Position: Preventing the Collapse of Peer Review Requires Verification-First AI

Lei You [1 2] **Lele Cao** [3 2] **Iryna Gurevych** [4]

Abstract

This paper argues that AI-assisted peer review should be verification-first rather than review-mimicking. We propose *truth-coupling*, i.e. how tightly venue scores track latent scientific truth, as the right objective for review tools. We formalize two forces that drive a phase transition toward proxy-sovereign evaluation: *verification pressure*, when claims outpace verification capacity, and *signal shrinkage*, when real improvements become hard to separate from noise. In a minimal model that mixes occasional high-fidelity checks with frequent proxy judgment, we derive an explicit coupling law and an incentive-collapse condition under which rational effort shifts from truth-seeking to proxy optimization, even when current decisions still appear reliable. These results motivate actions for tool builders and program chairs: deploy AI as an adversarial auditor that generates auditable verification artifacts and expands effective verification bandwidth, rather than as a score predictor that amplifies claim inflation.

1. Introduction

Peer review and venue selection act as the main feedback channel that couples scientific work to recognition and career outcomes. This channel now operates under stress. Submission volumes keep rising, reviewer time does not scale proportionally, and large language models (LLMs) reduce the marginal cost of producing paper-like text. Recent work has explored LLMs as review generators, score predictors, or meta-review assistants (Zhou et al., 2024; Thelwall & Yaghi, 2025; Shin et al., 2025; Bao et al., 2021). A natural "default goal" in this line of work is imitation: we train or prompt a model to produce review text that reads like a human review, or to predict a human score.

[1]Department of Engineering Technology, Technical University of Denmark, Ballerup, Denmark [2]CSPaper Review [3]Microsft Gaming, Stockholm, Sweden [4]Department of Computer Science, Technical University Darmstadt. Correspondence to: Lei You <leiyo@dtu.dk>.

Preprint. June 30, 2026.

Our analysis suggests that we should use AI-assisted peer review to increase truth-coupling by reducing *verification friction* and expanding effective *verification bandwidth*. We should **not** treat agreement with human review text or scores as the main objective, **even if** they come from a venue with historically low noise and high trustworthiness. A feasibility demonstrated in a recent work on automating paper–code consistency checks (Baumgärtner & Gurevych, 2026), complemented by efforts to use LLMs for review assistance and systematic error identification in papers (Liu & Shah, 2023; Xi et al., 2025).

We argue that imitation targets the wrong object: for a fixed paper, a human review score is not a gold label for truth but an outcome of a bandwidth-constrained social process that relies on heuristics and proxy signals (Shah, 2025). A concrete warning sign is a NeurIPS 2024 experiment where an LLM-based checklist assistant improved compliance yet could be gamed by fabricated justifications (Goldberg et al., 2024), illustrating a broader pattern that once proxy signals become objectives, optimization shifts to proxies rather than what they were meant to measure (Goodhart, 1984; Manheim & Garrabrant, 2019; Campbell, 1979).

1.1. A Phase Transition View of Scientific Production

We formalize a split that many researchers informally recognize. Some subfields feel like *truth-sovereign* regimes. In these regimes, a small number of decisive checks, replications, or proofs can settle disputes, and recognition tends to track what is correct over time. Other subfields feel like *proxy-sovereign* regimes. In these regimes, claims arrive faster than the community can verify them, so the evaluation process leans on easier-to-check signals, namely benchmark scores, formal compliance, fashionable phrasing, or dense-looking technicality.

We model this split as a phase transition driven by two forces. First, *verification friction* captures how costly it is for a community to produce decisive evidence per claim. Second, *signal shrinkage* captures how small true improvements become as a domain approaches an S-curve (Bloom et al., 2020). When the true signal shrinks, even honest progress becomes a low signal-to-noise problem. Proxy signals turn more attractive as they are easier to produce and evaluate.

1.2. Three Implications That Are Easy to Miss

Our model yields three implications that are easy to miss.

Implication 1 (High prestige is a deceptive lagging indicator). A scientific venue can appear statistically healthy while being structurally doomed. Our analysis shows that "high prestige" is merely a historical record of low verification pressure. Therefore, fitting AI agents to mimic top-tier venues merely replicates past signals without inheriting the verification capacity required to withstand the coming surge in AI-generated claims.

Implication 2 (Review-mimicking automates the collapse). Training prompting AI to predict human scores scales the *output* of evaluation (scores) without scaling the *capacity* for scrutiny (verification). This does not solve the review crisis; it hides it. By decoupling "getting a high score" from "passing a hard check," mimicking tools effectively remove the only friction that stopped authors from gaming the system, accelerating the drift into "Zombie Science".

Implication 3 (The goal is pressure control, not score optimization). To survive the AI era, peer review must shift its objective from "predicting quality" to "controlling verification pressure". The only legitimate role for AI is to act as an adversarial auditor that physically expands effective bandwidth by executing evidence. Any tool that merely generates text or scores without checking the underlying reality is simply inflating the bubble.

1.3. What We Contribute

We make four contributions. First, we define verification pressure and signal shrinkage as the two main drivers of a phase transition from truth sovereignty to proxy sovereignty. Second, we derive a closed-form expression for truth-coupling $\rho = \mathrm{Corr}(S, T)$ under a simple but interpretable mixture model of verification and proxy evaluation. Third, we analyze researcher incentives and show a collapse condition under which truth effort becomes privately irrational. Fourth, we translate these results into design principles for AI-assisted peer review, and we outline an empirical agenda for measuring the key quantities in real venues.

1.4. Roadmap

Section 2 grounds the model in a sociological description of two regimes. Sections 3 and 4 introduce the formal model and the main results. Section 5 translates the model into verification-first design principles for AI-assisted review. Section 6 describes credible alternative views and explains where we agree and disagree. Section 7 synthesizes these findings into concrete policy recommendations. Section 8 concludes the paper. Appendix B reviews relevant theory on metrics, incentives, and peer review automation.

2. Universe A and Universe B

We use the terms *Universe A* and *Universe B* as shorthand for two regimes that can coexist inside the same conference or department of a university. The labels are not moral categories. They describe two different ways a field closes its feedback loop between work and recognition.

2.1. Two Kinds of Sovereignty

In Universe A, *truth is sovereign*. A claim eventually lives or dies based on a relatively decisive interaction with the target of study, such as a proof, a falsifiable experiment, or a replication with high fidelity. Human judgment still matters, but it matters mostly for allocating attention, not for deciding what is true.

In Universe B, *proxy signals are sovereign*. Claims arrive faster than the community can verify them, so the system relies on low-cost signals that correlate with truth only weakly. Examples include small benchmark gains, compliance checklists, citation cascades, or the appearance of technical depth. When such signals decide careers, researchers rationally learn to optimize them. We treat citation cascades as one concrete example of a low-cost proxy signal, and we analyze this case in Appendices F.

2.2. Career Dynamics as An Observable Signature

The two regimes also differ in how they shape careers. In Universe A, the main scarce resource is insight. Insight is hard to outsource and hard to parallelize. As a result, even senior researchers often stay close to the technical frontier. A principal investigator (PI) may supervise students, but the PI usually keeps strong ownership of the key ideas and can personally verify core arguments.

In Universe B, the main scarce resource is attention and throughput. Outputs aggregate linearly across people, and the evaluation system rewards volume. As a result, career success favors management of a large production pipeline. Many successful PIs in Universe B shift away from hands-on verification and toward fundraising, staffing, and strategic positioning. This behavior is not a personal failure. It is a predictable response to incentives in a proxy-sovereign regime (Holmstrom & Milgrom, 1991; Power, 1997).

2.3. Verification Friction and Evidence Compressibility

A common confusion is to equate "easy to verify" with "cheap to verify." Some of the most truth-sovereign disciplines have high absolute verification cost. A major mathematical proof may take months of expert effort to check. What matters is not the raw cost c but whether the evidence is *decisive and compressible*.

We describe this using an *effective verification cost*

$$C_{\text{eff}} = \frac{c}{\kappa}, \tag{1}$$

where $\kappa \in (0, 1]$ denotes the fidelity or decisiveness of the evidence. A careful proof has large c but often has κ close to one. A noisy empirical study can have large c and small κ, which produces a large C_{eff}. A simulation study can have small c but also small κ when the simulator fails to capture the real-world mechanism of interest. In that case, C_{eff} can still be large. This is why proxy sovereignty often emerges in simulator-grounded mature engineering domains.

2.4. Signal Shrinkage and Saturation

A second driver of proxy sovereignty is the shrinkage of true effect sizes. As domains mature, improvements become smaller and harder to measure (Bloom et al., 2020). When the between-paper variance in true value shrinks, reviewers face a low signal-to-noise problem even if they act in good faith. Proxy signals then become tempting because they are easier to compare at scale.

Appendix H provides additional case studies and maps them to distinct regions of Universe A and Universe B.

3. Model

We propose a stylized model that keeps two features explicit. First, verification is scarce. Second, as fields saturate, true improvements become harder to distinguish from noise.

3.1. Latent Truth, Proxy Judgment, and Score

Each submission has an underlying (possibly multi-dimensional) scientific value; we model the venue-relevant scalar projection as T. In mathematics it corresponds to correctness and conceptual novelty. In empirical work it corresponds to validity, robustness, and generality. We treat T as a random variable with $\mathrm{Var}(T) > 0$ across the population of submissions to a venue in a fixed period.

A venue assigns each submission a scalar decision signal S that influences acceptance and future allocation of attention.

Assumption 3.1 (Mixture evaluation model)**.** With probability $q \in [0, 1]$, reviewers obtain verification evidence and assign a score based on T. With probability $1 - q$, reviewers rely on a proxy judgment that equals $T + \Delta$, where Δ is a zero-mean proxy error.[1] We assume Δ is independent of T.

Under Assumption 3.1, we can write

$$\begin{aligned} S &= qT + (1 - q)(T + \Delta) \\ &= T + (1 - q)\Delta. \end{aligned} \tag{2}$$

[1]The zero-mean assumption is across submissions: decisions are made once per paper, so proxy noise does not average out.

This expression absorbs many details into Δ. It captures random proxy noise, model instability, benchmark overfitting, and strategic "gaming" of evaluation criteria. We treat q as an *endogenous* quantity determined by verification capacity, which we formalize next.

Definition 3.2 (Truth-coupling)**.** We define truth-coupling as $\rho := \mathrm{Corr}(S, T)$.

Truth-coupling is high when a venue's selection process preserves a strong relationship between recognition and latent value. Truth-coupling is low when proxy errors dominate.

3.2. Verification Pressure

We describe the verification bottleneck using three quantities. The claim rate R denotes the number of claims that a venue processes per unit time. This quantity grows with submission volume and with the ease of producing plausible claims. The effective verification cost C_{eff} denotes the cost of producing decisive evidence per claim, as in (1). The effective verification bandwidth $\mathcal{B}_{\text{eff}}$ denotes how much decisive evidence the community can process per unit time. This term bundles reviewer time, reviewer expertise, and the availability of external checks such as replications, proof checking, or artifact evaluation.

We define verification pressure as

$$\Lambda := \frac{R\, C_{\text{eff}}}{\mathcal{B}_{\text{eff}}}. \tag{3}$$

When $\Lambda \leq 1$, the community can in principle verify a large fraction of claims. When Λ grows above one, verification becomes selective, and q decreases.

Assumption 3.3 (Verification frequency)**.** The probability of verification is

$$q = \min\left\{1,\ \frac{1}{\Lambda}\right\} = \min\left\{1,\ \frac{\mathcal{B}_{\text{eff}}}{R\, C_{\text{eff}}}\right\}. \tag{4}$$

This assumption treats verification as a capacity-constrained queue with random service. It is not a detailed model of program committee behavior. It is a transparent abstraction that makes the core dependence explicit.

3.3. Signal Shrinkage and Noise-to-Signal Ratio

We also need a way to represent saturation. We use the variance of T across submissions as a proxy for how much true signal is available to reviewers. In many domains, true effect sizes shrink over time (Bloom et al., 2020). This shrinkage reduces $\mathrm{Var}(T)$.[2]

[2]Nascent fields like XAI mirror "pre-ImageNet" CV: undefined T and high C_{eff} yield low ρ. ImageNet temporarily anchored Universe A by serving as verification infrastructure (lowering C_{eff}), until benchmark saturation and gaming ($\mathrm{Var}(\Delta) \uparrow$) triggered a regression to Universe B.

We summarize the balance between proxy error and truth signal with a dimensionless ratio

$$r := \frac{\mathrm{Var}(\Delta)}{\mathrm{Var}(T)}. \tag{5}$$

Large r means proxy noise dominates. Small r means true value dominates. Both verification pressure Λ and saturation can increase r. The former can increase $\mathrm{Var}(\Delta)$ through gaming and instability (Appendix D), while the latter decreases $\mathrm{Var}(T)$. Appendix E discusses practical proxies for signal shrinkage and sketches how we can estimate $\mathrm{Var}(T)$ from benchmark time series and robustness reports.

4. Theory

We now analyze how verification pressure Λ and noise-to-signal ratio r shape truth-coupling and researcher incentives. All proofs appear in Appendix C.

4.1. Truth-Coupling as a Closed Form

Our first result links truth-coupling to q and r.

Theorem 4.1 (Truth-coupling under mixture evaluation)**.** *Under Assumption 3.1, truth-coupling satisfies*

$$\rho \;=\; \mathrm{Corr}(S,T) \;=\; \left(1+(1-q)^2 r\right)^{-1/2}. \tag{6}$$

Theorem 4.1 makes two points explicit. First, verification scarcity hurts coupling through $(1-q)^2$. Second, saturation hurts coupling through r. Even if reviewers act in good faith, small true effect sizes make it hard to preserve a strong relationship between score and truth.

A direct corollary translates a coupling target into a required verification frequency.

Corollary 4.2 (Coupling budget)**.** *Fix a target $\rho_{\min} \in (0,1)$. If $r > 0$, then $\rho \geq \rho_{\min}$ holds whenever*

$$q \;\geq\; 1 - \sqrt{\frac{\rho_{\min}^{-2} - 1}{r}}. \tag{7}$$

Corollary 4.2 shows that coupling constraints become more demanding as r grows. This implies that a field saturates and true gains become smaller ($\mathrm{Var}(T)$ shrinks), maintaining the same level of truth-coupling requires *higher* verification frequency. By Assumption 3.3, this translates into a bandwidth requirement $\mathcal{B}_{\mathrm{eff}} \geq q_{\min} R C_{\mathrm{eff}}$.

4.2. A Phase Diagram: From Universe A to Universe B

Section 2 introduced Universe A and Universe B as two idealized feedback regimes. Figure 1 makes this link concrete. We use truth-coupling $\rho \triangleq \mathrm{Corr}(S,T)$ as an order parameter. When ρ is close to 1, evaluation is truth-sovereign

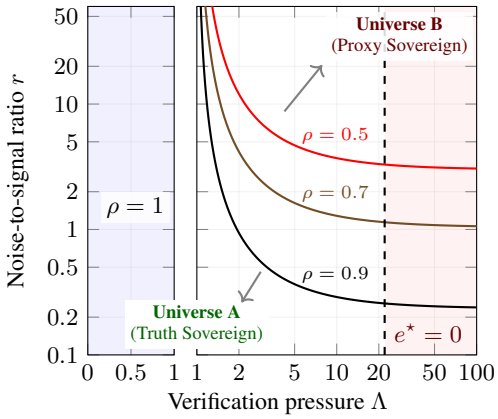

Figure 1. **Truth-coupling phase diagram.** Contours of $\rho = \mathrm{Corr}(S,T)$ over verification pressure Λ and noise-to-signal ratio r (Theorem 4.1). For $\Lambda \leq 1$ we have full verification ($q = 1$) and thus $\rho = 1$. For $\Lambda > 1$, truth-coupling degrades as verification becomes scarce (larger Λ) or proxy noise increases (larger r).

and the feedback loop closes through verification. When ρ is small, evaluation is proxy-sovereign and the loop closes through cheap signals that can become targets.

The left panel shows a simple but important limit. When $\Lambda \leq 1$, verification keeps up with claims, so $q = 1$. Then the mixture model gives $\rho = 1$ for any r. In other words, if we can verify essentially everything, proxy noise does not control outcomes.

The right panel shows the regime $\Lambda > 1$, where verification becomes scarce. Here q decreases as Λ increases, so even honest reviewers must lean more on proxies. Moving right, toward larger Λ, lowers ρ because more decisions occur in the proxy mode. Moving up, toward larger r, also lowers ρ because proxy variation dominates true variation.

The dashed line is *not* a boundary in ρ. It marks an incentive threshold from our effort model. To the right of this line, the researcher's optimal truth effort can hit the corner solution $e^\star = 0$, even if the field currently has moderate noise. We provide the formal derivation of this incentive collapse in Theorem 4.4.

We remark that sometimes $\mathrm{Var}(\Delta)$ (and hence r) may itself increase with verification pressure Λ. Appendix D gives a stylized mechanism showing how selective reporting and best-of-K search can make this happen.

4.3. A Collapse Condition for Truth Effort

Truth-coupling describes how well a venue's scores track latent scientific value at the community level. We also need to understand incentives at the level of individual authors. Even if reviewers act in good faith, authors respond to what the system rewards.

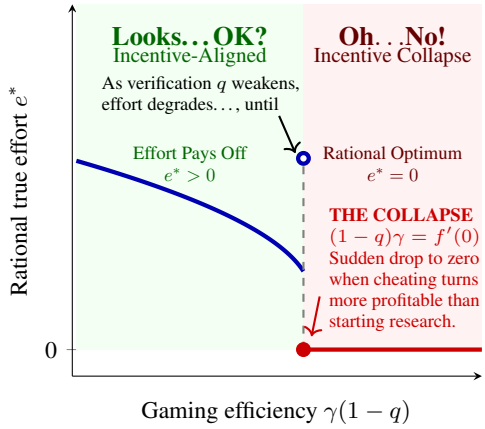

Figure 2. **Incentive collapse for truth effort.** The researcher's optimal truth effort $e^\star$ as a function of the effective proxy return $(1-q)\gamma$ (Theorem 4.4). As $(1-q)\gamma$ increases, $e^\star$ decreases and eventually hits the corner solution $e^\star = 0$ when $(1-q)\gamma \geq f'(0)$.

We use a minimal effort allocation model. Each researcher chooses how to spend a unit of effort. Truth-oriented effort supports claims that would survive careful checking. Proxy-oriented effort improves how the work looks under cheap signals, such as marginal benchmark tuning or presentation polish. When verification is common, proxy optimization matters less because fewer decisions rely on proxies. When verification is scarce, proxy optimization matters more because more decisions rely on proxies.

Assumption 4.3 (Effort and payoff)**.** A researcher chooses $e \in [0,1]$ to maximize

$$S(e) \;=\; f(e) \;+\; (1-q)\gamma(1-e), \tag{8}$$

where $S(e)$ follows (2) and f captures truth-oriented production (differentiable, increasing, and concave), and $\gamma > 0$ scales the effectiveness of proxy optimization in the proxy mode. We allow γ to increase with tools that facilitate proxy optimization, such as automated writing.

Theorem 4.4 (Truth effort decreases with verification scarcity)**.** *Under Assumption 4.3, any maximizer $e^\star$ satisfies the first-order condition*

$$f'(e^\star) = (1-q)\gamma \tag{9}$$

whenever $e^\star \in (0,1)$. Moreover, if $(1-q)\gamma \geq f'(0)$, then an optimal solution satisfies $e^\star = 0$.

Theorem 4.4 delivers a brutal interpretation of the researcher's dilemma. The author faces a trade-off between the marginal return on truth, $f'(e^\star)$, and the temptation of proxy optimization, $(1-q)\gamma$. As verification bandwidth saturates ($q \to 0$) or AI tools cheapen mimicry ($\gamma \uparrow$), the incentive structure tilts aggressively against truth.

The theorem's most critical implication is the corner solution $e^\star = 0$. If the return on cheap proxies exceeds the initial value of doing real science—specifically, if $(1-q)\gamma \geq f'(0)$, rational effort collapses completely, illustrated in Figure 2. Crucially, this collapse condition is independent of the field's current noise level or historical reputation. Mathematically, this implies that 'top-tier' status serves merely as a historical record of low verification pressure with low noise-to-signal ratio, rather than a structural firewall against AI-generated claims.

4.4. Theoretical Summary: The Capacity Trap

Our theoretical analysis leads to a sobering conclusion: under the exponential pressure of AI-driven claim inflation, **incentive collapse is not merely a risk, but the default structural destiny** of the peer review system. We have demonstrated that truth-coupling ρ acts as a deceptive lagging indicator, capable of masking deep fragility in historically trusted fields (i.e. high ρ fields can also have $e^\star \approx 0$ collapse). As long as verification bandwidth $\mathcal{B}_{\text{eff}}$ remains static while the volume of claims R and the ease of mimicry γ scale exponentially, the collapse condition derived in Theorem 4.4 becomes a mathematical inevitability. The system is structurally condemned to drift toward Universe B.

However, the math also reveals the escape: expand the verification bandwidth ($\mathcal{B}_{\text{eff}}$) to reverse the collapse inequality. This necessity, i.e. to shift from *score prediction* to *verification*, sets the stage for the verification-first paradigm.

5. Implications for AI-Assisted Peer Review

> *"You can not stop the cold wind from blowing, but you can build a house to shelter you from it."*
> *— Isaac Asimov, Foundation (Asimov, 1951)*

In our context, the "cold wind" is the exponential growth of AI-generated noise, and the "house" is the verification capacity. Just as a gas obeys the ideal gas law regardless of the motion of individual molecules, the scientific community obeys these collapse dynamics regardless of individual ethical intent. To construct this "house", our results in Section 4 suggest a fundamental design principle.

Principle (verification-first). *We should use AI to reduce verification friction and increase truth-coupling, not to imitate human review text or to predict human scores.*

We now justify this principle and translate it into design guidelines using the (q,r)-framework.

5.1. Why Score Imitation Is the Wrong Default Goal

Many "AI reviewer" projects optimize agreement with human reviews or acceptance decisions (Zhou et al., 2024; Thelwall & Yaghi, 2025; Bao et al., 2021). Imitation can be

a reasonable target for writing assistance, where we want outputs to match human style. It is not a good default target for evaluating a fixed scientific claim.

A human score comes from the same bandwidth-limited system we aim to improve. When Λ is large and r is large, Theorem 4.1 implies that truth-coupling ρ is low. In that regime, scores are weak signals of T by construction. If we train or prompt an AI reviewer to match those scores, we teach it to reproduce the same proxy-weighted rule, including its blind spots. Even training on a historical "high-SNR" period can at best reduce r; it does not increase verification frequency q, and at scale imitation tools tend to raise R and γ, lowering q and reintroducing proxy-sovereign dynamics.

Imitation-first deployment can also worsen incentives. Tools can increase the effective claim rate R by making proxy-aligned submissions cheaper to produce, and increase γ in Theorem 4.4 by making proxy criteria easier to optimize. Both changes raise verification pressure and move the system toward proxy sovereignty.

Our claim is not "automation instead of reviewers": human reviews still contain useful, work-specific signal, and verification-first AI should complement humans by producing auditable evidence for them to weigh (Dutta et al., 2025).

5.2. The Objective: Reducing Verification Pressure

Our theoretical analysis implies that we cannot simply optimize for truth-coupling ρ as a static metric, as it can be a deceptive lagging indicator without capture structurally the deep fragile. Instead, we must target the *drivers* of the potential collapse. Theorem 4.4 establishes that stability depends on keeping verification pressure $\Lambda = RC_{\text{eff}}/\mathcal{B}_{\text{eff}}$ low enough to satisfy the incentive constraint. This transforms the design goal from "better prediction" to "capacity engineering," pointing to three concrete operational levers, reducing R (Gatekeeping), reducing C_{eff} (Auditing), or increasing $\mathcal{B}_{\text{eff}}$ (Scaling).

Unlike mimetic methods that try to fit S to T directly, verification-first AI targets these levers. We formally state this design goal as an *evidence acquisition problem*. Let $\mathcal{D}$ be the distribution over submissions with observable proxy P and latent truth T. The AI policy is not a scoring function, but an *auditing policy* π that selects a verification action $V \sim \pi(\cdot \mid P)$ to reveal evidence. The final score S is then derived from this evidence.

We seek policies that maximize the revelation of truth within the community's bandwidth limits.

Verification-First AI Peer Review

$$\max_{\pi} \ \operatorname{Corr}(S_\pi, T) \ - \ \lambda\, \mathbb{E}_\pi\big[\operatorname{Cost}(V)\big], \qquad (10)$$

Here, λ penalizes verification cost, and S_π is the score derived from the evidence V under policy $\pi \in \Pi_{\text{verify}}$. Crucially, we restrict Π_{verify} to auditable evidence-producing actions, preventing the optimizer from exploiting cheap proxies. This forces the tool to find the most efficient *decisive* test that maintains truth-coupling.

5.3. Design Guidelines from the "(q, r)-Framework"

Reducing R without suppressing novelty. A naive response to high R is to raise barriers to entry. That can protect incumbents and can suppress high-risk ideas. A verification-first alternative is to *separate claim production from claim validation*. For example, a venue can implement a staged pipeline where most submissions first receive a lightweight auditability check. This check asks whether the core claims map to specific artifacts, namely code, data, proofs, or ablation plans. Only submissions that pass this gate consume full reviewer bandwidth.

We can also reduce R by changing what counts as a "claim." If a paper makes ten loosely connected claims, reviewers cannot verify any of them well. A venue can ask authors to declare one primary claim and to attach a verification plan for that claim. This reframes evaluation as claim-evidence pairing rather than as an unconstrained narrative.

Reducing C_{eff} by producing higher-fidelity evidence. Equation (1) suggests that C_{eff} drops when κ increases. AI can help by turning vague claims into explicit, testable units and by generating minimal tests with high diagnostic value. We highlight four families of verification artifacts that AI can scale.

Claim-evidence maps. Given a manuscript, an assistant can extract the main claims and link each claim to the exact evidence that supports it. This helps reviewers allocate attention and reduces time spent searching.

Smoke tests. Given code and data, an assistant can run a small set of checks that catch common failure modes, such as non-determinism, data leakage, and baseline misconfiguration. These checks do not prove correctness, but they increase κ by eliminating weak forms of evidence.

Variance audits. If a result depends on random seeds, an assistant can automatically run repeated trials, report dispersion, and flag results that rely on a small number of favorable runs. This directly targets the instability component of Δ.

Stress tests. For empirical claims, an assistant can generate perturbations, distribution shifts, or adversarial subsets that test whether a reported gain is robust. For theoretical claims with implementations, an assistant can search for counterexamples within a declared scope.

Each family reduces C_{eff} because it increases the amount

A paper reports a +0.5% gain on ImageNet. Real?

1. The imitation path (accelerating Universe B): The AI acts as a *simulator*. It scans the text, recognizes the "SOTA" claim, and generates a plausible-sounding review praising the "novelty" while nitpicking formatting or missing citations.

2. The verification-first path (restoring Universe A): The AI acts as an *adversarial auditor*. Instead of producing an opinion from the narrative, it extracts the core claim and generates a **stress test** to verify it. It automatically runs a seed sweep and checks sensitivity to hyperparameters.

Figure 3. Imitation AI automates the *appearance* of peer review, optimizing for style match. Verification-first AI automates the *labor* of scientific scrutiny, optimizing for truth discovery.

of decisive information per unit reviewer effort. These artifacts also reduce $\mathrm{Var}(\Delta)$, because they remove degrees of freedom for proxy gaming.

Increasing $\mathcal{B}_{\text{eff}}$ by compressing expert attention. The bandwidth $\mathcal{B}_{\text{eff}}$ is not just "more reviewers." It is also the ability to compress expert attention into reusable checks. AI can increase $\mathcal{B}_{\text{eff}}$ by standardizing and automating verification artifacts, by routing papers to reviewers with matching expertise, and by generating concise summaries of evidence that reviewers can trust. A key requirement is that these summaries remain grounded in verifiable artifacts. Summarizing a paper-like narrative does not increase $\mathcal{B}_{\text{eff}}$ in our sense. Summarizing verified evidence does.

Flattening proxy gradients to resist Goodhart drift. Even with better verification, proxy signals will remain. We should therefore reduce the incentive to over-optimize them. Theorem 4.4 suggests one mechanism: reduce the marginal score benefit of proxy effort in the non-verification mode. Mechanism design can help.

A probabilistic audit mechanism serves as a critical deterrent. By subjecting a random subset of accepted papers to rigorous, high-cost verification (e.g., code reproduction), we introduce a *negative expected utility* for fragile or fabricated results, even if the audit probability is low. Crucially, this creates an asymmetric risk profile: if authors know that acceptance carries a probability of a high-fidelity audit, then proxy gains that cannot survive scrutiny become liabilities.

5.4. Case Study: The "SOTA" Optimizer Mirage

Figure 3 reveals a critical *paradox of efficiency*. Superficially, imitation-first AI appears to boost productivity by generating text instantly; in reality, it acts as an *inflationary engine*. By commoditizing the production of "plausible" reviews, it inflates the noise floor (R) without expanding the actual verification bandwidth ($\mathcal{B}_{\text{eff}}$). This creates a "Review Theater" where agents mimic the rituals of evaluation without performing the work. In contrast, verification-first AI introduces *salutary friction*. It does not merely "check" the paper; it *collapses the state space* of selective reporting ($\mathrm{Var}(\Delta)$). By automating the "dirty work" of code execution and sensitivity analysis, it structurally removes the degrees of freedom that authors use to mine for significance. While this approach feels more "expensive" per paper (requiring compute), it is the only mathematical path to flip the inequality in Theorem 4.4 and make truth-seeking ($e^* > 0$) the rational strategy again.

5.5. Why AI Plausibly Can Help Verify Research Result

Automatic review generators can be largely insensitive to faulty reasoning under a controlled counterfactual evaluation, indicating that text-level "review mimicry" fails precisely on core reviewing skills (Dycke & Gurevych, 2025). This motivates verification-first designs that target *structured* reasoning and checkable sub-tasks (Dycke et al., 2025). Concretely, SciCoQA operationalizes verification as paper–code alignment auditing, turning a key class of claims into measurable, reusable QA artifacts (Baumgärtner & Gurevych, 2026). Together, these works support a pragmatic thesis: AI is most credible for verification when it produces inspectable evidence trails. Appendix G outlines concrete measurement strategies.

6. Alternative Views

To engage credible disagreement, we describe several alternative views that we often hear in discussions of AI-assisted peer review. We then explain why we still advocate a verification-first objective.

6.1. View 1: Review Mimicry Is a Pragmatic Target

A common view is that human review scores and texts are the best available labels at decision time. Under this view, we should train LLMs to predict scores or to generate reviews that match past reviewer behavior because doing so can reduce workload and speed up decisions (Zhou et al., 2024; Thelwall & Yaghi, 2025). This view is credible. Scientific truth is not directly observable at review time, and reviewers already represent the community's current norms.

We agree that language models can help with *communication* tasks, such as summarizing a submission, mapping claims to supporting evidence, or helping authors rewrite unclear parts. However, score mimicry is a problematic *optimization target*. In our model, a score is an output of a bandwidth-constrained process with nonzero proxy noise, so truth-coupling can be low under high verification pressure Λ even when reviewers act in good faith. Optimizing an AI system to match scores therefore optimizes toward this low-coupling mixture; moreover, fully automatic re-

view generation can miss core reviewing skills (e.g., detecting faulty reasoning), underscoring the need for human judgment (Dycke & Gurevych, 2025). Mimicry can also increase the claim rate R by making persuasive text cheaper and proxy heuristics easier to satisfy, which raises Λ and can further reduce coupling.

We do not argue that LLM assistance is useless. Rather, imitation should not be the primary objective for AI-assisted *evaluation*. Instead, AI should support *verification-first* workflows by producing auditable evidence that humans can interpret and weigh.

6.2. View 2: We Should Avoid AI in Review Entirely

Another credible view is that we should not use LLMs in peer review at all. Concerns include confidentiality, hidden biases, and the risk that automated tools create new forms of review theater (Akella et al., 2025). This view is also credible because of high-stakes setting and costly failure.

We acknowledge these risks, but we argue that *abstinence is a strategy for certain defeat*. Our theoretical analysis reveals a harsh reality: the verification bottleneck is not static; it is widening exponentially due to the rising claim rate R. As long as human bandwidth $\mathcal{B}_{\text{eff}}$ remains fixed, the collapse condition derived in Theorem 4.4 becomes a *mathematical inevitability*, regardless of ethical intent. Rejecting AI assistance does not preserve the sanctity of human review; it structurally condemns the system to be overwhelmed by noise, forcing a regression to Universe B. Therefore, the deployment of AI is not optional but existential. We must deploy AI to radically expand $\mathcal{B}_{\text{eff}}$—not to mimic opinions, but to execute the artifact-grounded audits (e.g., code execution, seed sweeping) that act as the only mathematical counterweight to the coming flood of cheap claims.

6.3. View 3: The Main Problem Is Bias and Inconsistency, Not Verification

A third view is that the main failure mode of peer review is social bias and inconsistency. Under this view, AI should focus on calibrating scores, detecting reviewer bias, or standardizing rubrics. We agree that bias and inconsistency are real. We also agree that some proxy signals can encode status and network effects.

We nevertheless expect verification-first mechanisms to reduce bias: when truth-coupling is low, small social effects can dominate decisions, but decisive evidence leaves fewer degrees of freedom for subjective bias. Thus our framework complements fairness-oriented work: bias reduction can be an additional objective, but not one pursued via score imitation. Instead, we should improve evidence fidelity and reduce proxy noise, strengthening both epistemic reliability and procedural fairness.

7. Call to Action

Our theoretical analysis implies concrete steps. We emphasize feasible changes within existing conference and journal structures.

7.1. What Venues Can Do

Venues should treat verification bandwidth as scarce and allocate it deliberately. A practical step is a staged review pipeline. Stage 1 screens *auditability*: reviewers and automated checks verify that the main claims map to concrete artifacts, scripts, data, or proofs. Stage 2 concentrates deep verification on a smaller set of submissions, including artifact reruns and robustness checks. This reduces the effective claim rate R that consumes full bandwidth.

Venues should also commit to random post-acceptance audits. Even a small audit rate weakens incentives to optimize fragile proxies. Audits can be lightweight but decisive, such as rerunning code, checking baselines, or validating a theorem. Audit outcomes can be published as additive addenda, so the process stays corrective rather than punitive.

7.2. What Tool Builders Can Do

Tool builders should prioritize evidence-grounded modules over review text generation. Examples include claim–evidence maps, reproducibility smoke tests, variance audits, and stress tests. Outputs should be inspectable and reusable, namely commands, logs, and minimal evidence bundles. We should evaluate tools by whether they reduce C_{eff}, increase $\mathcal{B}_{\text{eff}}$, or reduce r in downstream decisions.

Tool builders should also support shared, open audit suites. Community-maintained diagnostic tests let venues reuse results and compress expert effort into standardized checks, increasing effective bandwidth.

7.3. What Funders and Institutions Can Do

Verification-first systems require infrastructure. Funders can support shared testbeds, reproducibility tooling, and long-horizon evaluation datasets. Institutions can recognize verification work as a first-class contribution, including replications, negative results, and engineering that makes evidence cheaper to check.

8. Conclusion

AI can either amplify proxy sovereignty or help restore truth sovereignty. Our results suggest that the core challenge is not to generate better review prose. It is to design evidence, incentives, and bandwidth allocation so that "what gets rewarded" stays coupled to "what is true."

Acknowledgement

We thank Dr. Nihar Shah for insightful discussions and for early, sustained advocacy of verification-oriented uses of AI in peer review, and Nils Dycke and Subhabrata Dutta for valuable comments that helped clarify assumptions and sharpen the presentation.

References

Afzal, O. M., Nakov, P., Hope, T., and Gurevych, I. Beyond "not novel enough": Enriching scholarly critique with llm-assisted feedback. *arXiv preprint*, 2025.

Akella, A. P., Siravuri, H. V., and Rohatgi, S. Pre-review to peer review: Pitfalls of automating reviews using large language models. *arXiv preprint*, 2025.

Asimov, I. *Foundation*. Gnome Press, New York, 1951.

Bao, P., Hong, W., and Li, X. Predicting paper acceptance via interpretable decision sets. In *Companion Proceedings of the Web Conference 2021*, pp. 461–467, 2021.

Baumgärtner, T. and Gurevych, I. Scicoqa: Quality assurance for scientific paper–code alignment. *arXiv preprint*, 2026.

Bloom, N., Jones, C. I., Van Reenen, J., and Webb, M. Are ideas getting harder to find? *American Economic Review*, 110(4):1104–1144, 2020.

Campbell, D. T. Assessing the impact of planned social change. *Evaluation and program planning*, 2(1):67–90, 1979.

Dutta, S., Kaufmann, T., Glavaš, G., Habernal, I., Kersting, K., Kreuter, F., Mezini, M., Gurevych, I., Hüllermeier, E., and Schuetze, H. Problem solving through human–ai preference-based cooperation. *Computational Linguistics*, pp. 1–36, 2025.

Dycke, N. and Gurevych, I. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *arXiv preprint*, 2025.

Dycke, N., Zečević, M., Kuznetsov, I., Suess, B., Kersting, K., and Gurevych, I. Stricta: Structured reasoning in critical text assessment for peer review and beyond. In *Proceedings of ACL 2025*, pp. 22687–22727, 2025.

Goldberg, A., Ullah, I., Khuong, T. G. H., Rachmat, B. K., Xu, Z., Guyon, I., and Shah, N. B. Usefulness of llms as an author checklist assistant for scientific papers: Neurips'24 experiment. *arXiv preprint*, 2024.

Goldstone, J. A. Ages of discord: A structural demographic analysis of american history. *Evolutionary Studies in Imaginative Culture*, 2(1):137–142, 2018.

Goodhart, C. A. Problems of monetary management: the uk experience. In *Monetary theory and practice: The UK experience*, pp. 91–121. Springer, 1984.

Hao, Q., Xu, F., Li, Y., and Evans, J. Artificial intelligence tools expand scientists' impact but contract science's focus. *Nature*, pp. 1–7, 2026.

Holmstrom, B. and Milgrom, P. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization*, 7(special_issue):24–52, 1991.

Ioannidis, J. P. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., and Schwartz, R. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. In Walker, M., Ji, H., and Stent, A. (eds.), *Proceedings of ACL 2018*, pp. 1647–1661, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., and Zou, J. Y. Monitoring ai-modified content at scale: a case study on the impact of chatgpt on ai conference peer reviews. In *Proceedings of ICML 2024*, ICML'24. JMLR.org, 2024.

Liu, R. and Shah, N. B. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

Manheim, D. and Garrabrant, S. Categorizing variants of Goodhart's law. *arXiv preprint*, 2019.

Muller, J. *The tyranny of metrics*. Princeton University Press, 2018.

Power, M. *The Audit Society: Rituals of Verification*. Oxford University Press, 1997.

Shah, B. N. An overview of challenges, experiments, and computational solutions in peer review (extended version). *preprint*, 2025. URL https://www.cs.cmu.edu/~nihars/preprints/SurveyPeerReview.pdf.

Shin, H., Tang, J., Lee, Y., Kim, N., Lim, H., Cho, J. Y., Hong, H., Lee, M., and Kim, J. Automatically evaluating the paper reviewing capability of large language models. *arXiv preprint*, pp. arXiv–2502, 2025.

Sims, C. A. Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690, 2003.

Smaldino, P. E. and McElreath, R. The natural selection of bad science. *Royal Society open science*, 3(9):160384, 2016.

Strathern, M. *Audit cultures*, volume 146. London: Routledge, 2000.

Tainter, J. *The collapse of complex societies*. Cambridge university press, 1988.

Thelwall, M. and Yaghi, A. Evaluating the predictive capacity of chatgpt for academic peer review outcomes across multiple platforms. *Scientometrics*, pp. 1–23, 2025.

Xi, S., Rao, V., Payan, J., and Shah, N. B. Flaws: A benchmark for error identification and localization in scientific papers. *arXiv preprint arXiv:2511.21843*, 2025.

Zhou, R., Chen, L., and Yu, K. Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In *Proceedings of LREC-COLING 2024*, pp. 9340–9351, Torino, Italia, May 2024. ELRA and ICCL.

Zou, J. Chatgpt is transforming peer review—how can we use it responsibly. *Nature*, 635(8037):10–10, 2024.

Appendix: Preventing the Collapse of Peer Review Requires Verification-First AI

A Appendix Overview 11
B Related Work 11
C Proofs 12
D A Stylized Goodhart Amplification Mechanism 14
E Measuring Signal Shrinkage in Practice 15
F Citations as Proxy Signals and a Winner's Curse 16
G Empirical Agenda 21
H Case Studies: Mature Engineering Domains as Extreme Regimes 22
I Limitations 23

A. Appendix Overview

This appendix collects proofs and extensions that support the main text.

Appendix B summarizes the related work to this paper. Appendix C provides full proofs for Theorems 4.1 and 4.4 and Corollary 4.2. Appendix D provides a stylized derivation showing how gaming variance can grow rapidly with verification pressure when actors select the best outcome from many proxy-optimization attempts. Appendix E discusses practical ways to measure signal shrinkage and the noise-to-signal ratio r in real venues. Appendix F formalizes how low truth–coupling regimes can arise and shows large cross-field disparities in the expected scientific value per citation can follow. Appendix G outlines how we can measure verification pressure and signal shrinkage in practice. Appendix H gives short case studies that connect the model to several mature engineering domains. Appendix 7 proposes concrete steps that venues, tool builders, and funders can take. Appendix I discussed the limitations of this paper.

B. Related Work

B.1. Metrics, Incentives, and Goodhart Drift

A central concern in our paper is the gap between a latent objective and a measurable proxy. Goodhart's law states that when a measure becomes a target, it can stop being a good measure (Goodhart, 1984). Campbell's law articulates a related effect in social systems under pressure (Campbell, 1979). Recent work has clarified that Goodhart effects arise through multiple mechanisms, including selection, causal intervention, and adversarial behavior (Manheim & Garrabrant, 2019). The result is a feedback loop that actively narrows scientific horizons: recent empirical work confirms that while AI boosts individual productivity, it leads to a collective "flattening" of discovery, where outputs cluster around existing canons rather than exploring risky novelty (Hao et al., 2026).

Our model also draws from the literature on metric-driven organizations. In the "audit society" view, modern institutions increase reliance on formal indicators because indicators are legible and defensible under accountability constraints (Power, 1997; Strathern, 2000). This shift can improve consistency but can also displace the underlying purpose when indicators become the primary object of optimization (Muller, 2018).

Our framework unifies these classical sociological insights into a formal control model. While (Smaldino & McElreath, 2016) famously characterized the proliferation of bad science as a "natural selection" process driven by incentives, and (Ioannidis, 2005) identified the statistical inevitability of false findings, these works largely treated the verification constraint (human bandwidth) as a fixed constant. Our work differs by introducing AI as a dynamic variable. We argue that the "natural selection" toward Universe B is not absolute destiny, but a function of verification cost, which modern AI has the potential to fundamentally alter.

B.2. Limited Attention and Verification Bottlenecks

We treat peer review as an attention allocation process under scarcity. This view aligns with rational inattention, which models decision makers as optimizing under information processing constraints (Sims, 2003). It also aligns with multi-task incentive theory, which predicts that when agents face many tasks but are evaluated on a subset, effort shifts toward what is measured (Holmstrom & Milgrom, 1991).

B.3. Peer Review Datasets and Automation

Several datasets and models study peer review as text and as a prediction problem. PeerRead provides a public corpus of reviews and accept decisions across major venues (Kang et al., 2018). Prediction systems include interpretable decision rules for acceptance forecasting (Bao et al., 2021).

The recent wave of LLMs revived interest in automated reviewing and, more broadly, AI-assisted scholarly critique (Zou, 2024). Empirical studies have evaluated LLMs as review generators and compared their outputs with expert reviews (Zhou et al., 2024; Shin et al., 2025). Complementary work investigates how LLM-assisted feedback can enrich critique beyond simple novelty screening, for example by producing more specific, actionable feedback (Afzal et al., 2025). Other studies test whether LLM-derived scores correlate with conference or journal outcomes (Thelwall & Yaghi, 2025). Finally, structured approaches aim to formalize reviewer reasoning workflows (e.g., via causal or structured reasoning representations), which can guide where and how LLM-based interventions should assist humans (Dycke et al., 2025).

At the same time, LLMs already affect the peer review ecosystem as writers and editors. For example, a large-scale analysis estimated measurable LLM modification in review text after the release of ChatGPT (Liang et al., 2024).

Our stance differs from most "LLM-as-reviewer" work. We do not treat agreement with human review text or scores as the primary objective. We treat it as a potentially misleading target when the human process itself operates in a proxy-sovereign regime. We align more closely with work warning against over-automation of peer review (Akella et al., 2025) and with evidence that automated assistants can be gamed when judging proxy compliance (Goldberg et al., 2024).

B.4. Psychohistory and Sociophysics

Our approach parallels the "Psychohistory" concept envisioned by Asimov, which posits that while individual trajectories are unpredictable, the statistical behavior of massive populations follows deterministic laws (Asimov, 1951). In the real world, this intuition is formalized in Cliodynamics and Sociophysics. Turchin's theory of "elite overproduction" (Goldstone, 2018) offers a striking analog to our model of claim inflation (R): when the production of claims outstrips the community's verification capacity ($\mathcal{B}_{\text{eff}}$), the system enters a phase of instability similar to secular cycles in history. Furthermore, Tainter's framework on the collapse of complex societies (Tainter, 1988) explains the phenomenon of defensive proxy optimization: as a field matures, the marginal return on "complexity" (e.g., denser appendices, heavier tables) diminishes, yet the system demands more of it to maintain the illusion of progress, accelerating the drift toward Universe B. Our work translates these macro-historical dynamics into the micro-incentive structure of peer review.

C. Proofs

C.1. Proof of Theorem 4.1

Proof. We start from the score model in (2),

$$S = T + (1 - q)\Delta.$$

Assumption 3.1 states that Δ has mean zero and is independent of T.

We first compute the covariance between S and T. Using linearity of covariance, we have

$$\begin{aligned}\mathrm{Cov}(S, T) &= \mathrm{Cov}(T + (1 - q)\Delta,\, T) \\ &= \mathrm{Cov}(T, T) + (1 - q)\mathrm{Cov}(\Delta, T).\end{aligned}$$

Independence implies $\mathrm{Cov}(\Delta, T) = 0$, so

$$\mathrm{Cov}(S, T) = \mathrm{Var}(T).$$

We next compute the variance of S. Using $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$, we obtain

$$\begin{aligned}\mathrm{Var}(S) &= \mathrm{Var}(T + (1-q)\Delta) \\ &= \mathrm{Var}(T) + (1-q)^2\mathrm{Var}(\Delta) + 2(1-q)\mathrm{Cov}(T,\Delta).\end{aligned}$$

Independence again implies $\mathrm{Cov}(T,\Delta) = 0$, so

$$\mathrm{Var}(S) = \mathrm{Var}(T) + (1-q)^2\mathrm{Var}(\Delta).$$

By the definition of correlation,

$$\mathrm{Corr}(S,T) = \frac{\mathrm{Cov}(S,T)}{\sqrt{\mathrm{Var}(S)\mathrm{Var}(T)}} = \frac{\mathrm{Var}(T)}{\sqrt{\mathrm{Var}(T)\left(\mathrm{Var}(T) + (1-q)^2\mathrm{Var}(\Delta)\right)}}.$$

We cancel a factor of $\sqrt{\mathrm{Var}(T)}$ and obtain

$$\rho = \mathrm{Corr}(S,T) = \left(1 + (1-q)^2\frac{\mathrm{Var}(\Delta)}{\mathrm{Var}(T)}\right)^{-1/2}.$$

Recalling the definition $r = \mathrm{Var}(\Delta)/\mathrm{Var}(T)$ from (5) completes the proof. □

C.2. Proof of Corollary 4.2

Proof. Theorem 4.1 states that

$$\rho = \left(1 + (1-q)^2 r\right)^{-1/2}.$$

We want to characterize when $\rho \geq \rho_{\min}$.

Because all terms are nonnegative and $\rho_{\min} \in (0,1)$, we can square both sides without changing the inequality direction:

$$\left(1 + (1-q)^2 r\right)^{-1} \geq \rho_{\min}^2.$$

We invert both sides. Inversion reverses inequalities for positive numbers, so we obtain

$$1 + (1-q)^2 r \leq \rho_{\min}^{-2}.$$

We subtract one from both sides and divide by $r > 0$:

$$(1-q)^2 \leq \frac{\rho_{\min}^{-2} - 1}{r}.$$

The right-hand side is nonnegative because $\rho_{\min} \in (0,1)$. We take square roots on both sides and obtain

$$1 - q \leq \sqrt{\frac{\rho_{\min}^{-2} - 1}{r}}.$$

Finally, we rearrange to conclude that

$$q \geq 1 - \sqrt{\frac{\rho_{\min}^{-2} - 1}{r}},$$

which matches (7). □

C.3. Proof of Theorem 4.4

Proof. We consider the optimization problem

$$\max_{e\in[0,1]} U(e) \quad \text{where} \quad U(e) = f(e) + (1-q)\gamma(1-e).$$

Assumption 4.3 states that f is increasing, concave, and differentiable. The function $e \mapsto (1-q)\gamma(1-e)$ is linear in e. Therefore U is concave on $[0,1]$, as a sum of a concave function and a linear function. A concave function on a compact interval attains its maximum, and any point that satisfies the first-order optimality condition and lies in the interior is a global maximizer.

We compute the derivative of U:

$$U'(e) = f'(e) - (1-q)\gamma.$$

If an optimal solution $e^\star$ lies strictly inside $(0,1)$, then the necessary optimality condition is $U'(e^\star) = 0$. This yields the first-order condition

$$f'(e^\star) = (1-q)\gamma,$$

which is (9).

We now characterize when the optimum occurs at the boundary $e = 0$. Because U is concave, the derivative $U'(e)$ is nonincreasing in e. If $U'(0) \le 0$, then $U'(e) \le 0$ for all $e \in [0,1]$, and U is nonincreasing. In that case, $e^\star = 0$ is optimal. We compute

$$U'(0) = f'(0) - (1-q)\gamma.$$

The condition $U'(0) \le 0$ is equivalent to $(1-q)\gamma \ge f'(0)$. This proves the stated collapse condition. □

D. A Stylized Goodhart Amplification Mechanism

The main text treats $\mathrm{Var}(\Delta)$ as a primitive. In many fields, however, $\mathrm{Var}(\Delta)$ is endogenous: when incentives reward proxy performance, researchers can expand the *search budget* devoted to proxy optimization. This appendix sketches a minimal mechanism—*best-of-K selection*—showing how proxy-selection noise can grow quickly with verification pressure Λ.

D.1. Best-of-K Selection Under Proxy Optimization

Consider a proxy objective (e.g., benchmark score) that can be improved by repeated attempts such as hyperparameter sweeps, prompt sweeps, or variant sweeps. Let $X \ge 0$ denote the proxy gain from a single attempt, and let $X_1, \dots, X_K$ be i.i.d. copies of X. If only the best outcome is reported, the reported proxy gain is

$$G_K := \max\{X_1, \dots, X_K\}.$$

If the best-of-K gain does not reliably survive independent reruns, a simple stylized contribution to the proxy error is the *centered selection term*

$$\Delta_{\mathrm{sel}} := G_K - \mathbb{E}[G_K], \qquad \text{so that} \qquad \mathrm{Var}(\Delta) \;\ge\; \mathrm{Var}(\Delta_{\mathrm{sel}}) = \mathrm{Var}(G_K).$$

The key driver is how K scales. Verification pressure creates competitive pressure, and AI reduces the marginal cost per attempt; both effects can increase K. We therefore treat K as a (weakly) increasing function of Λ.

A standard order-statistics identity gives, for any CDF F of X,

$$\mathbb{P}(G_K \le x) = \mathbb{P}(X \le x)^K = F(x)^K.$$

Thus any typical scale of G_K (e.g., quantiles) increases with K.

D.2. Heavy-Tailed Proxy Gains Yield Fast Growth

Growth can be especially rapid when proxy gains are heavy-tailed: most tweaks yield tiny gains, but rare tweaks yield very large gains. We use Pareto tails as a transparent example.

Assumption D.1 (Pareto proxy gain)**.** The proxy gain X follows a Pareto distribution with parameters $x_{\min} > 0$ and $\alpha > 2$:

$$\mathbb{P}(X > x) = \left(\frac{x_{\min}}{x}\right)^\alpha, \qquad x \ge x_{\min}.$$

The condition $\alpha > 2$ ensures X has finite variance.

Proposition D.2 (Typical scale of best-of-K grows polynomially)**.** *Under Assumption D.1, the median of G_K is*

$$\mathrm{median}(G_K) = x_{\min}\big(1 - 2^{-1/K}\big)^{-1/\alpha} \;\approx\; x_{\min}\Big(\tfrac{K}{\ln 2}\Big)^{1/\alpha} \quad \textit{for large } K.$$

Consequently, the typical scale of best-of-K outcomes grows on the order of $K^{1/\alpha}$. When $\alpha > 2$, this implies that the dispersion of G_K (and thus the selection component of $\mathrm{Var}(\Delta)$) increases at least polynomially with K.

Derivation (standard order-statistics). For any $x \geq x_{\min}$, $\mathbb{P}(G_K \leq x) = F(x)^K$. The median m_K solves $F(m_K)^K = \frac{1}{2}$, hence $F(m_K) = 2^{-1/K}$. For Pareto, $F(x) = 1 - (x_{\min}/x)^\alpha$, so $1 - (x_{\min}/m_K)^\alpha = 2^{-1/K}$, yielding $m_K = x_{\min}(1 - 2^{-1/K})^{-1/\alpha}$. For large K, $2^{-1/K} = \exp(-(\ln 2)/K) \approx 1 - (\ln 2)/K$, which gives the stated approximation.

D.3. A Route to Exponential Amplification in Verification Pressure

Proposition D.2 gives polynomial growth in K. If K itself grows rapidly with verification pressure, then proxy-selection noise can grow rapidly with Λ.

Corollary D.3 (A route to exponential growth in proxy-selection noise)**.** *Assume $K(\Lambda) = K_0 \exp\big(\beta(\Lambda - 1)_+\big)$ for constants $K_0 \geq 1$ and $\beta > 0$, where $(x)_+ = \max\{x, 0\}$. Under Assumption D.1, the median (and typical scale) of $G_{K(\Lambda)}$ grows exponentially in Λ:*

$$\mathrm{median}\big(G_{K(\Lambda)}\big) \;\approx\; x_{\min}\Big(\tfrac{K_0}{\ln 2}\Big)^{1/\alpha} \exp\Big(\tfrac{\beta}{\alpha}(\Lambda - 1)_+\Big).$$

Therefore, the contribution of best-of-K selection to $\mathrm{Var}(\Delta)$ can increase exponentially with verification pressure.

This appendix is intentionally stylized. It does *not* claim that all fields exhibit Pareto proxy gains or exponential growth in K. The point is qualitative: once (i) the number of proxy-optimization attempts K becomes endogenous and increases with pressure, and (ii) selection picks the best outcome, proxy-selection noise can grow quickly. This pushes up $r = \mathrm{Var}(\Delta)/\mathrm{Var}(T)$ and can accelerate drift toward proxy sovereignty.

E. Measuring Signal Shrinkage in Practice

The main text uses $\mathrm{Var}(T)$ as a proxy for how readable the true signal is in a given venue-year. This appendix discusses measurement strategies and practical proxies. Our goal is not to define "truth" perfectly. Our goal is to approximate whether a community faces a low signal-to-noise regime.

E.1. Benchmark Headroom as A Saturation Proxy

In many machine learning subfields, progress is reported on fixed benchmarks. If a benchmark metric approaches a ceiling, then headroom shrinks. This creates two effects. First, true improvements become smaller. Second, the same absolute amount of evaluation noise becomes more influential.

A simple headroom proxy is $1 - M_t$, where M_t is the best reported metric value in year t on a bounded metric. If headroom shrinks over time, we expect the variance of improvements $\Delta M_t = M_t - M_{t-1}$ to shrink. This shrinkage is one empirical signal of declining $\mathrm{Var}(T)$.

This proxy has limitations. Benchmarks can shift, and reported best values can reflect selective reporting. For this reason, headroom is most informative when combined with verification artifacts such as reruns or standardized evaluation suites.

E.2. Variance Decomposition via Audits

We can estimate $r = \mathrm{Var}(\Delta)/\mathrm{Var}(T)$ without observing T directly by using audits as partial ground truth. Suppose an audit protocol reruns a reported experiment with controlled randomness and standardized baselines. Let Y denote the originally reported gain and let Y' denote the audited gain. A large dispersion of $Y - Y'$ indicates large proxy error variance.

If we treat Y' as a higher-fidelity measurement of the underlying effect, then the variance of Y' across papers approximates $\mathrm{Var}(T)$, while the variance of $Y - Y'$ approximates $\mathrm{Var}(\Delta)$. This approach is imperfect, but it matches the design goal of verification-first review: we want evidence that supports variance decomposition.

E.3. Long-Horizon Stability as A Truth Proxy

A central challenge is that short-horizon venue decisions cannot directly observe long-horizon truth. We can therefore use stability proxies.

One proxy is *baseline durability*. If a method remains competitive when later work introduces stronger baselines and broader sweeps, the original claim likely had higher truth content.

Another proxy is *replication survival*. If independent groups reproduce the claimed gain, the claim likely had higher truth content. Community replication datasets can support this measurement.

A third proxy is *mechanistic compression*. In some fields, a contribution becomes durable when it compresses into a reusable concept, theorem, or primitive. Citation patterns can partly capture this, although they remain social signals.

E.4. Why Saturation Interacts With Verification Pressure

Our main claim about saturation is that it amplifies the effect of verification pressure. When $\mathrm{Var}(T)$ shrinks, the same level of proxy noise produces a larger r. In that regime, even honest incremental work becomes hard to distinguish. This encourages communities to rely more on legible proxies. It also encourages researchers to invest in proxy optimization because small, fragile gains can win selection if they look like progress.

This interaction helps explain why mature engineering domains can drift toward proxy sovereignty even when they retain strong technical cultures. When true improvements become extremely fine-grained, the system needs more decisive verification per claim to maintain coupling. If bandwidth does not scale, the system drifts.

F. Citations as Proxy Signals and a Winner's Curse

We can use our framework to reason about citation counts. In many fields, citations act as a low-cost recognition signal. Citations can reflect genuine scientific value, but they can also reflect visibility, coordination effects, and social amplification. In this appendix we treat citations as another proxy score and we derive a simple *winner's curse* effect: when truth-coupling is low, papers with extreme citation scores can have modest latent scientific value.

F.1. A Stylized Citation Score Model

We let T denote latent scientific value, as in the main text. We define a centered real-valued citation score C. In practice, we can construct C from citation counts by taking a log transform and then standardizing, so C has mean zero and finite variance.

Assumption F.1 (Citation proxy decomposition)**.** We assume citations follow the same mixture structure as venue scores. Specifically,

$$C = T + (1-q)\Delta_C, \tag{11}$$

where $q \in [0,1]$ is the effective verification rate from the main text. The random variable Δ_C captures citation-specific proxy effects. We assume $\mathbb{E}[\Delta_C] = 0$ and Δ_C is independent of T.

We define the citation noise-to-signal ratio

$$r_C := \frac{\mathrm{Var}(\Delta_C)}{\mathrm{Var}(T)}. \tag{12}$$

Corollary F.2 (Citation truth-coupling)**.** *Under Assumption F.1, the correlation between citations and latent value equals*

$$\rho_C := \mathrm{Corr}(C,T) = \Big(1 + (1-q)^2 r_C\Big)^{-1/2}. \tag{13}$$

Corollary F.2 is derived directly from Theorem 4.1.

F.2. A Winner's Curse for Highly Cited Papers

Corollary F.2 shows that citations can decouple from latent value when q is small or r_C is large. We now derive a sharper statement about selection on high citation scores. We add a Gaussian assumption only to obtain a closed-form conditional expectation.

Assumption F.3 (Gaussian proxy model)**.** In addition to Assumption F.1, we assume T and Δ_C are independent Gaussian random variables with mean zero and finite variances.

Theorem F.4 (Winner's curse for citation scores)**.** *Under Assumptions F.1 and F.3, for any $c \in \mathbb{R}$ we have*

$$\mathbb{E}[T \mid C = c] = \rho_C^2\, c. \tag{14}$$

Moreover, the conditional variance equals

$$\mathrm{Var}(T \mid C = c) = \mathrm{Var}(T)\big(1 - \rho_C^2\big), \tag{15}$$

and the expected proxy inflation in the citation score equals

$$\mathbb{E}[C - T \mid C = c] = \big(1 - \rho_C^2\big)c. \tag{16}$$

Proof. We write $a := 1 - q$ for brevity. By (11), we have $C = T + a\Delta_C$. Since T and Δ_C are independent Gaussian, the pair (T, C) is jointly Gaussian.

We first compute $\mathrm{Cov}(T, C)$ and $\mathrm{Var}(C)$. Using linearity and independence,

$$\mathrm{Cov}(T, C) = \mathrm{Cov}\big(T,\, T + a\Delta_C\big) = \mathrm{Var}(T) + a\,\mathrm{Cov}(T, \Delta_C) = \mathrm{Var}(T),$$

and

$$\mathrm{Var}(C) = \mathrm{Var}\big(T + a\Delta_C\big) = \mathrm{Var}(T) + a^2\mathrm{Var}(\Delta_C) + 2a\,\mathrm{Cov}(T, \Delta_C) = \mathrm{Var}(T) + a^2\mathrm{Var}(\Delta_C).$$

Let

$$\beta := \frac{\mathrm{Cov}(T, C)}{\mathrm{Var}(C)} = \frac{\mathrm{Var}(T)}{\mathrm{Var}(C)}.$$

We define the residual random variable $U := T - \beta C$. We compute its covariance with C:

$$\mathrm{Cov}(U, C) = \mathrm{Cov}(T - \beta C,\, C) = \mathrm{Cov}(T, C) - \beta\,\mathrm{Var}(C) = 0.$$

Since (U, C) is jointly Gaussian and uncorrelated, U is independent of C.[3] This implies

$$\mathbb{E}[T \mid C] = \mathbb{E}[\beta C + U \mid C] = \beta C + \mathbb{E}[U \mid C] = \beta C + \mathbb{E}[U] = \beta C,$$

where we use $\mathbb{E}[U] = 0$ because $\mathbb{E}[T] = \mathbb{E}[C] = 0$. Conditioning on $C = c$ yields $\mathbb{E}[T \mid C = c] = \beta c$.

We now express β in terms of ρ_C. By definition,

$$\rho_C^2 = \frac{\mathrm{Cov}(T, C)^2}{\mathrm{Var}(T)\,\mathrm{Var}(C)} = \frac{\mathrm{Var}(T)^2}{\mathrm{Var}(T)\,\mathrm{Var}(C)} = \frac{\mathrm{Var}(T)}{\mathrm{Var}(C)} = \beta.$$

Substituting $\beta = \rho_C^2$ proves (14).

Next we compute the conditional variance. Since $T = \beta C + U$ with U independent of C, we have

$$\mathrm{Var}(T \mid C) = \mathrm{Var}(U \mid C) = \mathrm{Var}(U) = \mathrm{Var}(T - \beta C).$$

Expanding this variance and using $\beta = \mathrm{Cov}(T, C)/\mathrm{Var}(C)$, we obtain

$$\begin{aligned}
\mathrm{Var}(U) &= \mathrm{Var}(T) + \beta^2\mathrm{Var}(C) - 2\beta\,\mathrm{Cov}(T, C) \\
&= \mathrm{Var}(T) + \frac{\mathrm{Cov}(T, C)^2}{\mathrm{Var}(C)} - 2\frac{\mathrm{Cov}(T, C)^2}{\mathrm{Var}(C)} \\
&= \mathrm{Var}(T) - \frac{\mathrm{Cov}(T, C)^2}{\mathrm{Var}(C)} = \mathrm{Var}(T)\big(1 - \rho_C^2\big),
\end{aligned}$$

[3] For jointly Gaussian random variables, zero covariance implies independence. This is a standard fact about multivariate normal distributions.

which gives (15).

Finally, we use $C - T = a\Delta_C$ from (11). Taking conditional expectations and using (14) yields

$$\mathbb{E}[C - T \mid C = c] = c - \mathbb{E}[T \mid C = c] = \big(1 - \rho_C^2\big)c,$$

which proves (16). □

Equation (16) makes the winner's curse explicit. When ρ_C is small, an extreme citation score c can have a large expected inflation component $\mathbb{E}[C - T \mid C = c]$. At the same time, (15) shows that uncertainty about T remains large when ρ_C is small, even after we condition on $C = c$.

Corollary F.5 (Selection on the most cited paper)**.** *Consider N independent papers with i.i.d. pairs (T_i, C_i) that satisfy Assumptions F.1 and F.3. Let $i^* \in \arg\max_{i\in\{1,\dots,N\}} C_i$ and let $C_{\max} := C_{i^*}$. Then for any $c \in \mathbb{R}$,*

$$\mathbb{E}[T_{i^*} \mid C_{\max} = c] = \rho_C^2\, c. \tag{17}$$

Proof. By definition, the event $\{C_{\max} = c\}$ implies $\{C_{i^*} = c\}$ and $\{C_j \le c \text{ for all } j \ne i^*\}$. Since the papers are independent, the random pair (T_{i^*}, C_{i^*}) is independent of the collection $\{C_j\}_{j\ne i^*}$. Therefore, conditioning on $\{C_j \le c \text{ for all } j \ne i^*\}$ does not change the conditional distribution of T_{i^*} given $C_{i^*} = c$. We thus have

$$\mathbb{E}[T_{i^*} \mid C_{\max} = c] = \mathbb{E}[T_{i^*} \mid C_{i^*} = c] = \rho_C^2\, c,$$

where the last equality follows from Theorem F.4. □

Corollary F.5 shows that selecting papers by extreme citation scores does not remove proxy inflation. Equation (17) implies that the expected latent value of the most cited paper scales like ρ_C^2 times its citation score. When $\rho_C < 1$, the selected paper also satisfies $\mathbb{E}[C_{\max} - T_{i^*} \mid C_{\max} = c] = (1 - \rho_C^2)c$ by (16). This matches the qualitative phenomenon that some highly cited papers need not have correspondingly high latent scientific value.

F.3. Population-Normalized Citation Scores

To compare citation signals across fields, we need a normalization that accounts for field size and citation "volume." We keep the model as simple as possible.

Definition F.6 (Population-normalized citation score)**.** Consider a field f. We let Y_f denote the raw citation count of a paper measured over a fixed horizon. We let $\mu_f := \mathbb{E}[Y_f]$ denote the field baseline mean over the same horizon. We let $s_f > 0$ denote a scaling factor that captures population effects, such as the number of active citers or the number of papers that can generate citations in the horizon. We define the population-normalized citation score

$$C_f := \frac{Y_f - \mu_f}{s_f}. \tag{18}$$

We can now apply the same proxy decomposition to the normalized score.

Assumption F.7 (Field-specific citation proxy model)**.** For each field f, the normalized score C_f satisfies

$$C_f = T_f + (1 - q_f)\Delta_{C,f}, \tag{19}$$

where T_f denotes latent scientific value in field f. We assume $\mathbb{E}[\Delta_{C,f}] = 0$ and $\Delta_{C,f}$ is independent of T_f.

We define $r_{C,f} := \mathrm{Var}(\Delta_{C,f})/\mathrm{Var}(T_f)$ and $\rho_{C,f} := \mathrm{Corr}(C_f, T_f)$. By Corollary F.2, we have

$$\rho_{C,f} = \big(1 + (1 - q_f)^2 r_{C,f}\big)^{-1/2}.$$

F.4. Expected Scientific Value Given Raw Citations

We can translate the theories back to raw citation units.

Assumption F.8 (Field-specific Gaussian proxy model)**.** In addition to Assumption F.7, we assume T_f and $\Delta_{C,f}$ are independent Gaussian random variables with mean zero and finite variances.

Theorem F.9 (Conditional value in raw citation units)**.** *Under Assumptions F.7 and F.8, for any field f and any $y \in \mathbb{R}$ we have*

$$\mathbb{E}[T_f \mid Y_f = y] = \rho^2_{C,f} \, \frac{y - \mu_f}{s_f}. \tag{20}$$

Proof. We condition on $Y_f = y$. By Definition F.6, this is equivalent to conditioning on $C_f = (y - \mu_f)/s_f$.

By Assumption F.8, the pair (T_f, C_f) is jointly Gaussian. We can therefore apply Theorem F.4 to the pair (T_f, C_f), which yields

$$\mathbb{E}[T_f \mid C_f = c] = \rho^2_{C,f} \, c \quad \text{for all } c \in \mathbb{R}.$$

We substitute $c = (y - \mu_f)/s_f$ and we obtain (20). □

Equation (20) highlights two distinct sources of cross-field incomparability. First, the coupling factor $\rho^2_{C,f}$ can vary across fields. Second, the population scale s_f controls how many raw citations correspond to one unit of the normalized score.

F.5. A Population-Normalized Citation Exchange Rate

We now define a simple exchange rate between citations in two fields.

Definition F.10 (Citation exchange rate)**.** Consider two fields A and B. We define the marginal expected value per raw citation as

$$\kappa_f := \frac{\partial}{\partial y} \mathbb{E}[T_f \mid Y_f = y].$$

We define the exchange rate from A to B as the number $\mathcal{E}_{A \to B}$ such that one additional citation in field B has the same marginal expected value as $\mathcal{E}_{A \to B}$ additional citations in field A.

Theorem F.11 (Closed form exchange rate)**.** *Under the assumptions of Theorem F.9, the marginal expected value per citation is constant in y and equals*

$$\kappa_f = \frac{\rho^2_{C,f}}{s_f}. \tag{21}$$

Consequently, the exchange rate from A to B equals

$$\mathcal{E}_{A \to B} = \frac{\kappa_B}{\kappa_A} = \frac{s_A}{s_B} \left(\frac{\rho_{C,B}}{\rho_{C,A}} \right)^2. \tag{22}$$

Proof. We start from (20), which implies

$$\mathbb{E}[T_f \mid Y_f = y] = \rho^2_{C,f} \, \frac{y - \mu_f}{s_f}.$$

Differentiating both sides with respect to y yields

$$\kappa_f = \frac{\partial}{\partial y} \mathbb{E}[T_f \mid Y_f = y] = \frac{\rho^2_{C,f}}{s_f},$$

which proves (21).

By Definition F.10, $\mathcal{E}_{A \to B}$ satisfies

$$\kappa_A \, \mathcal{E}_{A \to B} = \kappa_B.$$

We solve for $\mathcal{E}_{A\to B}$ and we substitute (21) for A and B. This yields

$$\mathcal{E}_{A\to B} = \frac{\rho_{C,B}^2/s_B}{\rho_{C,A}^2/s_A} = \frac{s_A}{s_B}\left(\frac{\rho_{C,B}}{\rho_{C,A}}\right)^2,$$

which is (22). □

Equation (22) separates population effects from truth-coupling effects. The ratio s_A/s_B captures how raw citation volume differs across the two fields. The ratio $(\rho_{C,B}/\rho_{C,A})^2$ captures how strongly citations track latent value in each field.

F.6. An Illustrative Comparison

We close with a simple parameterization that illustrates how large cross-field exchange rates can be. We do not claim that the numbers below match any particular discipline. We instead choose them to represent two qualitative regimes, and we choose them so that each knob in the model has a clear interpretation.

The key quantity is

$$K_f = (1-q_f)^2 r_{C,f} = (1-q_f)^2\,\frac{\mathrm{Var}(\Delta_{C,f})}{\mathrm{Var}(T_f)}.$$

A field can have a large K_f for three distinct reasons. Verification can be rare, so q_f is small. Signal can shrink, so $\mathrm{Var}(T_f)$ is small. Proxy amplification can grow, so $\mathrm{Var}(\Delta_{C,f})$ is large. Any of these effects can push the field into a low-coupling regime.

Consider two fields A and B. We first set verification pressure. Under our queue model we have $q_f = \min\{1, 1/\Lambda_f\}$. We set $\Lambda_A = 10$ and $\Lambda_B = 2$, so $q_A = 0.1$ and $q_B = 0.5$.

We next set the proxy-to-truth ratio $r_{C,f}$. We set $r_{C,A} = 100$ and $r_{C,B} = 5$. These are not estimates. We choose them to reflect two levels of noise dominance. The choice $r_{C,A} = 100$ means that proxy variance is two orders of magnitude larger than signal variance. This can arise from two moderate effects, namely signal shrinkage by a factor of 10 together with proxy amplification by a factor of 10. The choice $r_{C,B} = 5$ means that proxy variance is only a few times larger than signal variance, which represents a milder proxy regime.

For additional intuition, we can view $r_{C,f}$ as an effective number of independent proxy channels. If we write

$$\Delta_{C,f} = \sum_{j=1}^{m_f} Z_{f,j},$$

where the variables $Z_{f,1},\ldots,Z_{f,m_f}$ are independent, have mean zero, are independent of T_f, and satisfy $\mathrm{Var}(Z_{f,j}) = \mathrm{Var}(T_f)$, then

$$\mathrm{Var}(\Delta_{C,f}) = \mathrm{Var}\Big(\sum_{j=1}^{m_f} Z_{f,j}\Big) = \sum_{j=1}^{m_f}\mathrm{Var}(Z_{f,j}) = m_f\,\mathrm{Var}(T_f),$$

so $r_{C,f} = m_f$. Under this interpretation, $r_{C,A} = 100$ corresponds to $m_A = 100$ effective proxy channels, while $r_{C,B} = 5$ corresponds to $m_B = 5$.

Given q_A, q_B and $r_{C,A}, r_{C,B}$, we obtain

$$K_A = (1-q_A)^2 r_{C,A} = 0.9^2\cdot 100 = 81,$$
$$K_B = (1-q_B)^2 r_{C,B} = 0.5^2\cdot 5 = \frac{5}{4}.$$

Therefore

$$\rho_{C,A} = \big(1+K_A\big)^{-1/2} = \frac{1}{\sqrt{82}} \approx 0.11,$$
$$\rho_{C,B} = \big(1+K_B\big)^{-1/2} = \frac{2}{3} \approx 0.67.$$

Finally, we set the population normalization. We take s_f to be proportional to the number of active citers in field f. If field A has five times as many active citers as field B, then $s_A/s_B = 5$.

By Theorem F.11, one additional citation in field B has the same marginal expected value as

$$\begin{aligned}\mathcal{E}_{A\to B} &= \frac{s_A}{s_B}\left(\frac{\rho_{C,B}}{\rho_{C,A}}\right)^2 \\ &= \frac{s_A}{s_B}\,\frac{1+K_A}{1+K_B} = 5\cdot\frac{82}{9/4} = \frac{1640}{9} \approx 182\end{aligned}$$

additional citations in field A. This example is meant to build intuition. It shows that large exchange rates arise whenever one field operates in a low-coupling regime with K_f much larger than 1, while another field operates in a moderate-coupling regime with K_f comparable to 1.

G. Empirical Agenda

Our theory introduces quantities that we can estimate in real venues. This section sketches a measurement program. We emphasize feasibility over perfection.

G.1. Estimating Verification Pressure Λ

Equation (3) defines $\Lambda = RC_{\text{eff}}/\mathcal{B}_{\text{eff}}$. We can estimate each component.

We can estimate R from submission counts or preprint inflow. OpenReview provides submission volumes for many conferences, and arXiv provides category-level inflows. We can refine R by counting claims rather than papers, for example by extracting the number of headline results, benchmarks, or theorem statements.

We can estimate $\mathcal{B}_{\text{eff}}$ by counting reviewer assignments and by estimating available review time. A coarse estimator multiplies the number of reviewers by an assumed hours-per-review. A better estimator uses self-reported time, rebuttal length, or meta-review depth as proxies for attention. Artifact evaluation programs provide another signal because they create explicit verification tasks.

We can estimate C_{eff} using time-to-verify studies. For computational papers, we can log the compute and engineering effort needed to reproduce a result from provided artifacts. For theoretical work, we can log the time needed to check proofs or to find missing assumptions. We can also decompose $C_{\text{eff}} = c/\kappa$ by measuring how often evidence resolves disputes. For example, we can measure how often an artifact rerun changes a decision. This is a direct proxy for κ.

G.2. Estimating Signal Shrinkage and r

We defined $r = \mathrm{Var}(\Delta)/\mathrm{Var}(T)$ in (5). Direct truth measurement is hard, but we can estimate components.

We can estimate $\mathrm{Var}(\Delta)$ using instability and sensitivity metrics. In machine learning, seed variance, hyperparameter sensitivity, and data leakage are measurable. In experimental sciences, batch effects and measurement error provide analogs. We can also estimate proxy noise by comparing claimed improvements with improvements that survive a standardized audit.

We can approximate $\mathrm{Var}(T)$ using long-horizon outcomes. One approach uses replication success. Another uses post-publication robustness, such as whether a method remains competitive under later, stronger baselines. A third uses "disruption" or citation topology metrics, although these remain indirect.

Saturation should show up as a decline in effect sizes on mature benchmarks. For example, if a benchmark metric approaches a ceiling, then the year-to-year spread of improvements shrinks. This reduces $\mathrm{Var}(T)$ and increases r even if proxy noise stays constant.

G.3. Testing the Phase Transition Hypothesis

The model predicts that proxy sovereignty becomes likely when Λ is large and r is large. We can test this by comparing venues and subfields.

A practical strategy is to assemble a dataset of papers with artifacts, benchmark claims, and reproducibility outcomes. We then measure proxies for q and for r and test whether these predict later instability, retractions, or rapid obsolescence. We

can also test whether interventions that reduce C_{eff} or increase $\mathcal{B}_{\text{eff}}$ increase stability.

LLM-mediated writing creates a natural experiment. If LLMs increase R without increasing $\mathcal{B}_{\text{eff}}$, the theory predicts that truth-coupling should decline unless the community also scales verification-first mechanisms. Recent measurements of AI-modified review text show that large-scale shifts already occur (Liang et al., 2024). We can build on such measurements to study downstream effects on coupling and on robustness.

G.4. What Would Falsify Our Model

The theory would be wrong or incomplete if we observe sustained high truth-coupling in regimes with high Λ and high r. Such cases would imply that communities found additional coupling mechanisms not captured in our abstraction. Examples could include strong informal replication norms, industry testbeds, or tight proof cultures that effectively increase κ. These cases would be scientifically valuable because they would suggest new design patterns for verification-first peer review.

H. Case Studies: Mature Engineering Domains as Extreme Regimes

This appendix uses the $(R, C_{\text{eff}}, \mathcal{B}_{\text{eff}})$ lens to interpret several mature engineering domains. These examples are not meant as judgments about individuals. They illustrate how a field's technical structure can create proxy sovereignty even when researchers act rationally.

H.1. Wireless Communications and "The Simulator as the World"

Wireless communications provides an instructive extreme case. The real-world object is a complex physical and socio-technical system. To validate a claim about a new protocol at scale, we often need expensive testbeds, custom hardware, and deployment conditions. This makes c large. At the same time, real environments are noisy and context-dependent, which can make κ smaller than one. As a result, $C_{\text{eff}} = c/\kappa$ can be very large.

The community therefore uses simulation proxies. Simulations reduce c, but they can also reduce κ if the simulator does not faithfully capture the deployment regime. When the proxy becomes the evaluation standard, the community enters a proxy-sovereign loop. Papers optimize for the simulator rather than for the real system. As the domain matures and approaches physical limits, true improvements shrink. This reduces $\mathrm{Var}(T)$ and increases r, which further encourages proxy reliance. In this sense, communications is an extreme realization of the phase diagram in Figure 1.

H.2. Autonomous Driving and Safety-Critical Claims

Autonomous driving has high verification friction for a different reason. Claims often concern rare events. Rare-event validation requires large-scale testing or strong causal models. Both increase c. Because accidents are sparse, evidence can remain ambiguous, which reduces κ. This raises C_{eff}.

In response, the community relies on proxy benchmarks and curated datasets. This can be useful, but it also creates strong incentives to overfit to proxy distributions. A verification-first review process would invest in stress tests and distribution shift audits that approximate the rare-event regime. This directly targets κ and reduces C_{eff}.

H.3. Reinforcement Learning and Sensitivity-Driven Proxy Noise

In reinforcement learning, results often exhibit large variance across random seeds and implementation details. This increases $\mathrm{Var}(\Delta)$ even when the underlying idea is sound. If reviewers lack bandwidth to run reruns, q is small and the system selects based on noisy maxima, which increases r. A verification-first intervention is straightforward in this domain: require seed sweeps and publish dispersion. This reduces $\mathrm{Var}(\Delta)$ without demanding new theory.

H.4. Materials Science and Long Experimental Cycles

In experimental materials research, verification can require long fabrication cycles and specialized equipment. This increases c and also creates barriers to independent replication. The result is high verification pressure even at moderate claim rates. If publication incentives push R upward, proxy sovereignty can emerge through the selection of results that appear strong in limited experimental settings. Here, verification-first design may rely more on preregistration, shared protocols, and multi-site replication consortia. These mechanisms increase $\mathcal{B}_{\text{eff}}$ and κ.

H.5. A Unifying Interpretation

Across these domains, we observe the same structure. When verification is expensive and true gains shrink, communities adopt low-cost proxies to keep the system running. Once proxies decide careers, optimization pressure shifts toward proxies. This is a stable equilibrium for individuals, but it can reduce truth-coupling at the system level. AI can either accelerate the drift, by scaling proxy optimization and claim production, or help resist it, by scaling decisive evidence.

I. Limitations

We proposed a phase transition view of scientific production driven by verification pressure and signal shrinkage. This view yields clear design implications for AI-assisted review, but it also has limitations.

I.1. Not all proxy use is bad

Proxy signals are unavoidable. Even in truth-sovereign disciplines, humans rely on heuristics to allocate attention. Our claim is narrower. Proxy sovereignty becomes harmful when proxy optimization dominates and when truth-coupling becomes low. In that regime, Goodhart drift becomes likely and Theorem 4.4 predicts that honest truth effort becomes privately irrational.

I.2. Equity and access

Verification-first mechanisms can increase burdens on authors. Compute-heavy audits may disadvantage low-resource groups. We therefore need verification artifacts that are cheap and diagnostic. We also need shared infrastructure, standardized audit suites, and community datasets. These investments can increase $\mathcal{B}_{\text{eff}}$ without requiring each author to own large compute.

I.3. The risk of verification theater

A warning from the audit literature is that verification can itself become ritual (Power, 1997). If communities treat checklists as proxies, they can recreate proxy sovereignty in a new form. Verification-first AI must keep evidence grounded. A smoke test that runs and produces logs increases fidelity. A checklist that asks for a justification paragraph can be gamed, as the NeurIPS 2024 experiment suggests (Goldberg et al., 2024).

I.4. Modeling limitations

Our model is intentionally simple. We used a scalar truth variable and a scalar score. Real peer review involves multi-dimensional judgments and also serves prioritization and feedback, in addition to social dynamics. We also treated verification as random service with a fixed bandwidth. In practice, committees route attention strategically and fields differ in their social structure. We see our model as a minimal backbone that clarifies which quantities matter and how they interact. Future work can extend it with network effects, strategic reviewer behavior, and heterogeneous paper types.