Fighting Fire with Fire: Infusing AI into Peer Review to Sustain Quality Scholarship

Hemant K. Bhargava Sarah H. Bana Zhe Zhang Pantelis Loupos
Daniel Zantedeschi Laura Brandimarte Vidyanand Choudhary
J. Frank Li

March 2, 2026

Abstract

The accelerating use of generative AI tools in research raises a looming crisis in academic publishing: while author productivity increases, peer review capacity remains constrained, threatening the integrity and timeliness of scholarly evaluation. In this paper, we propose a hybrid review framework that integrates large language models (LLMs) into the journal review process as structured first-line reviewers. Our approach outlines a concrete editorial workflow wherein LLM-generated reviews are presented to authors for their response before human review. We then analyze our proposal through a formal analytical model and demonstrate that this AI-augmented process reduces the reviewer's burden and improves decision accuracy, particularly in response to increasing submission volumes. We also address key concerns regarding confidentiality, accountability, and accuracy by advocating for journal-specific AI systems and human-in-the-loop oversight. Our framework provides a path forward for adapting academic publishing to the AI era while preserving its normative and epistemic standards. More broadly, our advocacy is not for a particular approach for infusing AI, but about the vital need for journals to find some suitable approach versus living in a status quo where many reviewers engage in isolated use with unvetted AI tools that are not tuned to the journals' objectives.

Keywords: Large Language Models, Hybrid Human–AI Workflows, Generative AI in Science,

Academic Publishing.

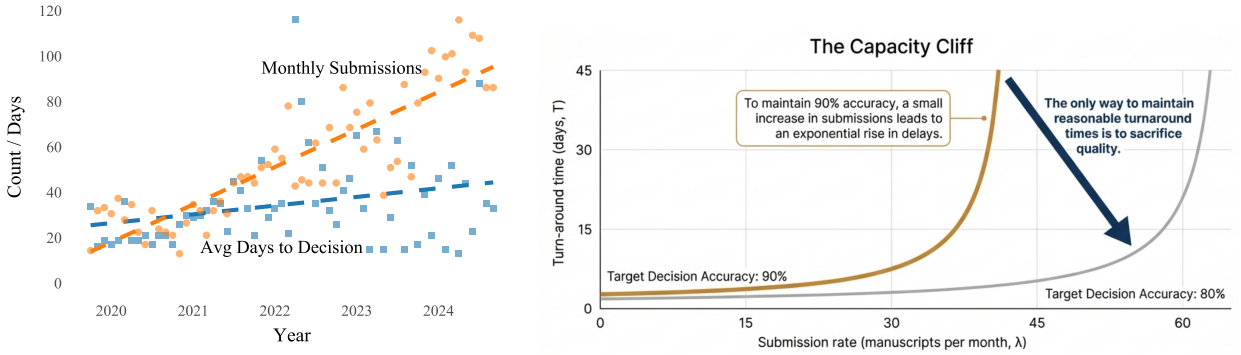
1 Introduction and Motivation

Peer review is essential to academic research and the profession. It requires multiple expert reviewers, with specialized knowledge, to devote substantial time in assessing the quality of submitted articles. As any operations system, peer review faces substantial stress when supply of submissions overcomes the capacity to process them. Such a scenario is occurring now as artificial intelligence (AI) tools, especially generative AI and large language models, have accelerated authors' ability to generate research articles. For instance, Google's AI Co-Scientist has successfully generated novel, literature-grounded hypotheses for biology labs, some of which aligned with hypotheses developed independently by the same labs (Gottweis et al., 2025). Similarly, Gans (2025b) reports that an AI co-author helped draft a full paper within a single day. Mathematician Terence Tao has likewise explored AI collaboration on proofs and analyses (Tao, 2024). Ludwig and Mullainathan (2024) and Manning et al. (2024) indicate that AI holds potential for generating and testing hypotheses within the social sciences.

Recent editorial commentary in *Information Systems Research* underscores that generative AI is already reshaping research practices and disciplinary expectations while warning that the community must develop new norms to manage these shifts responsibly (Gopal et al., 2025). From a recent report by a top publisher: "Peer review ... is under unprecedented pressure. Too many submissions are funnelled through a limited pool of reviewers, creating unsustainable workloads and threatening the effectiveness, speed and fairness of peer review" (Cambridge University Press, 2025). This article articulates the operations crisis confronting academic peer review, develops a formal model to capture the essence of the process, and proposes an AI-infused peer review workflow and applies an extended model to examining whether, under what conditions, and to what extent, this new approach can help address the crisis. Our process and model exemplify how AI can be integrated, not mandate the only way to do so.

1.1 Impending Crisis: Operational Strain and the "Shadow AI" Problem

As AI increases submissions, academic journals are finding it harder to distinguish high-quality work while maintaining timely decisions. At an anonymous journal used for illustration, submissions increased by roughly 300% between 2019 and 2025 while the effective reviewer pool remained essentially flat. Over the same period, time-to-first-decision rose from about 20 days to about 40 days (Figure 1a). A similar concern is noted by government agencies and in the popular and academic press (National Institutes of Health, 2025; Richardson et al., 2025). Operations theory suggests a nonlinear congestion effect: as utilization approaches one, small inflow shocks translate into large delay increases—a *capacity cliff*. Fig. 1b illustrates this pattern based on a stylized operations model of journals (presented in §2), with nonlinear deterioration in the critical metrics that define a journal's performance, namely turnaround time and decision accuracy.



(a) Capacity pressure and volatility (anonymous journal)

(b) Congestion nonlinearity: the capacity cliff

Figure 1: Capacity pressure and a governance hypothesis. Panel (a) shows rapid growth in submissions and rising time-to-first-decision, followed by apparent stabilization with higher volatility. Panel (b) provides a stylized illustration of the nonlinear delay–quality trade-off near capacity: as utilization rises, journals face pressure to either accept longer delays or relax effective standards to maintain throughput.

A related problem is that reviewers (and editors) are already engaged in a *Shadow AI* process, using AI tools in an opaque and un-governed manner to assist in their review tasks. As policy, many academic journals and publishers restrict or ban AI tools in peer review due to concerns over confidentiality, accountability, and accuracy (Bhargava et al., 2025; Bhargava and Tokarskaya, 2025;

Peters and Chin-Yee, 2025). In reality, however, informal reports and detection-based studies estimate that a nontrivial fraction of reviews contain AI-generated text (Latona et al., 2024; Liang et al., 2024a). These estimates are best interpreted as lower bounds: "light-touch" use (outlining, polishing, summarizing) may leave weak traces even when it materially changes effort and content. For journal leadership, the operational implication of this uneven adoption is opaque and hard to measure. Moreover, the journal is no longer observing the same process it *believes* it is managing. When multiple reviewers privately rely on the same general-purpose LLMs to evaluate a manuscript, their reviews become structurally correlated, eroding the independence of judgment that journals rely upon. Unlike a managed process where this correlation is observable, shadow AI introduces hidden biases and unobserved correlations that secretly degrade decision accuracy. Reviewer-initiated use of public interfaces can create uncontrolled retention, third-party logging, or jurisdictional ambiguity.

1.2 Solution: Managed, AI-Augmented, Peer Review

Although AI use for peer review is problematic when it is opaque, uncontrolled, heterogeneous, and undisclosed, there is credible evidence of upside in using AI. Comparative studies report substantial overlap between LLM critiques and human critiques, with performance often strongest on lower-quality submissions where screening and triage dominate (Liang et al., 2024b). LLMs tend to perform best on *auditable diagnostics* and worst on *disciplinary valuation*. They can be helpful in identifying clarity problems, missing reporting elements, internal inconsistencies, and other structured checks; they are less reliable for novelty, contribution, construct validity, and the appropriate tolerance for intellectual risk (Berente and Recker, 2025).

The managerial problem is therefore not "*AI or no AI*" but how to move from opaque adoption to *managed* adoption of AI: a journal-provided, journal-scoped AI diagnostic tool applied *before* human review, paired with a short author right-of-response, and instrumented through structured

logging. The AI output is explicitly *non-final*: it is an input to editorial judgment, not a recommendation and not a substitute for reviewers. Crucially, we are asking for a *journal-specific* AI reviewer and process. General-purpose AI tools lack alignment with a particular journal's standards, methodologies, and values. Therefore, journals should *scope* AI to tasks where criteria can be specified and outputs can be validated, and should prevent scope creep into value judgments. The AI model must be tailored, trained, and vetted according to the specific criteria of the target journal, focusing its capabilities on aspects appropriate for that journal (e.g., mechanical checks, writing quality, or methodological soundness), while reserving intellectual labor for human reviewers. Confidentiality and IP concerns further strengthen the case for a journal-led workflow.

Table 1: AI-in-Peer-Review Literature: Overlap and Gaps.

Theme	**What existing evidence supports**	**What remains missing (this paper's focus)**
Prevalence of AI use	Nontrivial AI-generated content in reviews, often undisclosed (Latona et al., 2024; Liang et al., 2024a).	An auditable definition of "AI involvement" and a way to monitor drift over time.
Helpfulness and overlap	LLM critiques can overlap with human critiques and appear most reliable for screening/triage contexts (Liang et al., 2024b).	A translation from "helpful feedback" into a decision-and-capacity operating model.
Limits of judgment	LLMs are weaker on novelty, contribution, and field-specific valuation (Berente and Recker, 2025).	A scoped workflow that assigns AI to auditable diagnostics and prevents scope creep.
Integrity tools in practice	Some venues already deploy bounded AI checks (screening/compliance) (Hosseini and Resnik, 2025).	A governed extension with structured outputs, versioning, and estimation of dependence/overlap.
Hidden dependence risk	Common-mode influence is plausible when reviewers consult similar tools.	A measurable overlap parameter plus levers (sequencing, guidance, blinding) to control it.
Policy posture	Allow/ban framing dominates; enforcement and measurement are limited.	An adoption rule stated in estimable quantities, plus a pilot blueprint to learn before scaling.

Table 1 summarizes the goals and principles underlying our proposal for a governed AI in peer review. Crucially, managed AI (vs. both shadow AI or permissive policies) enables the measurement of the effects of using AI. It enables treatment of AI as an informational input with two

trackable properties: *reliability* (noise/precision of the AI signal) and *overlap* with human judgment (the degree to which reviewer assessments move with the AI once the AI is observed). These quantities can be estimated from structured records and used to design workflows that preserve independence. We ask the operational questions: *When does AI add independent information, when does it become redundant, and what workflow choices keep the system in the high-value regime?* To formalize our measurement framework, we extend our journal operations model by infusing AI into the process. §3 turns this into a concrete workflow design and a pilotable adoption rule. Manuscripts arrive over time; humans (or humans and AI in the governed-AI case) produce noisy signals; and the editor aggregates them to meet a target evidentiary standard. The central governance risk is *overlap*: when reviewers see the AI report, their errors may become correlated with the AI's errors (anchoring). When overlap is low, the AI contributes independent information and can reduce burden or improve precision. When overlap is high, the AI becomes redundant and should receive little (or no) weight. This yields a simple operational implication: journals should scale AI only when measured overlap remains below the redundancy boundary, and use workflow controls (sequencing, guidance, rubric discipline, scope) to keep the system in that informative region.

2 A Benchmark Model of Journal Operations

We frame an academic journal's operations as a machine designed to distinguish high-quality scholarship from faulty or less impactful work, subject to the journal's and community's values and objectives. This machine receives a flow of submissions as its input, and through a resource-constrained evaluation process, it produces an output of accepted or rejected papers (see Fig. 2). It has two performance elements, *speed* (turnaround time) and *accuracy* (correctly accepting good papers and rejecting flawed ones), which depend on the *per capita* resources available. As mentioned earlier in §1.1, the increase in submission rate threatens journal integrity, presenting a poten-

tial capacity cliff in which the journal must face either worsening turnaround times or worsening decision accuracy (or both). Journals will be confronted with a Hobson's choice: either become less adept at distinguishing high-quality submissions or tolerate untenable delays.

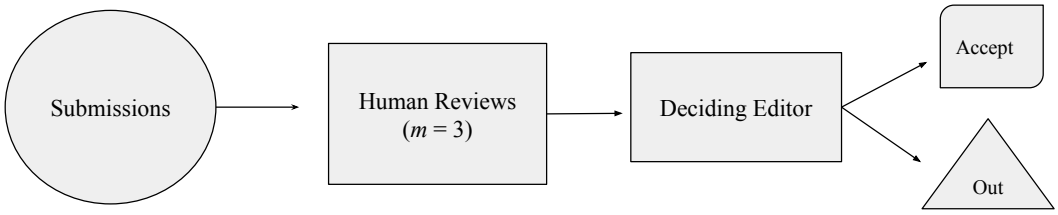


Figure 2: Simplified model of the current review process. For clarity, we model the review process as a single stage process. We also visually omit potential desk rejection for clarity.

Formally, we model journal review as a stylized single-round multi-step process yielding an "in" or "out" decision. Manuscript arrival is treated as a Poisson process with submission rate λ. Each manuscript has latent quality that follows distribution F, i.e., $Q \sim F$. Each manuscript is assigned to m human reviewers, each of whom provides a quality signal Q_{H_i} to a Deciding Editor (DE). Reviewer signals can be noisy, with the error rate increasing in volume (or temporal pressure) and reducing with reviewer skill. The DE's error rate depends on the noisiness of the reviewer signals being aggregated.[1] This parsimonious setup blends multi-round review into one stage so we can focus on decision accuracy (whether a manuscript exceeds a journal quality threshold) rather than iterative paper improvement. A valuable extension may be to consider improvements to paper quality endogenous to the review process. All notation introduced in this section is summarized in Table 2 for ease of reference.

Human reviewer signals and aggregation. When reviewing a paper, each reviewer i on the review team uses time τ to estimate an unbiased but noisy signal of the paper's quality,

$$Q_{H_i} \;=\; Q + \varepsilon_{H_i}, \qquad \varepsilon_{H_i} \sim \mathcal{N}\left(0,\; \frac{\sigma_0^2}{s\,\tau}\right), \tag{1}$$

[1] While some journals involve an Associate Editor, we omit this for simplicity.

Table 2: Notation summary for the analytical model

Symbol	*Definition*
Q	Latent quality of each manuscript, following distribution $\sim F$
λ	Submission arrival rate (Poisson process)
m	Number of human reviewers per manuscript
s	Reviewer skill parameter
τ	Reviewer time allocated per manuscript
H_{tot}	Total reviewer-hour budget per period
q^*	Journal's acceptance quality threshold
α	Significance level of DE's decision rule
σ_0^2	Baseline variance of human review errors
σ_H^2	Variance of human reviewer error under review time τ: $\frac{\sigma_0^2}{s\tau}$
σ^2_{post}	Posterior variance of DE's quality estimate
J_H	Posterior precision under human-only review: $\frac{ms\tau}{\sigma_0^2}$
	Additional notation for extended AI model:
$\sigma_A^2(\theta)$	Variance of AI reviewer error (decreasing in AI quality θ)
θ	AI quality parameter (higher θ = higher precision)
ρ_{HA}	Correlation between AI and human errors (anchoring)
ρ_{red}	Maximum value of ρ_{HA} at which AI provides net benefit
J_A	Precision of AI signal: $1/\sigma_A^2(\theta)$
J_{post}	Posterior precision with humans + AI
σ^2_{tar}	Target posterior variance set by the editor
$\tau(\theta, \rho_{\text{HA}})$	Reviewer time needed under AI quality θ and correlation ρ
$\Lambda(m, \theta, \rho_{\text{HA}})$	Journal throughput (capacity): $\frac{H_{\text{tot}}}{m\tau(\theta,\rho_{\text{HA}})}$
θ_{crit}	Minimum AI quality where 2 reviewers + AI = 3 reviewers

where s is reviewer skill and σ_0^2 is an exogenous parameter for review noise. The goal is to have a precise review by minimizing the error term ε_{H_i}. The higher the reviewer's skill and the more time they spend on the review, the more precise the review is. Signals are presumed independent across i in this base case, so the DE aggregates reviews by the sample mean

$$\bar{Q}_H \;=\; \frac{1}{m}\sum_{i=1}^{m} Q_{H_i}, \qquad \sigma^2_{\text{post}} = \mathrm{Var}(\bar{Q}_H \mid Q_i) \;=\; \frac{\sigma_0^2}{m\,s\,\tau} \tag{2}$$

where σ^2_{post} is the posterior variance that the DE observes after aggregating the reviews. This posterior variance shrinks (at diminishing rate) with more m reviewers. Likewise, it shrinks at

diminishing rate with reviewer skill and the time allocated to the review task. It is convenient to define the inverse, which is the DE's posterior precision

$$J_H \;\equiv\; \frac{1}{\sigma^2_{\text{post}}} \;=\; \frac{m\,s\,\tau}{\sigma_0^2}. \tag{3}$$

J_H is subscripted with H to denote that this is the base case with humans only. For analytical convenience, we use precision and variance interchangeably throughout the paper, with $J = 1/\sigma^2$. When appropriate, we refer to J as the "posterior precision" and σ^2 as its inverse, the "posterior variance."

Decision rule. The DE applies a hypothesis-testing rule. The journal has a quality threshold q^* for publication, which a paper's estimated quality must be statistically-significantly above at level α. This yields the decision rule

$$\text{Accept if} \quad \bar{Q}_H \;\geq\; q^* + z_{1-\alpha}\,\sigma_{\text{post}}.$$

Operationally for the journal, the DE wants their posterior variance on estimating a paper's quality to be below a minimal target posterior variance $\sigma^2_{post} \leq \sigma^2_{\text{tar}}$. To achieve this, they must choose a sufficiently high time allocation to reviewers, τ, such that:

$$\left(\sigma^2_{\text{post}} \;=\; \frac{\sigma_0^2}{m\,s\,\tau} \;\leq\; \sigma^2_{\text{tar}}\right) \;\equiv\; \left(\tau \;\geq\; \frac{\sigma_0^2}{m\,s\,\sigma^2_{\text{tar}}}\right). \tag{4}$$

Throughput under a review-time budget. Let H_{tot} be the total reviewer-hour budget per month. With m reviewers per paper and τ hours per reviewer assigned by the DE, monthly throughput (capacity) in the human-only regime is

$$\Lambda(m, 0) \;=\; \frac{H_{\text{tot}}}{m\,\tau}$$

where 0 denotes the lack of AI. Given a submission rate λ, the DE selects m and the maximum noise they are willing to tolerate in estimating paper quality σ^2_{tar}. This determines the necessary time per paper review, τ, via the constraint above. The DE thus trades off decision accuracy and turnaround time under the fixed budget H_{tot}. The following implication is evident due to the non-linearity embedded in Eq. 1.

Proposition 1 (Capacity Cliff)**.** *A journal's (average) response time, given a desired accuracy (quality) level J for editorial decisions, deteriorates at an increasing rate with submission volume. Let $T(\lambda; J) = \frac{1}{\Lambda(J)-\lambda}$ represent the expected journal response time, which is increasing and strictly convex in λ:*

$$\frac{\partial T}{\partial \lambda} > 0, \qquad \frac{\partial^2 T}{\partial \lambda^2} > 0,$$

and diverges as $\lambda \uparrow \Lambda(J)$

The formal proof of Proposition 1 is provided in Appendix A.

Fig. 1 from §1 illustrates the emerging crisis journals face when relying exclusively on human reviewers, particularly as manuscript submission rates increase due to authors adopting generative AI. In this figure, we depict the scenario of assigning three reviewers per manuscript, a common standard in current practice. At relatively low submission volumes, such as $\lambda = 20$, journals comfortably maintain high decision accuracy (e.g., target at 90%) with short turnaround times. As the submission rate grows however, the system reaches a sharply defined limit, termed the *capacity cliff*. At this point, reviewer workloads fully consume the journal's reviewing capacity, leading to rapidly escalating turnaround times and significant processing backlogs. Crucially, as the submission rate increases, the capacity cliff shifts leftward, constraining the journal's ability to sustain both rapid processing and high accuracy. For instance, at a higher submission rate ($\lambda = 40$), the cliff emerges, implying that maintaining current quality standards would significantly prolong manuscript processing times. Thus, journals face the difficult choice of increasing reviewer resources, reducing the number of reviewers per manuscript, or accepting lower decision accuracy (as shown in Fig. 1 with a worse 80% target accuracy instead).

3 Fighting Fire with Fire: AI-Infused Peer Review

In its recent review of operational pressures facing academic journals, *Cambridge University Press* underline the role of technology in improving productivity [of peer review]. Calling for "new solutions that address peer review at scale," the review argues that "responsible use of AI [can] streamline administrative and technical aspects of review, allowing human expertise to focus on what makes research interesting and important." (Cambridge University Press, 2025) We propose an AI-infused hybrid editorial workflow which strategically incorporates AI early in the review process, reconfiguring the role of human reviewers to focus their efforts on high-level, critical tasks. AI influence is explicit, logged, constrained, and auditable. There is no intent to automate editorial judgment, instead our approach relies on *principled augmentation*: leveraging AI to support human judgment, preserve scholarly values, and make the peer review process sustainable for the future.

3.1 Proposed AI-Augmented Journal Operations Workflow

The core of our proposal is a two-stage process in which the evaluative role of AI is tailored to the domain, methods, and value systems of the journal.

1. *Initial AI Review & Author Response:* Upon submission, every paper undergoes an immediate, journal-specific AI review. The AI tool generates structured critique and feedback for the authors, who then prepare a response document (or withdraw the paper) to address or rebut the points raised by the AI. This initial exchange serves to identify and correct basic issues, increases clarity, and ensures a baseline level of quality before the paper enters the human review process.
2. *Human Review with AI Context:* When the paper proceeds, human reviewers receive not only the original manuscript but also the AI-generated review and the authors' response. This augmented context allows them to bypass mechanical checks and focus their intellectual energy on the manuscript's originality, contribution to the literature, and prospective impact.

Our proposal specifically implies that the AI review is limited and targeted to a well-defined subset of review activities which have been vetted with respect to that journal. For instance, a particular journal might trust AI to examine writing quality, statement of research objectives and contribu-

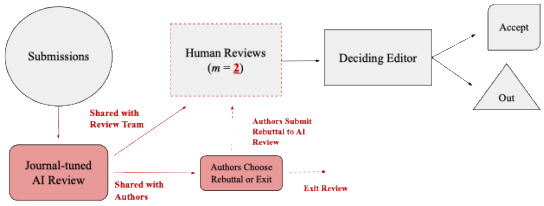
Diagram	**(Simplified) Current Process:** 	**AI-Assisted Review Proposal:** 
Author Submission	1. Author submits paper. *(optional: DE may desk reject)*	1. Author submits paper. 2. AI-assisted review is generated and sent to author. 3. Author chooses to either **exit** the review process, or **write a rebuttal** to AI-review, addressing concerns and correcting any AI mistakes. *(optional after Step 3: DE may desk reject)*
Review Process	2. Reviewers review the paper. 3. They write a review to the "Deciding Editor".	4. Reviewers review the paper **AND** the AI review alongside the authors' AI-rebuttal. 5. They write a review to the "Deciding Editor".
Deciding Editor	4. Deciding Editor assesses the reviews. 5. DE decides whether to accept or reject the submission.	6. Deciding Editor assesses the reviews (**including the AI review and AI-rebuttal**). 7. DE decides whether to accept or reject the submission.

Figure 3: Governed AI workflow (proposed) vs. status quo. The AI output is non-final and paired with an author right-of-response before human evaluation.

tions, existence of testable research claims, and a rigorous methodology that provides evidence of these claims; but not to hypothesis development or analysis of mathematical proofs or novelty with respect to existing literature. Crucially, this proposal shifts the use of AI from an informal, unobserved tool used idiosyncratically by individuals, to a *journal-provided* and *journal-scoped* assessment that every submission sees in the same way. The AI and human reviewers work together as explained below, and depicted in Fig. 3. The manuscript itself remains fixed at this stage; the response is an attachment, not a revision. Reviewers receive three objects, the manuscript, the AI assessment, and the author response, so that their effort can shift away from mechanical checks already performed and toward higher-order judgment.

3.2 A Model of AI-Infused Peer Review

We now extend our model of peer review to include an AI reviewer alongside humans, allowing us to analyze how AI affects decision accuracy, reviewer effort, and journal throughput. The model captures the interplay between these three key forces: (1) The use of AI increases the productivity of scholars, captured as increase in submission rate λ, (2) an AI reviewer of quality θ which can perform a subset of reviewer tasks, and (3) the adverse potential for *dependence* between the AI report and human reviewers, captured as correlation ρ_{HA} between human reviewers and the AI review (with ideal value $\rho_{\text{HA}} = 0$).

Signals of Submission Quality. Each manuscript continues to have latent quality $Q \sim F$. Human reviewers generate noisy signals of Q as in the benchmark model, while the AI produces its own evaluation (Q_A). Specifically,

$$Q_{H_i} = Q + \varepsilon_{H_i}, \quad \varepsilon_{H_i} \sim \mathcal{N}\big(0, \sigma_H^2\big), \quad \sigma_H^2 = \frac{\sigma_0^2}{s\,\tau},$$
$$Q_A = Q + \varepsilon_A, \quad \varepsilon_A \sim \mathcal{N}\big(0, \sigma_A^2(\theta)\big),$$

where $\sigma_A^2(\theta)$ decreases in AI quality (θ). This is important because as AI quality improves, it provides a more precise signal of manuscript quality. As this signal becomes more precise, this force allows the AI to alleviate some of the reviewer burden. However, to account for potential anchoring , where human reviewers may be influenced (or anchored) by the initial AI review , we allow correlation between human (H) and AI (A) errors:

$$\mathrm{Cov}(\varepsilon_A, \varepsilon_{H_i}) = \rho_{\text{HA}}\,\sigma_A \sigma_H, \quad \text{where } \sigma_H^2 = \tfrac{\sigma_0^2}{s\,\tau}.$$

In our base model, we make the conservative assumption that human reviews do not have correlation with each other. If we relax this assumption, as we discuss in the deteriorating base case

extension, this would reduce the performance of our base model.

Aggregation of Signals. The DE forms a generalized least squares (GLS) estimator based on the stacked signals $(Q_A, \bar{Q}_H)$, where $\bar{Q}_H$, as defined earlier, is the aggregated human reviews. The posterior variance with an AI reviewer is now

$$\mathrm{Var}(\hat{Q}) = \frac{ab - c^2}{a + b - 2c}, \qquad a = \sigma_A^2, \;\; b = \tfrac{\sigma_H^2}{m}, \;\; c = \rho_{\text{HA}}\, \sigma_A \sigma_H.$$

Reviewer effort. Similar to the human-only base case, the DE has a desired target posterior variance σ^2_{tar}. To achieve this, the DE also chooses reviewer time $\tau(\theta, \rho)$, where τ now depends on θ and ρ. Solving the GLS condition yields

$$\tau(\theta, \rho_{\text{HA}}) = \frac{\sigma_0^2}{s\, h^2},$$

where h is determined by the quadratic

$$\sigma^4_{\text{tar}}(m - m\rho^2)h^4 + (2\rho\sigma_A\sigma^2_{\text{tar}} - m\sigma_A^2)h^2 + m\sigma_A^2\sigma^2_{\text{tar}} = 0$$

where, as a reminder, σ_A is a function of AI quality θ. This expression shows that the necessary reviewer time to meet the target decreases with AI quality ($\partial\tau/\partial\theta < 0$), but increases with anchoring correlation ($\partial\tau/\partial\rho_{\text{HA}} > 0$).

System Throughput. With an AI-assisted review process, journal capacity under total reviewer-hours H_{tot} is now

$$\Lambda(m, \theta, \rho_{\text{HA}}) = \frac{H_{\text{tot}}}{m\, \tau(\theta, \rho_{\text{HA}})}.$$

The throughput illustrates the intuition in the forces of the model. Journal capacity rises when AI quality increases (allowing AI to reduce human effort). Higher values of AI quality (θ) allow

the editor to reduce the time they require for human reviewers (τ). This allows the same budget to process more manuscripts. However, as we show below, editors should be wary of potential anchoring correlation of reviewer opinion in ρ_{HA}.

Correlation Threshold for using AI. A key insight is that anchoring affects incremental information: as the AI and human assessments become more dependent, the AI signal becomes increasingly redundant with the human panel. We assume that the DE uses variance-minimizing GLS aggregation using the known, true covariances. Thus, including Q_A cannot worsen posterior precision; at worst, GLS assigns it zero weight. Accordingly, there exits an operational threshold

$$\rho_{\text{red}} \equiv \frac{\sigma_H}{m\,\sigma_A}. \tag{5}$$

Therefore, adding AI yields a *strict* precision improvement if and only if

$$\rho_{\text{HA}} < \rho_{\text{red}}. \tag{6}$$

At $\rho_{\text{HA}} = \rho_{\text{red}}$, the optimal GLS weight on the AI becomes exactly zero and the system reverts to the human-only benchmark.

Importantly, this threshold provides a testable adoption rule: measuring ρ_{HA}, σ_A, and σ_H allows editors to determine whether AI meaningfully improves decision accuracy. We discuss how a journal may measure these in the empirical operationalization subsection.

Figure 4 illustrates the effect of anchoring correlation in the model. In the informative region $\rho_{\text{HA}} < \rho_{\text{red}}$, Governed AI (via the DE's GLS aggregation) strictly improves precision; at the redundancy point $\rho_{\text{HA}} = \rho_{\text{red}}$ the AI receives zero weight and the system returns to the human-only baseline. This visualization highlights a central insight: *AI is capacity-enhancing, reducing the number of reviewers needed per paper, when it does not dramatically anchor reviewers' opinions.*

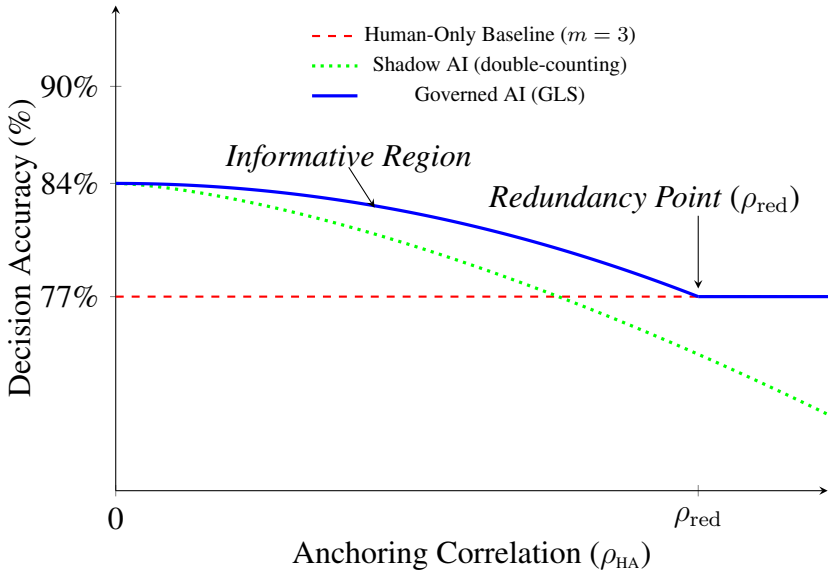


Figure 4: Anchoring Correlation. The horizontal dashed line is the human-only baseline (illustrated for $m = 3$ reviewers). As anchoring correlation ρ_{HA} rises, the AI signal becomes less incremental. In the *informative region* $\rho_{\text{HA}} < \rho_{\text{red}}$, Governed AI (solid blue, GLS aggregation) strictly improves precision. At the *redundancy point* $\rho_{\text{red}} = \sigma_H/(m\sigma_A)$, the AI contributes zero marginal information and the optimal DE decision under GLS aggregation sets its weight to zero, returning to baseline. Under *misspecified* aggregation (dotted-green), dependence can degrade accuracy via double-counting correlated signals if human reviews also become correlated with each other due to the presence of shadow AI reviews.

3.3 Main Analytical Results

In this section, we summarize the model's core theoretical implications for reviewer effort, informational value, and operational capacity. All proofs appear in Appendix A.

Theorem 1 (Effort-Saving)**.** *Human reviewer time per manuscript, $\tau(\theta, \rho_{HA})$, decreases monotonically with AI quality θ in the informative (non-redundant) region:*

$$\frac{\partial \tau(\theta, \rho_{HA})}{\partial \theta} < 0 \quad \textit{for} \quad \rho_{HA} < \rho_{\text{red}}.$$

Formally, this holds under the regularity conditions stated in Appendix A, requiring that AI and human signals lie in a balanced-precision *region where the posterior variance is increasing in both noise sources.*

Higher-quality AI reduces the human effort needed to achieve any fixed accuracy target. Critically, this effect depends on informational independence: as anchoring rises toward ρ_{red}, the AI signal

becomes increasingly redundant and the gains shrink (Figure 4). Slack in reviewer-hours can be used to reduce turnaround time or shift effort to tasks outside the AI rubric.

Theorem 2 (Reviewer Substitution Threshold)**.** *There exists a critical AI quality level θ_{crit} such that two human reviewers assisted by AI ($m = 2$) achieve the same posterior precision as three human reviewers without AI ($m = 3$):*

$$\frac{3\,\tau(0)}{2\,\tau(\theta_{crit})} \geq 1.$$

This identifies the point at which one human reviewer can be removed without sacrificing accuracy. It provides an operational benchmark for when "AI substitution" is feasible.

Theorem 3 (Anchoring Redundancy)**.** *Assume the Deciding Editor combines the AI and human signals using the variance-minimizing GLS rule with the true covariance. Then including the AI signal cannot worsen posterior precision:*

$$\mathrm{Var}(\hat{Q}_{AI+H}) \;\leq\; \mathrm{Var}(\hat{Q}_{H\text{-}only}).$$

Moreover, the inequality is strict whenever the AI signal is marginally informative, equivalently $\rho_{HA} < \rho_{\mathrm{red}}$, and becomes an equality at the redundancy point $\rho_{HA} = \rho_{\mathrm{red}}$ (where GLS assigns the AI signal zero weight).

As anchoring rises, the AI signal becomes less incremental. At $\rho_{\mathrm{HA}} = \rho_{\mathrm{red}}$, under GLS aggregation, the AI provides no marginal information beyond the human panel and the joint information collapses to the human-only benchmark level (Figure 4). This emphasizes that editorial guidance and reviewer training can matter as much as AI quality.

Corollary 1 (No-Collapse under Correct Aggregation)**.** *Assume the joint covariance of $(Q_A, Q_{H_1}, \ldots, Q_{H_m})$ is positive definite and the Deciding Editor aggregates $(Q_A, \bar{Q}_H)$ using the variance-minimizing GLS rule with the true covariance. Then adding the AI signal cannot reduce posterior precision:*

$$\mathrm{Var}(\hat{Q}_{AI+H}) \;\leq\; \mathrm{Var}(\hat{Q}_{H\text{-}only}).$$

Moreover, the improvement is strict whenever the AI signal is marginally informative ($\rho_{HA} < \rho_{\mathrm{red}}$).

When the editor knows the true covariances and aggregates signals optimally, adding AI cannot make things worse. GLS aggregation automatically down-weights the AI signal as its correlation

with human reviewers increases, assigning it zero weight exactly at the redundancy threshold. This contrasts with Theorem 3, which characterizes a local phenomenon—how the AI's marginal contribution shrinks as ρ_{HA} approaches redundancy. The corollary is a global guarantee: across the entire admissible parameter space, properly aggregated AI never degrades precision below the human-only benchmark. Consequently, degradation below baseline requires process mis-specification such as treating correlated signals as independent (e.g., shadow AI use, naive averaging, or other forms of double-counting). Figure 4 illustrates this distinction.

Theorem 4 (Frontier Shift)**.** *For a fixed total reviewer-hour budget H_{tot}, increasing AI quality θ in the informative region ($\rho_{HA} < \rho_{\text{red}}$) shifts the journal's delay–accuracy frontier outward:*

$$\frac{\partial \Lambda(m, \theta, \rho_{HA})}{\partial \theta} > 0, \qquad \frac{\partial \sigma^2_{\text{post}}}{\partial \theta} < 0.$$

Improved AI quality simultaneously expands capacity and reduces posterior uncertainty. This creates a Pareto improvement in the review pipeline—faster turnaround and higher accuracy—so long as anchoring remains below the redundancy point.

3.4 Reality of Shadow AI

Our analysis has assumed the peer review *status quo* comprises human reviews with independent signals Q_{H_i}. Emerging evidence of *shadow AI* in the review process (Gans, 2025a) suggests that reality is worse than the "No AI" analysis presented in 2. With *ad hoc* use of AI, human reviewers become increasingly correlated with each other as they mutually (and secretly) rely on similar AI-assisted review tools. In our model, this scenario corresponds to positive correlation between the human review signals, e.g., $\rho_{\text{HH}} = \text{Cov}(Q_{H_i}, Q_{H_j})$ where $i \neq j$. This implies that the DE forms a worse aggregated signal than the benchmark case of $\rho_{\text{HH}} = 0$. With $\rho_{\text{HH}} > 0$, the DE's posterior variance after aggregating the m human reviews becomes $\text{Var}(\bar{Q}_H \mid Q_i) = \frac{\sigma_H^2}{m}(1 + (m-1)\rho_{\text{HH}})$. Compared with shadow AI in the review process, our proposed framework has an additional benefit of making the AI-assistance *transparent*. By the journal providing its own AI-assisted review, this

makes the AI review transparent not only for the authors, but the reviewers as well. The DE is able to study and observe ρ_{HA} whereas ρ_{HH} is unobserved in the shadow AI scenario where human reviewers act individually and with their own tools and prompts, rather than on an agreed upon AI review that the journal provides.

4 Evaluation of our Proposal

This article emerged from the imperative to address operational stress faced by journals from increased submission volume, combined with the potential harms from hidden uncontrolled use of AI in peer review. Our specific proposal (laid out in §3) for how one might use AI tools to counter this AI-induced challenge is a call to action rather than necessarily perfect in every aspect of it. Below we discuss its merits and limitations, emphasizing the need for careful experimentation (detailed in Appendix B).

4.1 Benefits Across Stakeholders

The proposed workflow's front-end orientation delivers benefits across stakeholders by framing AI output as transparent, non-final observations that authors can address.

- **For Reviewers:** AI handles lower-level methodological and presentation concerns early, freeing reviewers for literature context and high-level evaluation. Authors' responses to AI feedback clarify obscure elements, saving time.
- **For Editors:** AI screening for methodological soundness enables better-informed desk rejections and improved reviewer allocation. Structured AI feedback provides a consistent baseline for identifying overlooked issues or divergent opinions.
- **For Authors:** Immediate AI reviews dramatically reduce wait times. Required response documents create a correction loop, allowing authors to clarify or rebut concerns before human review—ensuring AI serves as triage, not replacement.
- **For Journal:** The workflow addresses a critical capacity gap: insufficient qualified reviewers relative to submission volumes, widening as AI tools lower writing costs. Crucially,

reviewers already use AI privately and inconsistently, creating unpredictable feedback quality. Journal-approved AI introduces transparency and standardization, aligning expectations across stakeholders.

The workflow satisfies accountability requirements (human reviewers make final decisions, consistent with EU AI Act mandates), maintains confidentiality through secure infrastructure with reviewer-level access controls, and enables measurement. Unlike shadow AI use, journal-provided assessment generates standardized records linkable to review outcomes, making key parameters estimable: AI signal noise (σ_A^2), human signal noise (σ_H^2), and induced correlation (ρ_{HA}). Journals can even hold out reviewers from seeing AI output to assess anchoring effects.

One potential concern for an adverse effect from using an AI reviewer (and sharing the AI review with human reviewers) is correlation between human reviewers and AI output—specifically, reviewer complacency. However, this risk exists to some extent even when reviewers independently use AI tools. Journal-approved AI makes this transparent and observable. To mitigate anchoring, AI output should be: (i) constrained to auditable rubric dimensions, (ii) prohibited from accept/reject recommendations, and (iii) presented in structured, tunable formats (e.g., flags-first with expandable evidence). AI functions as a structured auditor, surfacing clarity failures, missing reporting elements, and claim-evidence mismatches. Humans retain responsibility for contribution, novelty, impact, and publication decisions. Additional concerns are summarized in Table 3.

4.2 Why This Design Enables Measurement

The workflow's advantage over shadow AI is observability. Journal-provided AI assessment enables structured, standardized output that can be logged and linked to downstream reviews. Three key quantities become estimable from routine editorial data: **AI signal noise** (σ_A^2), tracked through dispersion of AI diagnostics relative to outcomes and independent human assessments; **human signal noise** (σ_H^2), measured via reviewer disagreement patterns conditional on manuscript features; and **induced dependence** (ρ_{HA}), assessed by comparing assessments with/without AI context via

Table 3: Strategic responses and implementation concerns.

Concern	Risk	Mitigation / monitoring
Reviewer complacency	Over-reliance on AI comments; neglect of dimensions presumed "covered" by the AI rubric.	Evaluate via randomized holdouts; constrain AI to auditable rubrics; give explicit reviewer guidance on non-rubric dimensions.
Reviewer over-correction	Shifting standards upward in response to AI-flagged issues, potentially increasing false negatives.	Measure decision shifts under AI exposure; calibrate rubric thresholds; monitor rejection reasons and downstream citation/replication signals.
Author gaming	Pre-optimizing to anticipated AI feedback; strategic compliance without substantive improvement.	Limit response length; vary prompts/rubrics over time; monitor abnormal divergence in AI scores and reviewer language.
Excessive author effort	Increased response burden without commensurate quality gains.	Track revision cycles and author time-to-resubmit; audit which AI-identified issues predict editorial outcomes.
Limited improvement value	AI feedback does not materially improve manuscripts (weak treatment effect).	Treat AI as triage/screening even if developmental value is limited; reallocate human effort to higher-variance judgments.
Prompt injection / manipulation	Attempts to exploit the AI system (instructions embedded in manuscripts, adversarial formatting).	Journal-side sanitization (strip instructions/metadata), retrieval isolation, and red-teaming; log anomalies and auto-flag suspicious inputs.

controlled pilots or random reviewer holdouts. These parameters determine whether AI improves or degrades decision quality. The workflow transforms hidden system risk—unmeasured correlation from shadow AI—into an observable, manageable process parameter. Appendix B describes a pilot experiment to measure effects of introducing a front-line AI review.

5 Conclusion

The accelerating use of AI in academic research presents a structural challenge to traditional peer review. As AI-assisted submissions increase, reviewer-dependent workflows risk not only delays but a fundamental erosion of independent judgment due to the rise of shadow AI. This paper offers both a theoretical model and a practical workflow to integrate AI into peer review, not as a substitute for human judgment but as a structured and transparent enhancement. Rather than

responding defensively or resisting AI, the academic community has the opportunity to lead in shaping systems that uphold scholarly values in this new era. We offer one such design and invite continued dialogue as implementation and empirical evaluation proceed.

We propose employing AI as a first-line reviewer to provide early feedback and request author revisions prior to human evaluation. This process raises the bar for submission readiness and reduces the burden on reviewers. Crucially, this workflow transforms the use of AI from a hidden liability into a transparent asset. Under the status quo, unmonitored AI use by reviewers introduces hidden correlations that degrade the epistemic independence of the review process. By formalizing AI as a journal-tuned agent, our framework converts these unobserved biases into managed operational levers. This process also makes the use of AI transparent, whereas under the status quo, it is masked and out of the journal's influence if reviewers are using their own AI reviews and prompts. Our model demonstrates that even moderately capable AI systems can improve throughput and decision accuracy, particularly in a configuration involving two human reviewers alongside AI support.

To summarize, ours is not a call for AI-based automated reviews, but for principled augmentation. Our framework preserves reviewer agency, supports editorial judgment, and prepares journals for the realities of AI-augmented scholarship. We have examined one possible workflow to take advantage of advances in AI capability, not just in research and research volume, but also in the editorial process. We recognize that this is only one of many possible ways to adapt the process for increasing submission rates. Further, the existence of a journal-specific AI reviewer raises a question regarding public (and pre-submission) accessibility to authors: doing so could enhance reproducibility and allow authors to improve upon on their manuscripts, but authors might exploit this openness to optimize submissions specifically for the AI reviewer prior to the formal review process. Moreover, this proposal aims to establish a foundation for the management sciences scholarly community to build upon and improve.

References

Berente, N. and Recker, J. (2025). Let's all cheer for the journal of the association for information systems. This IS Research Podcast. Podcast episode featuring Monideepa Tarafdar, Editor-in-Chief of JAIS.

Bhargava, H. K., Brown, S., Ghose, A., Gupta, A., Leidner, D., and Wu, D. J. (2025). Exploring generative ai's impact on research: Perspectives from senior scholars in management information systems. *ACM Transactions on Management Information Systems*, 16(2):1–9.

Bhargava and Tokarskaya (2025). Ai and the scholarly review across different journals. Internal working paper. Unpublished manuscript.

Cambridge University Press (2025). Publishing futures: Working together to deliver radical change in academic publishing.

Gans, J. (2025a). What to do about AI referee reports? https://joshuagans.substack.com/p/what-to-do-about-ai-referee-reports.

Gans, J. (2025b). What will ai do to (p)research? https://joshuagans.substack.com/p/what-will-ai-do-to-presearch.

Gopal, R. D., Li, J., Riemer, K., Sarker, S., Singh, P. V., Susarla, A., Bichler, M., and Thatcher, J. B. (2025). Inventing with machines: Generative AI and the evolving landscape of IS research. *Information Systems Research*, 0(0). Forthcoming.

Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., Palepu, A., Sirkovic, P., Myaskovsky, A., Weissenberger, F., Rong, K., Tanno, R., Saab, K., Popovici, D., Blum, J., Zhang, F., Chou, K., Hassidim, A., Gokturk, B., Vahdat, A., Kohli, P., and Natarajan, V. (2025). Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*.

Hosseini, M. and Resnik, D. B. (2025). Guidance needed for using artificial intelligence to screen journal submissions for misconduct. *Research Ethics*, 21(1):1–8.

Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., and West, R. (2024). The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint*.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., et al. (2024a). Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *International Conference on Machine Learning (ICML)*, pages 29575–29620.

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., and Zou, J. (2024b). Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196.

Ludwig, J. and Mullainathan, S. (2024). Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827.

Manning, B. S., Zhu, K., and Horton, J. J. (2024). Automated social science: Language models as scientist and subjects. Technical report, National Bureau of Economic Research.

National Institutes of Health (2025). Supporting fairness and originality in NIH research applications. NIH Notice NOT-OD-25-132. Effective for applications submitted to the September 25, 2025, receipt date and beyond.

Peters, U. and Chin-Yee, B. (2025). Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776.

Richardson, R. A. K., Hong, S. S., Byrne, J. A., Stoeger, T., and Amaral, L. A. N. (2025). The entities enabling scientific fraud at scale are large, resilient, and growing rapidly. *Proceedings of the National Academy of Sciences*, 122(32):e2420092122.

Tao, T. (2024). Machine assisted proof. Notices of the American Mathematical Society.

A Extended Proofs and Derivations

Standing assumptions (signals, anchoring, and induced moments). Unless stated otherwise, we assume the conservative base case in which human reviewers are conditionally independent given Q (no shadow-AI correlation), i.e., $\rho_{\text{HH}} = 0$. Anchoring is symmetric and is indexed by the *pairwise* correlation

$$\rho_{\text{HA}} \;:=\; \mathrm{Corr}(\varepsilon_A, \varepsilon_{H_i}), \qquad i = 1, \ldots, m.$$

Let $\bar{\varepsilon}_H := \frac{1}{m}\sum_{i=1}^{m} \varepsilon_{H_i}$ denote the mean human error. Under conditional independence of $\{\varepsilon_{H_i}\}$,

$$\mathrm{Var}(\bar{\varepsilon}_H) = \frac{\sigma_H^2}{m}.$$

Moreover, symmetric anchoring implies

$$\mathrm{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{Cov}(\varepsilon_A, \varepsilon_{H_i}) = \rho_{\text{HA}}\, \sigma_A \sigma_H.$$

Equivalently, the induced AI–panel-mean correlation is

$$\rho_{A\bar{H}} \;:=\; \mathrm{Corr}(\varepsilon_A, \bar{\varepsilon}_H) = \frac{\rho_{\text{HA}}\, \sigma_A \sigma_H}{\sigma_A(\sigma_H/\sqrt{m})} = \sqrt{m}\, \rho_{\text{HA}}.$$

Finally, feasibility (positive definiteness of the 2×2 covariance matrix of $(Q_A, \bar{Q}_H)$) implies $|\rho_{A\bar{H}}| < 1$, hence

$$|\rho_{\text{HA}}| < \frac{1}{\sqrt{m}}.$$

Proposition 1 (Capacity Cliff)**.** *Fix an evidentiary (accuracy) target J for editorial decisions, and let $\tau(J)$ denote the human reviewer time per reviewer required to meet J. Given m reviewers per manuscript and a total reviewer-hour budget per unit time H_{tot}, define the induced service capacity (service rate)*

$$\mu(J) \;\equiv\; \Lambda(J) \;\equiv\; \frac{H_{\text{tot}}}{m\,\tau(J)}.$$

Assume submissions arrive at rate $\lambda < \mu(J)$. Under an M/M/1 approximation with arrival rate λ and service rate $\mu(J)$, the expected response time (time in system) is

$$T(\lambda; J) \;=\; \frac{1}{\mu(J) - \lambda},$$

which is increasing and strictly convex in λ:

$$\frac{\partial T}{\partial \lambda} = \frac{1}{(\mu(J) - \lambda)^2} > 0, \qquad \frac{\partial^2 T}{\partial \lambda^2} = \frac{2}{(\mu(J) - \lambda)^3} > 0,$$

and diverges as $\lambda \uparrow \mu(J)$.

Proof. Fixing J fixes the required per-reviewer effort $\tau(J)$ and hence the service rate $\mu(J) = H_{\text{tot}}/(m\tau(J))$. For an M/M/1 queue with arrival rate λ and service rate $\mu(J)$, stability requires $\lambda < \mu(J)$ and the expected time in system is $T(\lambda; J) = 1/(\mu(J) - \lambda)$. Differentiating yields

$$\frac{\partial T}{\partial \lambda} = \frac{1}{(\mu(J) - \lambda)^2} > 0, \qquad \frac{\partial^2 T}{\partial \lambda^2} = \frac{2}{(\mu(J) - \lambda)^3} > 0,$$

establishing monotonic deterioration, strict convexity, and divergence as $\lambda \uparrow \mu(J)$. □

Restatement of Theorem 1 (formal version)

Theorem 1 (Effort-Saving Law (formal restatement))**.** *Fix a target posterior variance $\sigma^2_{\text{tar}} > 0$ and a number of human reviewers $m \geq 1$. Let $a(\theta) \equiv \sigma^2_A(\theta)$ be strictly decreasing in θ. Let $h \equiv \sigma_H > 0$ denote the human error standard deviation (endogenous through reviewer time $\tau = \sigma_0^2/(sh^2)$).*

For each AI quality θ and anchoring level $\rho_{HA} \in [0, 1)$, define

$$\begin{aligned} V(a, h; \rho_{HA}, m) &= \frac{ab - c^2}{a + b - 2c}, \\ b &= \frac{h^2}{m}, \\ c &= \text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \rho_{HA}\sqrt{a}\, h, \qquad \bar{\varepsilon}_H := \frac{1}{m}\sum_{i=1}^{m} \varepsilon_{H_i}. \end{aligned}$$

Assume the operating point lies in the regular (informative) region

$$\mathcal{R} := \left\{ (a, h, \rho_{HA}) : \begin{array}{l} 0 \leq \rho_{HA} < 1/\sqrt{m}, \\ \rho_{HA} = 0 \text{ or } \rho_{HA} < \rho_{\text{red}}(a, h) \end{array} \right\}, \qquad \rho_{\text{red}}(a, h) := \frac{h}{m\sqrt{a}}.$$

and that the editor attains the target variance:

$$V\big(a(\theta), h(\theta); \rho_{HA}, m\big) = \sigma^2_{\text{tar}}.$$

Then, along this iso-variance curve,

$$\frac{\partial \tau(\theta, \rho_{HA})}{\partial \theta} < 0.$$

Remark (Regular region and its meaning). The proof uses the Implicit Function Theorem and requires that $V(a, h; \rho_{\text{HA}}, m)$ is locally increasing in both a (AI noise) and h (human noise) at the operating point.

(i) Feasibility / no singularity. For $\rho_{\text{HA}} < 1/\sqrt{m}$ the GLS denominator is strictly positive:

$$a + \frac{h^2}{m} - 2\rho_{\text{HA}} h\sqrt{a} = (\sqrt{a} - \rho_{\text{HA}} h)^2 + \frac{h^2}{m}\big(1 - \rho_{\text{HA}}^2 m\big) > 0.$$

(ii) Correct monotonicity signs. On $\mathcal{R}$ we have $b - c = \frac{h^2}{m} - \rho_{\text{HA}} h\sqrt{a} > 0$, which implies $V_a > 0$ and $V_h > 0$ at the operating point.[2] Intuitively, $\mathcal{R}$ excludes the redundancy boundary where the AI becomes a re-expression of the human mean.

Connection to the adoption rule. The condition $\rho_{\text{HA}} < \rho_{\text{red}}(a, h)$ is equivalent to $\rho_{\text{HA}} < \sigma_H/(m\sigma_A)$, i.e., the value-added regime in which the AI signal is not redundant under GLS.

Sign logic (why effort falls). Along the iso-variance curve $V(a(\theta), h(\theta); \rho_{\text{HA}}, m) = \sigma_{\text{tar}}^2$, implicit differentiation gives $V_a a_\theta + V_h h_\theta = 0$. Since $a_\theta < 0$ and $V_a, V_h > 0$ on $\mathcal{R}$, we have $h_\theta > 0$. Because $\tau = \sigma_0^2/(sh^2)$, it follows that $\tau_\theta < 0$.

Proof of Theorem 1

Proof. Fix $\sigma_{\text{tar}}^2 > 0$, $m \geq 1$, and $\rho_{\text{HA}} \in [0, 1)$. Write $a = \sigma_A^2(\theta) > 0$ (strictly decreasing in θ) and $h = \sigma_H > 0$ (endogenous through reviewer time). Under the standing assumptions, the GLS posterior variance based on $(Q_A, \bar{Q}_H)$ is

$$V(a, h; \rho_{\text{HA}}, m) = \frac{ab - c^2}{a + b - 2c}, \qquad b = \frac{h^2}{m}, \qquad c = \rho_{\text{HA}}\sqrt{a}\, h.$$

Assume the operating point lies in the regular region $\mathcal{R}$ (as defined in Remark A), so that feasibility holds ($\rho_{\text{HA}} < 1/\sqrt{m}$) and the signal is informative ($b - c > 0$) and away from near-collinearity (so the derivatives below have the correct signs).

The editor chooses h to meet the target $V(a, h; \rho_{\text{HA}}, m) = \sigma_{\text{tar}}^2$. Let $F(a, h) = V(a, h; \rho_{\text{HA}}, m) -$

[2] A direct calculation shows $V_a = (b - c)^2/(a + b - 2c)^2 > 0$ whenever $a + b - 2c > 0$, and similarly $V_h > 0$ on the same informative region.

σ^2_{tar}. By the Implicit Function Theorem on $F(a,h)=0$,

$$\frac{\partial h}{\partial \theta} = \frac{\partial h}{\partial a}\frac{\partial a}{\partial \theta} = -\frac{V_a}{V_h}\, a_\theta, \qquad a_\theta < 0.$$

Thus it suffices to show that $V_a > 0$ and $V_h > 0$ at the operating point.

A direct differentiation yields (writing $k := \frac{1}{m} - \rho^2_{\text{HA}}$)

$$V_a = \frac{kh^2\big(\frac{h^2}{m} - \rho_{\text{HA}} h\sqrt{a}\big)}{(a + \frac{h^2}{m} - 2\rho_{\text{HA}} h\sqrt{a})^2}, \qquad V_h = \frac{2kah\big(a - \rho_{\text{HA}} h\sqrt{a}\big)}{(a + \frac{h^2}{m} - 2\rho_{\text{HA}} h\sqrt{a})^2}.$$

On $\mathcal{R}$ we have $k > 0$ (since $\rho_{\text{HA}} < 1/\sqrt{m}$), the denominator is strictly positive (Remark A), and the bracketed factors are positive by definition of $\mathcal{R}$. Hence $V_a > 0$ and $V_h > 0$.

Therefore $\partial h/\partial a = -V_a/V_h < 0$. Since $a_\theta < 0$, it follows that

$$\frac{\partial h}{\partial \theta} > 0.$$

Finally, since $\tau = \sigma_0^2/(sh^2)$,

$$\frac{\partial \tau}{\partial \theta} = -\frac{2\sigma_0^2}{sh^3}\,\frac{\partial h}{\partial \theta} < 0,$$

which is the effort-saving law. □

Proof of Theorem 2

Proof. Fix a target variance $\sigma^2_{\text{tar}} > 0$. Assume the human-only benchmark with $m = 3$ reviewers and per-reviewer time τ_0 attains this target:

$$\sigma^2_{\text{tar}} = \frac{\sigma_0^2}{3s\tau_0}.$$

Let $h_0^2 := \sigma_H^2(\tau_0) = \sigma_0^2/(s\tau_0)$, so that

$$\frac{h_0^2}{3} = \sigma^2_{\text{tar}}.$$

Now consider the AI+$m = 2$ regime while holding per-reviewer effort fixed at τ_0 (i.e., $h = h_0$). Write $a = \sigma_A^2(\theta)$ and let ρ_{HA} denote the pairwise anchoring. Under the standing assumptions (conditionally independent reviewers and symmetric anchoring),

$$b = \text{Var}(\bar{\varepsilon}_H) = \frac{h_0^2}{2}, \qquad c = \text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \rho_{\text{HA}}\sqrt{a}\, h_0, \qquad \bar{\varepsilon}_H := \frac{1}{2}\sum_{i=1}^{2} \varepsilon_{H_i}.$$

The GLS posterior variance based on $(Q_A, \bar{Q}_H)$ is

$$V_2(a) = \frac{ab - c^2}{a + b - 2c}.$$

Step 1: A bracket for the target. As $a \to 0^+$, the AI becomes arbitrarily precise and $V_2(a) \to 0$. If $\rho_{\text{HA}} > 0$, define the redundancy boundary

$$a_{\text{red}} := \Big(\frac{h_0}{2\rho_{\text{HA}}}\Big)^2,$$

so that $\sqrt{a_{\text{red}}} = h_0/(2\rho_{\text{HA}})$ and hence $c = b$. Substituting $c = b$ yields $V_2(a_{\text{red}}) = b = h_0^2/2$. Therefore,

$$\lim_{a \to 0^+} V_2(a) = 0 < \sigma_{\text{tar}}^2 = \frac{h_0^2}{3} < \frac{h_0^2}{2} = V_2(a_{\text{red}}).$$

Since $V_2(\cdot)$ is continuous on $(0, a_{\text{red}}]$, the Intermediate Value Theorem implies there exists at least one $a^\star \in (0, a_{\text{red}})$ such that

$$V_2(a^\star) = \sigma_{\text{tar}}^2.$$

If $\rho_{\text{HA}} = 0$, then $c \equiv 0$ and $V_2(a) = \frac{ab}{a+b}$ is continuous and strictly increasing in a with $\lim_{a\to 0^+} V_2(a) = 0$ and $\lim_{a\to\infty} V_2(a) = b$, so the same existence conclusion holds on $(0, \infty)$.

Step 2: Uniqueness via monotonicity on the informative region. Assume feasibility $\rho_{\text{HA}} < 1/\sqrt{2}$ (positive definiteness under $m = 2$). For $a \in (0, a_{\text{red}})$ we have $b - c > 0$ and $a + b - 2c > 0$. A direct differentiation accounting for $c = \rho_{\text{HA}} h_0 \sqrt{a}$ yields

$$\frac{d}{da} V_2(a) = \frac{(b - c)^2}{(a + b - 2c)^2} > 0 \qquad \text{for all } a \in (0, a_{\text{red}}),$$

so $V_2(a)$ is strictly increasing on $(0, a_{\text{red}})$ and the solution $a^\star$ is unique. In particular,

$$a \le a^\star \implies V_2(a) \le V_2(a^\star) = \sigma_{\text{tar}}^2.$$

Step 3: Define θ_{crit} and conclude substitution. Since $a(\theta) = \sigma_A^2(\theta)$ is strictly decreasing in θ, define θ_{crit} by $a(\theta_{\text{crit}}) = a^\star$. Then for any $\theta \ge \theta_{\text{crit}}$ we have $a(\theta) \le a^\star$ and thus

$$V_2(a(\theta)) \le \sigma_{\text{tar}}^2 \quad \text{while holding human effort at } \tau_0.$$

Consequently, the minimal per-reviewer time required to meet the target under AI+$m = 2$ satisfies $\tau(\theta, \rho_{\text{HA}}) \le \tau_0$ for $\theta \ge \theta_{\text{crit}}$. Therefore the total human time under AI+$m = 2$ at θ_{crit} is no greater

than the baseline total human time under $m = 3$ without AI:

$$2\tau(\theta_{\text{crit}}, \rho_{\text{HA}}) \leq 2\tau_0 < 3\tau_0 = 3\tau(0),$$

which proves the substitution threshold claim. □

Proof of Theorem 3

Proof. Fix $m \geq 1$, $h = \sigma_H > 0$, $a = \sigma_A^2 > 0$, and write

$$b = \frac{h^2}{m}, \qquad c = \text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \rho_{\text{HA}} \sqrt{a}\, h,$$
$$V(a, h; \rho_{\text{HA}}, m) = \frac{ab - c^2}{a + b - 2c}, \qquad V_H = b.$$

Then

$$\begin{aligned} V - V_H &= \frac{ab - c^2}{a + b - 2c} - b \\ &= \frac{ab - c^2 - b(a + b - 2c)}{a + b - 2c} \\ &= -\frac{(c - b)^2}{a + b - 2c} \;\leq\; 0. \end{aligned} \tag{7}$$

To justify the sign, note that (since $c = \text{Cov}(\varepsilon_A, \bar{\varepsilon}_H)$)

$$a + b - 2c = \text{Var}(\varepsilon_A) + \text{Var}(\bar{\varepsilon}_H) - 2\,\text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \text{Var}(\varepsilon_A - \bar{\varepsilon}_H) \;\geq\; 0.$$

Moreover, feasibility (positive definiteness of the 2×2 covariance matrix of $(Q_A, \bar{Q}_H)$) implies $ab - c^2 > 0$ and hence rules out $a + b - 2c = 0$ (since $a + b - 2c = 0$ would force $c = (a + b)/2$ and thus $ab - c^2 = -\frac{1}{4}(a - b)^2 \leq 0$). Therefore $a + b - 2c > 0$, and (7) implies $V \leq V_H$.

Equality in (7) holds if and only if $c = b$, i.e.,

$$\rho_{\text{HA}} = \rho_{\text{red}} := \frac{b}{\sqrt{a}\, h} = \frac{h}{m\sqrt{a}} = \frac{\sigma_H}{m\sigma_A},$$

which is precisely the redundancy point where GLS places zero weight on the AI signal.

Finally, the *formal* denominator singularity for the two-signal GLS expression occurs at

$$\rho_{\text{crit}} := \frac{a + b}{2\sqrt{a}\, h} = \frac{\sigma_A^2 + \sigma_H^2/m}{2\sigma_A \sigma_H}.$$

Under symmetric anchoring with $\rho_{\text{HA}} = \text{Corr}(\varepsilon_A, \varepsilon_{H_i})$, positive definiteness implies $|\rho_{\text{HA}}| \leq 1/\sqrt{m}$, while $\rho_{\text{crit}} \geq 1/\sqrt{m}$ by AM–GM. Hence ρ_{crit} lies on or beyond the feasibility boundary and cannot be reached under admissible covariances (Corollary 1). □

Proof of Corollary 1

Proof. Under the standing assumptions (conditionally independent reviewers and symmetric anchoring),

$$\bar{Q}_H = Q + \bar{\varepsilon}_H, \qquad \text{Var}(\bar{\varepsilon}_H) = \frac{\sigma_H^2}{m}, \qquad Q_A = Q + \varepsilon_A, \qquad \text{Var}(\varepsilon_A) = \sigma_A^2.$$

Let $\rho_{\text{HA}} = \text{Corr}(\varepsilon_A, \varepsilon_{H_i})$ for each i. Then

$$\text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \frac{1}{m} \sum_{i=1}^{m} \text{Cov}(\varepsilon_A, \varepsilon_{H_i}) = \rho_{\text{HA}}\, \sigma_A \sigma_H,$$

and therefore

$$\text{Corr}(Q_A, \bar{Q}_H) = \text{Corr}(\varepsilon_A, \bar{\varepsilon}_H) = \frac{\rho_{\text{HA}}\, \sigma_A \sigma_H}{\sigma_A(\sigma_H/\sqrt{m})} = \rho_{\text{HA}} \sqrt{m}.$$

Feasibility (positive definiteness of the 2×2 covariance matrix of $(Q_A, \bar{Q}_H)$) implies $|\,\text{Corr}(Q_A, \bar{Q}_H)| < 1$, hence

$$|\rho_{\text{HA}}| < \frac{1}{\sqrt{m}}.$$

Now write $a = \sigma_A^2$, $h = \sigma_H$, $b = h^2/m$, and $c = \text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \rho_{\text{HA}} \sqrt{a}\, h$. The GLS posterior variance is

$$V = \text{Var}(\hat{Q}) = \frac{ab - c^2}{a + b - 2c}, \qquad V_H = \text{Var}(\bar{Q}_H) = b.$$

From the algebra in Theorem 3,

$$V - V_H = -\frac{(c - b)^2}{a + b - 2c} \;\leq\; 0,$$

and feasibility implies $a + b - 2c > 0$ (equivalently $\text{Var}(\varepsilon_A - \bar{\varepsilon}_H) > 0$), so the inequality holds throughout the feasible (non-degenerate) region. Moreover, equality holds iff $c = b$, i.e.,

$$\rho_{\text{HA}} = \rho_{\text{red}} = \frac{b}{\sqrt{a}\, h} = \frac{h}{m\sqrt{a}} = \frac{\sigma_H}{m\sigma_A},$$

which is the redundancy point where GLS sets the AI weight to zero.

Finally, the *formal* singularity of the two-signal GLS expression occurs at

$$\rho_{\text{crit}} = \frac{\sigma_A^2 + \sigma_H^2/m}{2\sigma_A\sigma_H}.$$

By AM–GM,

$$\rho_{\text{crit}} = \frac{\sigma_A^2 + \sigma_H^2/m}{2\sigma_A\sigma_H} \geq \frac{1}{\sqrt{m}}.$$

Hence ρ_{crit} lies on or beyond the feasibility boundary $|\rho_{\text{HA}}| < 1/\sqrt{m}$ and cannot be reached under admissible covariances. Therefore the GLS variance is always well-defined in the feasible region (no blow-up), and adding Q_A under GLS weakly improves precision, with equality only at redundancy. □

Proof. Fix $m \geq 1$ and ρ_{HA} in the informative (non-redundant) region $\rho_{\text{HA}} < \rho_{\text{red}} = \sigma_H/(m\sigma_A)$, equivalently $c < b$. Let $a(\theta) = \sigma_A^2(\theta)$ be strictly decreasing in θ. Write $h = \sigma_H > 0$ and define

$$b := \frac{h^2}{m}, \qquad c := \text{Cov}(\varepsilon_A, \bar{\varepsilon}_H) = \rho_{\text{HA}}\sqrt{a}\, h, \qquad V(a, h; \rho_{\text{HA}}, m) := \frac{ab - c^2}{a + b - 2c}.$$

(i) Precision improvement at fixed human effort. Holding human effort fixed (equivalently holding h fixed), we have

$$\frac{\partial V}{\partial \theta} = V_a(a, h; \rho_{\text{HA}}, m)\, a_\theta(\theta).$$

A direct differentiation yields

$$V_a(a, h; \rho_{\text{HA}}, m) = \frac{(b - c)^2}{(a + b - 2c)^2} \; > \; 0,$$

and feasibility (positive definiteness of the 2×2 covariance of $(Q_A, \bar{Q}_H)$) implies $a + b - 2c > 0$. Since $a_\theta(\theta) < 0$ by assumption, it follows that

$$\frac{\partial V}{\partial \theta} < 0,$$

so improving AI quality strictly lowers posterior variance at the same human effort.

(ii) Throughput improvement at fixed precision target. Fix a target variance σ_{tar}^2. Define $\tau(\theta, \rho_{\text{HA}})$ as the minimal reviewer time per reviewer required to attain

$$V\big(a(\theta), h(\tau); \rho_{\text{HA}}, m\big) \leq \sigma_{\text{tar}}^2, \qquad h(\tau) = \sigma_H(\tau), \qquad \tau = \frac{\sigma_0^2}{s\, h(\tau)^2}.$$

By Theorem 1, $\tau(\theta, \rho_{\text{HA}})$ is weakly decreasing in θ (and strictly decreasing whenever the effort constraint binds in the regular region). Therefore throughput

$$\Lambda(m, \theta, \rho_{\text{HA}}) = \frac{H_{\text{tot}}}{m\,\tau(\theta, \rho_{\text{HA}})}$$

is weakly increasing in θ (and strictly increasing under the same regularity).

Combining (i) and (ii), increasing AI quality shifts the feasible delay–precision frontier outward: it strictly improves precision at fixed human effort and weakly increases throughput at fixed precision, with strict throughput gains whenever Theorem 1 applies with strict inequality. □

B The Need for Experimentation

Although our proposal relies on mathematical analysis of an operational model, the efficacy of the model depends on several real-world aspects of AI capability for review tasks as well as the impact on human reviewers. Fortunately, the model's key parameters are empirically measurable, allowing future validation using real editorial data. Specifically, the AI quality parameter θ can be estimated as the alignment between AI-generated evaluations and historical editorial outcomes (e.g., acceptance or rejection decisions). The correlation parameter ρ_{HA} can be inferred from the overlap between AI and human review texts or numeric evaluation scores—for instance, using cosine similarity or correlation across structured review dimensions. The AI variance σ_A^2 and human variance σ_H^2 can be approximated from review consistency across multiple referees for the same manuscript. Finally, the submission inflow rate λ is directly observable from journal data. These quantities enable the empirical estimation of model inputs and the calibration of ρ_{crit}, θ_{crit}, and $\Lambda(m, \theta, \rho_{\text{HA}})$. As a result, the model can be tested in practice—linking theory to measurable editorial processes and providing a pathway for evidence-based AI policy in academic publishing.

To support this perspective, we propose an experiment design to test the effects of our suggested workflow. For instance, AI reviews could be included in the first round of a journal's special issue review process. Authors can be informed ahead of submission that they may be subject to receiving an initial AI review on their manuscript to help reviewers decide which papers reach the next stage in the creation of the special issue (e.g., a paper presentation workshop). The special issue also allows all those papers that make it to the workshop to then return to follow the journal's current procedures.

The first step consists in training a LLM (e.g., fine-tuning an open-source model such as mistral or llama with recently accepted papers) and generating a prompt based on the set of criteria that the editors of the special issue defined. Then, of N paper submissions received, a fraction, e.g.,

two-thirds are randomly assigned to the treatment condition, which consists in the paper being reviewed by the AI and receiving an AI-generated report. Authors receive the AI-generated report immediately and decide whether to respond to it and provide an additional brief response while holding their original submission fixed, or to withdraw the submission (e.g., if the AI-generated report contains critical issues which the authors agree with but may not be able to address). The response document allows authors to address concerns raised in the AI-generated report and clarify their reasoning for any points they choose not to incorporate. The remaining third of the submissions constitute the control group. Each submission—together with the AI-generated report and author response for papers in the treatment group—is screened by a human AE, who may either desk-reject it or assign it to human reviewers. Among papers sent to review, those in the treatment condition provide reviewers with the paper, the AI-generated report, and the authors' response; those in the control condition provide only the paper. All papers receive human-generated reviews, but only treatment papers receive a first-round AI-generated assessment.

Table 4: Metrics Reported by Human Reviewers

Metric	**Description**
Review time (all papers)	Minutes spent completing the review.
Perceived AI accuracy (treatment only)	Rating of accuracy of the AI-generated report (1 = very low, 5 = very high).
Helpfulness of AI report (treatment only)	Rating of usefulness of the AI report in the review process (1 = very unhelpful, 5 = very helpful).
Qualitative reflection (treatment only)	Brief open-ended assessment of the reviewer's experience with the AI report.

The AE receives all human-generated reviews together with the reviewer-provided metrics and decides whether to move the paper forward to the workshop or reject it. The AE additionally reports the metrics in Table 5. To evaluate the impact of AI-assisted review, we compare papers across conditions using the outcomes in Table 6. Such an experiment would provide the empirical foundation needed to assess whether AI-assisted review can reduce review time and reviewer workload while maintaining decision quality and reviewer satisfaction.

Table 5: Metrics Reported by the Associate Editor

Metric	**Description**
Decision time	Minutes required for the AE to reach a decision.
Perceived quality of human reviews	Rating of the helpfulness and quality of the human-generated reviews.
Workflow assessment	Brief qualitative reflection on the workflow (including AI involvement for treatment papers).

Table 6: Primary Outcomes for Evaluation

Outcome	**Description**
Review time per paper	Broken down by stage: desk decision, reviewer assignment, human review, and final AE decision.
Perception of AI value	Reviewer and AE assessments of the accuracy and usefulness of the AI-generated report.
Perceived decision quality	Whether reviewers and the AE believe the process improved identification of the manuscript's underlying quality.
Acceptance rate	Fraction of papers advanced to the workshop or accepted.
Subjective reflections	Reviewer and AE qualitative commentary on workflow smoothness and tool utility.