E_H) = \mathrm{diag}(1/2, 1) \preceq \Omega_H = I$.

For the assessment problem of Figure 1, $c = (1, 25)$ and Equation (13) reads

$$G(e_1, e_2) = \Big(1 - \frac{1}{\omega_1(e_1)}\Big) + 25\Big(1 - \frac{1}{\omega_2(e_2)}\Big), \qquad \omega_1(e) = \tfrac{1}{2}(1-e)^2 + e^2, \quad \omega_2(e) = \tfrac{11}{9}(1-e)^2 + e^2.$$

Input on the second attribute alone: $G(0, e_2) = 0$ if and only if $\omega_2(e_2) = 25/24$, that is, $160e_2^2 - 176e_2 + 13 = 0$, whose root in $[0, 1]$ is $e_2 = (22 - \sqrt{354})/40 \approx 0.080$; since $\omega_2 \le 25/24$ on $[e_2, 1]$, the excess workload stays removed beyond it. On the uniform path, $G(e, e) = 0$ at $e \approx 0.073$, and $G(e, e) \le 0$ for all larger e: both terms are nonincreasing on $[0, 1/3]$, and for $e \ge 1/10$ both are nonpositive because $\omega_1 \le 1$ on $[0, 1]$ and $\omega_2 \le 1$ on $[1/10, 1]$. Hence $e_W \approx 0.07$, against $e_H = 1/10$.

D EXPERIMENT DETAILS

D.1 DATASET

Of the 82 released papers, 54 have at least three human reviews with complete item text and three released AI reviews (one each from GPT-5.2, Claude Opus 4.5 and Gemini 3.0 Pro). When a paper has more than three human reviews, first three reviews in the dataset are chosen.

D.2 REPORTS AND ITEMS RANKING

All model calls use `gpt-6-sol` with medium reasoning effort.

Capped human report H. The released AI reviewers were instructed to write at most five significant items. Human reviews are longer and more variable (mean 8.69 items, median 7). To compare reports of similar length, we cap human reviews at five items. For review with more than five items, the model selects exactly five items with the prompt below.

```
You are prioritizing an existing scientific peer-review report
that contains more than five review items. Select EXACTLY
FIVE of its items, ordered from highest to lowest priority,
so that the five retained items preserve as much
of the reviewer's substantive information as possible.

Prioritize concerns that, if valid, materially affect the
paper's central claims, technical validity, conclusions,
originality, or interpretation. Use the support and
specificity present in the supplied text to distinguish
substantive concerns from generic or unsupported
assertions. Minor style, formatting, and purely
administrative comments have lower priority.
Prefer items that express different concerns; select two
items with the same specific concern and stance only if no
other item remains.

You have the review text, not an independent investigation
of the paper. Do not invent evidence, repair a claim,
rewrite an item, or infer expert annotations. Select only
the supplied item IDs. Do not combine items. Return JSON
with selected_ids and a short selection_reason.
```

Human-informed revision M_α**.** We randomly draw $\lfloor \alpha n + u \rfloor$ items from human reviews H, where n is the number of items in H and $u \sim \mathrm{Uniform}(0, 1)$. The expected packet size is exactly αn. The editor receives the AI draft and the packet, each item shown with its source role, and returns at most five original identifiers. The prompt is shown as below.

```
An AI draft is the starting point for a scientific peer-
review report. A human reviewer has supplied the additional
original items below. Revise the selection of items for the
final report using both sources.

Select AT MOST FIVE of its most scientifically significant
substantive review items, ordered from highest to lowest
priority.

Prioritize concerns that, if valid, materially affect the
paper's central claims, technical validity, conclusions,
originality, or interpretation. Use the support and
specificity present in the supplied text to distinguish
substantive concerns from generic or unsupported
assertions. Minor style, formatting, and purely
administrative comments have lower priority.
Do not select two items that express the same specific
concern and stance. Aim to retain four or five items: when
a lower-priority item still raises a
distinct substantive concern, include it rather than
stopping early. Retain fewer than four only if the report
contains fewer than four distinct substantive concerns.

Keep strong draft items; replace or omit them when the
supplied human material provides a more significant, better-
supported concern. Do not prefer either source merely
because it is human or AI. There is no quota for items from
either source. Identical or paraphrased repetitions are not
independent corroboration.

Select only original IDs from this draft and this human
packet. Do not invent, rewrite, combine, or split items.
Return selected_ids in priority
order and a short selection_reason.
```

Item group To count distinct items, all items of a paper are partitioned once into items groups by the model, using a strict same-concern rule.

```
You are indexing the review items written about one
scientific paper. Partition ALL supplied items into groups.
Two items may share a group only when they express the same
specific scientific concern with a compatible stance.
Referring to the same experiment, section, broad topic, or
weakness category is not sufficient. A validity critique
and a novelty critique are different concerns. Do not merge
opposing positions. Avoid chains in which
A resembles B and B resembles C but A and C express
different concerns.

Every supplied item ID must appear in exactly one group; an
item without a match forms a group of its own. Do not
rewrite, summarize, or judge the
items. Return JSON with groups, a list of lists of item IDs.
```

Panel coverage N_{FP} counts the groups of a panel that contain at least one FP item.

Simulated AC For a panel condition the simulated AC receives every concern group visible to that panel and prompted by

```
You are preparing a prioritized extractive feedback list
for an area chair. The supplied groups collect variants of
the same specific concern.
For each group, choose the one eligible ORIGINAL item that
best states a scientifically important, credible, specific,
and adequately supported concern. Then rank all groups from
highest to lowest priority for feedback.

Judge only the material supplied. Do not prefer an item
because of human or AI origin. Repeated text inherited from
a draft is not independent corroboration.
Pay attention to substantive corrections or qualifications
when present.

The first five items will form the short feedback list.
Include lower-priority groups later so that the entire
ordered list is available. Choose exactly
one eligible original variant per group. Do not rewrite or
combine items. Some records may be marked context_only:
they may inform your judgment
but their IDs may not be selected.

Return ranked_items, each with group_id and item_id, in
priority order.
```

inference. We report means over papers and pointwise 95% percentile intervals from 10,000 paired paper-bootstrap resamples.