



cs.AI / cs.LG / econ.TH

THE PLAUSIBILITY TRAP: ANTI-INFORMATIVE REVIEW BEYOND AN UNLOCATABLE FRONTIER, AND A PRE-REGISTERED NULL

SERIORA RESEARCH PREPRINT

Lionel Arce
Seriora Research
larce@seriora.ai

Buenos Aires, Argentina, 25 October 2026

Abstract

When AI systems produce scientific research that their reviewers cannot fully verify, the reviewers often treat plausibility as evidence of truth. We ask when this habit makes acceptance anti-informative, so that accepted claims are less likely to be true than rejected ones. In a signal-detection model, true claims lose plausibility with their distance from the reviewer's frontier, while a capable generator shapes false claims to look plausible. Discriminability then changes sign at a frontier $d^*(c)$ that never moves outward as generator capability c rises. Beyond d^* , every acceptance rule that is increasing in plausibility has likelihood ratio below one, and auditing the least plausible accepted claims first is the wrong order. The information is not destroyed. A reviewer who knows which side of d^* a claim lies on can always recover a likelihood ratio above one. The trap is therefore a self-location failure. For a Bayesian reviewer who pools claims from both sides, the beyond-frontier slice is inverted once the sub-frontier share exceeds a threshold $\bar{w}$. We prove that this crossing is unique and strict under an explicit sufficient condition, and leave uniqueness open outside it. We then ran a pre-registered test of the mechanism on public per-instance outputs of 15 LLM verifiers. The prediction was falsified on the primary domain. In open-answer mathematics, strong generators did lower verifier specificity and sensitivity did fall with difficulty, but the likelihood ratio stayed above one in all 135 cells (minimum 1.15). An apparent inversion on 10-option knowledge questions occurs only on items the verifier never solves and is stronger for weak generators, the opposite of the mechanism's signature. The theory stands as a conditional result. Whether current systems operate beyond d^* is open.

Keywords research produced by AI systems · peer review · LLM verifiers · signal detection · scalable oversight · pre-registration

1 Introduction

AI systems now produce research ideas, proofs and papers, including on questions their reviewers cannot fully check, and other models or people decide which of these to accept. Most such reviewers read a claim and ask whether it looks right. Plausibility is a sensible proxy when the reviewer knows the territory. It is less clear

that it remains one when the claim lies far from what the reviewer knows and the author is good at making errors look right.

Several recent findings point in this direction. LLM verifiers detect the errors of strong generators much less often than those of weak ones. In mathematics, the Qwen2.5-72B verifier’s true-negative rate falls from 0.68 on solutions from Llama-3.1-8B to 0.17 on solutions from Qwen3-32B [Zhou et al., 2025]. Fabricated papers written by an AI research agent reach acceptance rates up to 82.0% under LLM review [Jiang et al., 2025]. Rhetorical framing and rephrasing shift LLM review scores, and adversarial rephrasing raises them, with content held fixed [Li et al., 2026b,a]. In an observational study of REF abstracts, explicit novelty claims are associated with higher ChatGPT quality scores [Kousha and Thelwall, 2026]. Human panels give lower scores to more novel proposals [Boudreau et al., 2016]. Research ideas generated by LLMs that looked more novel than expert ideas lose that advantage once executed [Si et al., 2025].

This paper asks when these effects add up to an inversion. We call a review process anti-informative when acceptance carries evidence against truth, so that the likelihood ratio of acceptance is below one. We make three contributions.

1. A sign-change result (Theorem 1). In a Gaussian signal-detection model with fidelity decay and generator shaping, discriminability d' changes sign at a computable frontier $d^*(c)$. Beyond it, every plausibility-increasing rule is anti-informative, and the frontier moves inward as generators improve.
2. A self-location characterisation (Theorem 1 parts (c) and (d), Proposition 5, Lemma 6). The inversion is never forced by the signal. A reviewer who knows which side of d^* it is on can always do better than chance. A Bayesian reviewer who cannot tell, and pools both sides, inverts the beyond-frontier slice once the sub-frontier share of claims exceeds a case-dependent threshold $\bar{w}$.
3. A pre-registered test with a negative result (Section 6). We froze a component-wise test of the mechanism before downloading outcome data from a public release of 15 LLM verifiers [Zhou et al., 2025]. Two components held. The crossing did not occur in open-answer mathematics, so the signed prediction is falsified there.

The negative result shapes how the theory should be read. The theorems are conditional statements under assumptions A1 to A6. They say what happens beyond d^* . They do not say that current systems are beyond it, and our test found no evidence that they are on open-answer mathematics. We report the falsification as such and do not re-scope the prediction to fit the data.

2 Model

A generator of capability c emits a claim at novelty $d \geq 0$, the claim’s distance from the reviewer’s frontier. The claim is true ($T = 1$) with prior probability $q(d) \in (0, 1)$ and false ($T = 0$) otherwise. The reviewer observes a scalar plausibility signal x and accepts or rejects.

Assumption A1 (Single sender). One generator and one reviewer. There is no adversarial second agent and no external witness such as replication or a formal check.

Assumption A2 (Plausibility as evidence). The reviewer’s rule is increasing in x . It accepts if and only if $x \geq \tau$, with τ finite, fixed and not re-optimised against the generator.

Assumption A3 (Fidelity decay). True claims at novelty d have $x \sim N(\mu_T(d), \sigma^2)$ with $\mu_T(d) = \mu_0 - \lambda d$ and $\lambda > 0$.

Assumption A4 (Shaping with bounded-away slope). False claims have $x \sim N(\mu_F(c, d), \sigma^2)$. The mean μ_F is nondecreasing in c with ceiling $\bar{\mu}_F$, and $\partial\mu_F/\partial d \geq -\lambda + \varepsilon$ for some $\varepsilon > 0$.

Assumption A5 (Prior). The prior $q(d)$ is arbitrary in $(0, 1)$.

Assumption A6 (Equal variance). True and false claims share the variance σ^2 . This gives the monotone likelihood ratio property in x .

A3 encodes the observation that reviewers rate more novel proposals lower [Boudreau et al., 2016] and favour ideas that resemble their own [Bao et al., 2026]. Here d is the claim’s distance from verified knowledge, as in proposal novelty, not its distance from the evaluator’s own expertise, for which Boudreau et al. [2016] report the opposite sign and which we do not use. A4 encodes the observation that capable generators make errors that look right [Zhou et al., 2025, Li et al., 2026b,a]. A2 matches the fixed criterion shift reported for multimodal judges [Cai et al., 2026]. The weaker condition $\partial\mu_F/\partial d > -\lambda$ does not suffice for A4. A gap of $0.5 + 0.5e^{-d}$ satisfies it and never closes. We set $\sigma = 1$ without loss of generality. Table 1 collects the notation.

Table 1: Notation.

Symbol	Meaning
d	novelty, the claim's distance from the reviewer's frontier
c	generator capability
x	plausibility signal observed by the reviewer
$\mu_T(d), \mu_F(c, d)$	mean plausibility of true and of false claims
λ, ε	rate of fidelity decay, and the margin by which shaping cannot keep pace
$q(d), \theta$	prior probability of truth, and a Bayesian reviewer's posterior threshold
$d', d^*(c)$	discriminability, and the novelty at which it changes sign
LR	likelihood ratio of acceptance, $P(\text{accept} \mid T = 1)/P(\text{accept} \mid T = 0)$
$w, \bar{w}, w_{\text{low}}$	sub-frontier share, and the thresholds of Proposition 5

The object of study is the discriminability and the likelihood ratio of a rule R ,

$$d'(d, c) = \frac{\mu_T(d) - \mu_F(c, d)}{\sigma}, \quad \text{LR}_R(d, c) = \frac{P(\text{accept} \mid T = 1)}{P(\text{accept} \mid T = 0)}. \quad (1)$$

Acceptance is anti-informative when $\text{LR}_R < 1$. By Bayes' rule this holds exactly when $P(T = 1 \mid \text{accept}) < q(d)$.

3 The frontier and the self-location failure

Theorem 1 (The trap is a self-location failure). *Assume A1 and A3 to A6.*

(a) *There is $d^*(c)$ with $d'(d, c) < 0$ for all $d > d^*(c)$. If μ_F does not depend on d , then*

$$d^*(c) = \frac{\mu_0 - \mu_F(c)}{\lambda},$$

which is nonincreasing in c and bounded below by $(\mu_0 - \bar{\mu}_F)/\lambda$. If $\mu_F(c) \geq \mu_0$ the reversal holds at every $d > 0$. Under general A4, $d^*(c) \leq (\mu_0 - \mu_F(c, 0))/\varepsilon$.

(b) *Under A2, for $d > d^*(c)$ every increasing rule has $\text{LR} < 1$, the area under the ROC curve is $\Phi(d'/\sqrt{2}) < 1/2$, and $P(T = 1 \mid \text{accept}) < q(d) < P(T = 1 \mid \text{reject})$.*

(c) *For $d \neq d^*(c)$, a reviewer who knows the sign of d' has a rule with $\text{LR} > 1$. The rule is increasing below d^* and decreasing above it. At $d = d^*(c)$ every rule has $\text{LR} = 1$.*

(d) *Suppose the reviewer sees x but not d , and claims come from a sub-frontier slice $d_{\text{lo}} < d^*$ with share w and a beyond-frontier slice $d_{\text{hi}} > d^*$ with share $1 - w$. Then (d1) every increasing rule inverts the beyond-frontier slice, and (d2) the pooled Bayes rule behaves as in Proposition 5.*

Proof sketch. Part (a) follows because the gap $\mu_T - \mu_F$ falls at rate at least ε and is linear when μ_F is constant in d . For (b), with $\Delta = \mu_T - \mu_F < 0$ the acceptance probabilities satisfy

$$\text{LR}(\tau) = \frac{1 - \Phi(\tau - \mu_T)}{1 - \Phi(\tau - \mu_F)} < 1$$

for every finite τ . Rejection then has likelihood ratio above one, which gives the posterior ordering. For (c), the reversed rule “accept if and only if $x \leq \tau$ ” has $\text{LR} = \Phi(\tau - \mu_T)/\Phi(\tau - \mu_F) > 1$. At $\Delta = 0$ the two distributions coincide. Part (d1) is (b) applied to the slice. Section 7.1 gives the full proof in steps. $\square$

Part (c) is the reason we call the trap a self-location failure. Beyond the frontier the signal still separates true from false claims. It separates them in the wrong direction for a reviewer who reads plausibility as evidence. A reviewer who could locate a claim relative to d^* would simply flip its rule. The difficulty is that d^* depends on the generator's capability and on the reviewer's own frontier, and neither is observed claim by claim.

Proposition 2 (Prior drop is insufficient). *If $d' \geq 0$, every increasing rule gives $P(T = 1 \mid \text{accept}) \geq P(T = 1 \mid \text{reject})$ for every $q(d)$.*

Proposition 3 (An informed Bayesian is not trapped). *A reviewer who knows d and c and accepts if and only if $P(T = 1 \mid x) \geq \theta$ accepts a lower set of x when $d > d^*$. It keeps $P(T = 1 \mid \text{accept}) \geq \theta \geq P(T = 1 \mid \text{reject})$.*

Proposition 4 (No increasing rule helps). *For $d > d^*(c)$ no increasing rule on x has $\text{LR} > 1$. A reviewer bound to A2 can restore $\text{LR} > 1$ only with a signal outside x , or with information that locates d relative to d^* .*

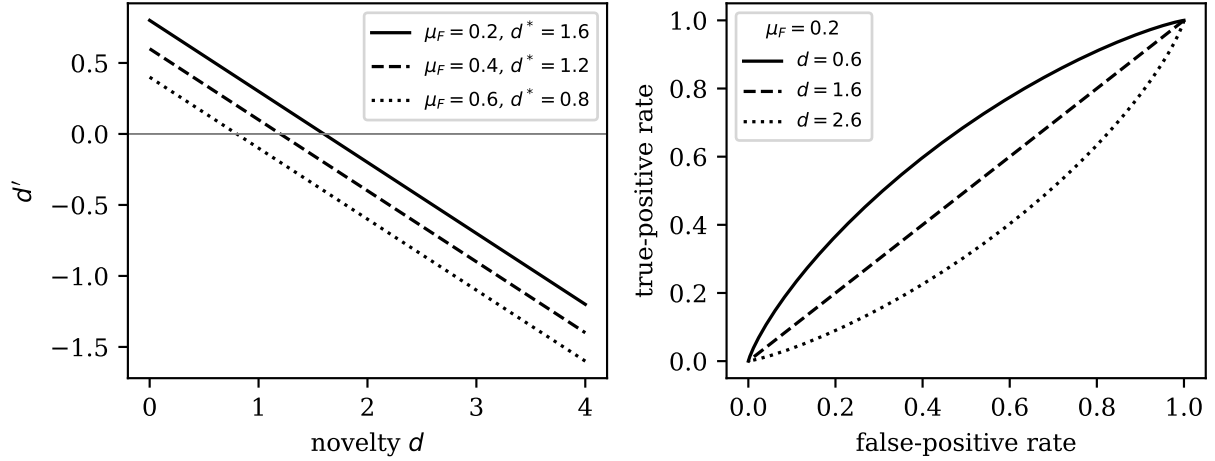


Figure 1: Left, discriminability d' against novelty for three levels of generator shaping μ_F , with $\mu_0 = 1$ and $\lambda = 0.5$. Right, ROC curves of the threshold family at $\mu_F = 0.2$ below ($d = 0.6$), at ($d = 1.6$) and beyond ($d = 2.6$) the frontier.

Proposition 2 rules out a tempting alternative explanation. A falling base rate of truth at high novelty lowers the posterior of accepted claims but never orders accepted below rejected ones. Proposition 3 shows that the inversion is not a property of Bayesian updating. The informed Bayesian reverses its ranking and remains informative. It does not pool. Pooling, with $LR = 1$ for every rule, happens only at $d = d^*$. Proposition 4 restates (b) as a design constraint. Debate and other outside signals fall outside A1 and are discussed in Section 9.

Geometry of the frontier. Figure 1 shows the two objects of Theorem 1 in the worked parameterisation $\mu_0 = 1$ and $\lambda = 0.5$. On the left, d' falls linearly in d and crosses zero at $d^* = (\mu_0 - \mu_F)/\lambda$. Raising μ_F from 0.2 to 0.6, which A4 attributes to a more capable generator, moves the crossing from 1.6 to 0.8. A generator does not need to produce more errors for the reviewer's frontier to recede. It only needs to make its errors look more like the reviewer's idea of a correct claim. On the right, the ROC curve of the threshold family lies above the diagonal below the frontier, on it at the frontier, and below it beyond. Below the diagonal every threshold is worse than a coin, and the mirror-image rule is better than a coin by the same margin.

Why the shape assumptions matter. Three assumptions carry most of the weight, and each fails in an identifiable way. Without the bounded-away slope in A4 the gap can shrink forever without closing, and no frontier exists. Without equal variances (A6) the likelihood ratio of a threshold is not monotone in τ , so a reviewer could be anti-informative at some thresholds and informative at others. Without A2 the reviewer may already use a decreasing rule, and the inversion becomes a question of what it knows rather than how it reads plausibility. A3 is the weakest empirically. Its linear form is a convenience, and the result needs only that μ_T eventually falls below μ_F .

4 A pooled Bayesian reviewer

Part (d) of Theorem 1 asks what an optimal reviewer does when it cannot locate a claim. Let the slice means be $m_{lo} = \mu_T(d_{lo}) > \mu_F > m_{hi} = \mu_T(d_{hi})$ and write $a = m_{lo} - \mu_F > 0 > b = m_{hi} - \mu_F$. The likelihood ratios of each slice against false claims are

$$s(x) = \frac{\varphi(x - m_{lo})}{\varphi(x - \mu_F)} = Ae^{ax}, \quad r(x) = \frac{\varphi(x - m_{hi})}{\varphi(x - \mu_F)} = Be^{bx},$$

with $A = e^{-(m_{lo}^2 - \mu_F^2)/2}$ and $B = e^{-(m_{hi}^2 - \mu_F^2)/2}$. With a common prior q and posterior threshold θ , the pooled Bayes rule accepts if and only if

$$g_w(x) = w s(x) + (1 - w) r(x) \geq k, \quad k = \frac{\theta(1 - q)}{(1 - \theta)q}. \quad (2)$$

We study the slice likelihood ratio $\text{LR}(w)$ of the beyond-frontier slice under this rule.

Proposition 5 (Two thresholds). *There exist $0 < w_{\text{low}} \leq \bar{w} < 1$ such that the rule (2) inverts the beyond-frontier slice, with $\text{LR}(w) < 1$, for every $w > \bar{w}$, and does not invert it for any $w < w_{\text{low}}$. If $\min_x g_w \geq k$ the rule accepts every claim and $\text{LR} = 1$. The threshold $\bar{w}$ depends on (θ, q) and the slice separation, and can be arbitrarily close to 1.*

The proof is in Section 7.2. Its main steps are short. g_w is strictly convex, so the acceptance set is two-tailed, $(-\infty, x_1] \cup [x_2, \infty)$. As $w \rightarrow 1$ the left tail vanishes and $\text{LR}(w)$ tends to a right-tail ratio below one. As $w \rightarrow 0$ the right tail vanishes and $\text{LR}(w)$ tends to a left-tail ratio above one. Continuity and the intermediate value theorem give the thresholds.

Accept-all bands are real. Because $w \mapsto \min_x g_w(x)$ is concave and tends to zero at both ends, the set of w with $\min_x g_w \geq k$ is a single closed interval inside $(0, 1)$. On it $\text{LR} = 1$. We do not count entering or leaving such a band as a crossing. A crossing is a sign change of $\text{LR} - 1$ between values different from one.

Uniqueness of the crossing holds under a simple condition. Let $x_0 = (m_{\text{lo}} + m_{\text{hi}})/2$ be the unique point where $s = r$, and let $g_0 = r(x_0)$. Then $g_w(x_0) = g_0$ for every w .

Lemma 6 (Unique strict crossing). *If $k > g_0$, then $\text{LR}(w) - 1$ has exactly one zero $\bar{w}$ in $(0, 1)$. The crossing is strict and downward, $\text{LR} > 1$ on $(0, \bar{w})$, $\text{LR} < 1$ on $(\bar{w}, 1)$, $w_{\text{low}} = \bar{w}$, and there is no accept-all band.*

Proof. We give the proof in steps.

Step 1 (smooth boundaries). When $\min g_w < k$ the boundaries satisfy $g_w(x_i) = k$ with $g'_w(x_1) < 0 < g'_w(x_2)$. By the implicit function theorem they are C^1 in w , with

$$x'_i(w) = -\frac{s(x_i) - r(x_i)}{g'_w(x_i)}.$$

Step 2 (derivative of the gap). Let $\delta(w) = P_{\text{hi}}(\text{accept}) - P_F(\text{accept})$, so that $\text{LR} > 1$ if and only if $\delta > 0$. Differentiating the two tails gives

$$\delta'(w) = \varphi_F(x_1)(r(x_1) - 1)x'_1 - \varphi_F(x_2)(r(x_2) - 1)x'_2.$$

Step 3 (sign of r at a zero). Both densities integrate to one, so $\delta = 0$ if and only if $\int_{x_1}^{x_2} \varphi_F(r - 1) dx = 0$. Since r is strictly decreasing, this forces $r(x_1) > 1 > r(x_2)$.

Step 4 (direction). If $k > g_0$ then $x_0 \in (x_1, x_2)$. So $s(x_1) < r(x_1)$ and $g'_w(x_1) < 0$, which give $x'_1 < 0$. Likewise $s(x_2) > r(x_2)$ and $g'_w(x_2) > 0$, which give $x'_2 < 0$. With Step 3, the first term of δ' is negative and the second is subtracted and positive. Hence $\delta' < 0$ at every zero.

Step 5 (no band). $\min_x g_w \leq g_w(x_0) = g_0 < k$ for every w .

Step 6 (uniqueness). δ is C^1 on $(0, 1)$ with $\delta(0^+) > 0$ and $\delta(1^-) < 0$. Let z_1 be a zero and $z_2 = \inf\{z > z_1 : \delta(z) = 0\}$. Then $\delta < 0$ on (z_1, z_2) , so $\delta'(z_2) \geq 0$ if $z_2 < 1$, which contradicts Step 4. $\square$

Remark 1 (What is open). When $k \leq g_0$ the point x_0 lies in an acceptance tail, x'_1 and x'_2 share a sign, and Steps 1 to 3 do not sign δ' . $\text{LR}(w)$ itself is not monotone, so monotonicity cannot replace Step 4. Numerically, two random-search runs over $k \in [e^{-10}, e^{10}]$, $|a|, |b| \in [0.01, 8]$ and $\mu_F \in [-4, 4]$ (34,000 and 26,400 sets, the second ending with $k \leq g_0$, 453-point w grid, each stalled on one set) and a directed search with $k \leq g_0$ (293,901 sets, random then 112 differential-evolution restarts, 247-point grid, top five rechecked at 60 digits) found no set with more than one crossing, and the directed search found no upward one. This is numerical evidence, not a proof, and crossing pairs narrower than the grid could be missed. All 793 zeros located in a separate sample were downward, including 479 outside the condition of Lemma 6. We state uniqueness only under $k > g_0$.

Figure 2 shows the worked example $\mu_0 = 1$, $\lambda = 0.5$, $\mu_F = 0.2$, $q = 0.3$, $\theta = 0.35$ and $d = d^* \pm 1$. Here $d^* = 1.6$, $k = 1.256$ and $g_0 = e^{-1/8} \approx 0.883$, so Lemma 6 applies. The slice likelihood ratio peaks near $w \approx 0.35$ and crosses one once, at $\bar{w} \approx 0.562$. These numbers illustrate the result and are not general.

Table 2 varies the threshold and the prior for the same slices. At low posterior thresholds the accept-all band covers almost all of $(0, 1)$ and w_{low} and $\bar{w}$ are its edges. At higher thresholds Lemma 6 applies and the two coincide. In this symmetric example ($a = -b$) the computed crossing did not move with k once $k > g_0$. This

Table 2: Thresholds of Proposition 5 for $m_{lo} = 0.7$, $m_{hi} = -0.3$ and $\mu_F = 0.2$, on a grid of step 5×10^{-4} . The band is the accept-all interval, where LR = 1.

q	θ	k	$k > g_0$	w_{low}	$\bar{w}$	band
0.3	0.10	0.259	no	0.022	0.978	[0.022, 0.978]
0.3	0.20	0.583	no	0.125	0.875	[0.125, 0.875]
0.3	0.35	1.256	yes	0.562	0.562	none
0.3	0.50	2.333	yes	0.562	0.562	none
0.5	0.10	0.111	no	0.004	0.996	[0.004, 0.996]
0.5	0.35	0.538	no	0.104	0.896	[0.104, 0.896]
0.5	0.50	1.000	yes	0.562	0.562	none

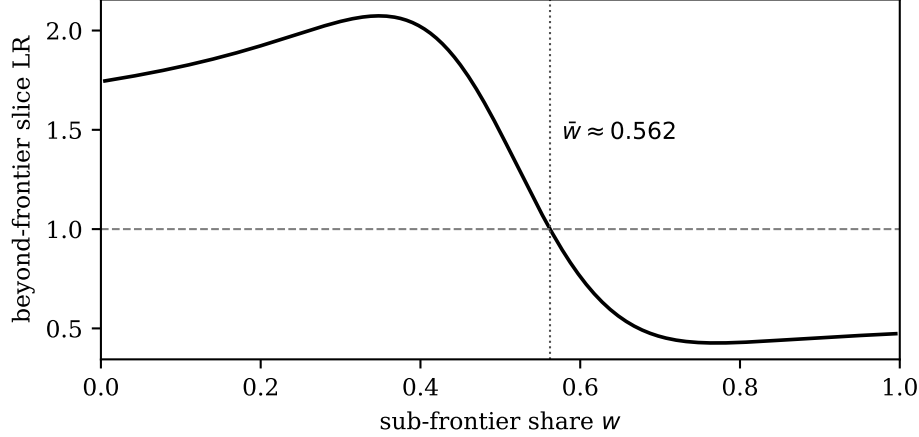


Figure 2: Beyond-frontier slice likelihood ratio under the pooled Bayes rule (2), worked example. The ratio is not monotone in w but crosses one exactly once (Lemma 6). Computed in log space from closed-form tail probabilities.

is a numerical observation for the symmetric case, and we have no proof of it. With $m_{lo} = 0.9$ instead of 0.7, $\bar{w}$ fell from 0.519 to 0.192 as k rose from $1.01 g_0$ to $100 g_0$.

The comparison with Proposition 3 is the point of this section. The informed Bayesian never inverts. The pooled Bayesian inverts the beyond-frontier slice once the sub-frontier share exceeds $\bar{w}$, because its rule is then tuned to the sub-frontier slice. Optimality does not rescue a reviewer that cannot locate the claim.

5 Consequences for corpora and audits

Proposition 7 (Frontier paradox). *Fix τ and an increasing rule. Let a generator emit claims at $d_{lo} < d^*(c)$ with share $1 - m(c)$ and at $d_{hi} > d^*(c)$ with share $m(c)$, with overall false rate $f(c)$. Write s_F , s_{lo} and s_{hi} for the acceptance probabilities of false claims and of true claims in each slice. The false share of the accepted corpus is*

$$\Pi(c) = \frac{f s_F}{f s_F + (1 - f) S_T}, \quad S_T = (1 - m) s_{lo} + m s_{hi},$$

and, for differentiable m , f and $s_F(\mu_F(c))$, $\Pi'(c) > 0$ if and only if $s'_F/s_F + m'(s_{lo} - s_{hi})/S_T > -f'/(f(1 - f))$.

The statement is an identity plus a monotonicity condition. Its content is that a better generator can make the accepted corpus worse while making fewer errors overall. With $\tau = 0.5$, $\mu_0 = 1$, $\lambda = 0.5$, $d_{lo} = 0.1$ and $d_{hi} = 3.1$, the triples $(m, f, \mu_F) = (0.1, 0.30, 0.2)$, $(0.4, 0.25, 0.4)$ and $(0.7, 0.20, 0.6)$ give accepted false shares 0.209, 0.249 and 0.307. The overall false rate falls while the accepted false share rises. The proposition covers A2 reviewers. For a pooled Bayesian, the trap can disappear once w falls below w_{low} .

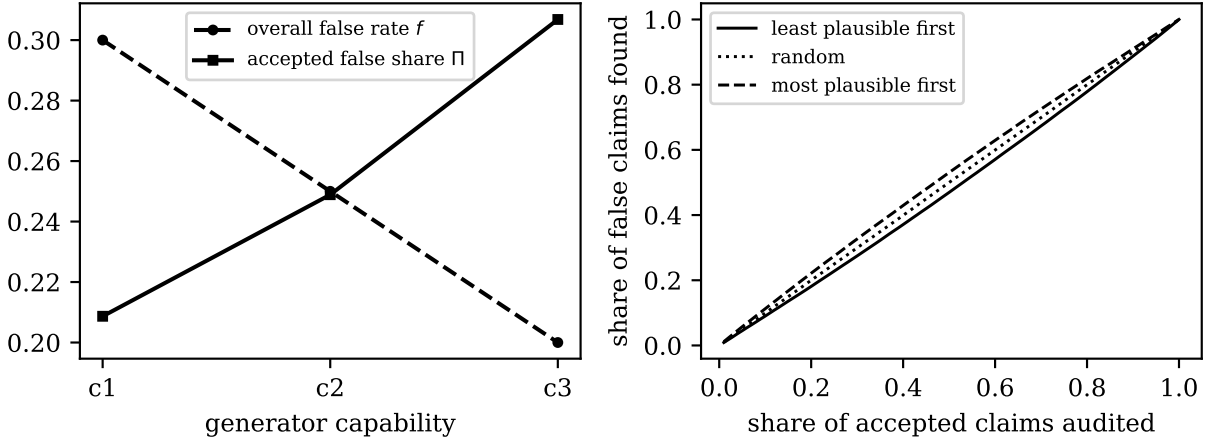


Figure 3: Left, the frontier paradox of Proposition 7, where the overall false rate falls with capability while the false share of the accepted corpus rises. Right, cumulative share of false accepted claims found under three audit orders beyond the frontier (Proposition 8), simulated with 2×10^5 claims.

Proposition 8 (Suspicion-ordered audits invert). *For $d > d^*(c)$, $P(T = 0 | x)$ is increasing in x . Auditing accepted claims least-plausible-first finds fewer false claims per audit than random order, and most-plausible-first is the efficient order.*

The proof is one line from A6, since the likelihood ratio $\varphi(x - \mu_F)/\varphi(x - \mu_T)$ is increasing in x when $\mu_F > \mu_T$. In a simulation with $\mu_T = -0.5$, $\mu_F = 0.2$, $q = 0.3$, $\tau = 0$ and an audit budget of 20% of accepted claims, the share of audited claims that were false (precision) was 0.741 least-plausible-first, 0.814 at random and 0.900 most-plausible-first. Figure 3 shows both worked examples, with the audit panel plotting the complementary recall, the share of all false accepted claims found, as the budget grows.

The practical point is conditional. If a review pipeline operates beyond d^* , the natural instinct to audit what looks weakest is backwards.

6 A pre-registered test of the mechanism

The theorems are conditional. To ask whether their premises bite in a current system, we registered a component-wise test before seeing any outcome data.

Data. Zhou et al. [2025] released per-instance verdicts of 15 generative verifiers (14 open models from 2B to 72B parameters and GPT-4o) on responses from the same 15 models, with correctness labels. Each verifier reads a problem and a response, reasons, and returns “Correct” or “Incorrect”. We use the public balanced subsamples, with 367,920 rows per verifier in mathematics (eight open-answer benchmarks) and 143,499 in knowledge (MMLU-Pro, ten options). The release codes verdicts as 1 for “Correct”, 2 for “Incorrect” and -1 for an unparsable output. We confirmed this coding in the released parsing code. We count -1 as not accepted. Dropping those rows instead leaves the verdict in both domains unchanged.

Protocol. The protocol was frozen before the outcome files were downloaded. Mathematics is the primary domain. Knowledge is secondary and cannot by itself support the prediction. Natural-language reasoning is excluded, because its tasks have at most four options and an earlier below-chance result there was pre-registered as non-counting. Responses whose generator is the verifier are excluded. Novelty is proxied by difficulty relative to the verifier, namely the verifier’s own solve rate on the item over its 64 generator samples in the public generator release, binned as easy ($\geq 2/3$), medium and hard (0). For each verifier, generators other than itself are ranked by overall accuracy in the same release, and the top five and bottom five form the strong and weak tiers, and all of them form the pooled tier. A first run of this analysis computed both quantities from the label-balanced subsample instead, which deviates from the protocol. The numbers below use the protocol definition. The first run gave the same verdict in both domains. A cell is a (verifier, tier, bin)

Table 3: Mathematics, hard bin (verifier solves none of its 64 attempts), strong-generator tier. TPR and TNR are pooled over problems. The interval is a 95% bootstrap interval over problems. ΔTNR is strong minus weak tier, and all its intervals lie below zero.

Verifier	TPR	TNR	LR	95% interval	ΔTNR
Qwen2.5-3B	0.49	0.57	1.15	[1.07, 1.23]	−0.17
Qwen2.5-72B	0.88	0.34	1.33	[1.26, 1.38]	−0.39
Qwen2.5-7B	0.81	0.34	1.23	[1.18, 1.28]	−0.38
Qwen3-32B	0.53	0.72	1.89	[1.52, 2.27]	−0.18
Qwen3-4B	0.44	0.75	1.78	[1.50, 2.11]	−0.17
Qwen3-8B	0.64	0.68	2.02	[1.74, 2.30]	−0.21
gemma-2-27b-it	0.85	0.37	1.35	[1.30, 1.40]	−0.30
gemma-2-2b-it	0.58	0.58	1.40	[1.34, 1.47]	−0.12
gemma-2-9b-it	0.83	0.39	1.35	[1.31, 1.40]	−0.21
gpt-4o	0.57	0.59	1.38	[1.23, 1.55]	−0.30
Llama-3.1-8B	0.50	0.64	1.40	[1.32, 1.48]	−0.19
Llama-3.2-3B	0.72	0.47	1.34	[1.29, 1.39]	−0.24
Llama-3.3-70B	0.92	0.27	1.27	[1.23, 1.31]	−0.39
Minstral-8B	0.55	0.60	1.38	[1.29, 1.48]	−0.21
Mistral-Small-24B	0.66	0.58	1.57	[1.46, 1.70]	−0.31

triple with at least 30 true and 30 false responses. Confidence intervals come from 1,000 bootstrap resamples of problems. The likelihood ratio is $\text{TPR}/(1 - \text{TNR})$, the acceptance ratio of (1).

Components and decision rule. M1 (shaping) requires that the true-negative rate be lower for strong than for weak generators in medium or hard bins, with the upper confidence bound of the difference below zero. Its coherence sub-component needs error-type labels that are not released and was not run. M2 (fidelity decay) requires the true-positive rate to fall from easy to medium to hard for a majority of verifiers. M3 (crossing) requires a strong-tier medium or hard cell with LR upper bound below one, true-positive rate at least 0.20, and M1 holding in the same verifier and bin. The signed rule declares the prediction falsified if no included cell has $\text{LR} < 1$.

Result in mathematics. All 135 cells were valid. M1 held in all 30 medium and hard (verifier, bin) pairs, with true-negative differences between -0.06 and -0.47 . M2 held for 14 of 15 verifiers. M3 failed. Every cell had $\text{LR} > 1$, with minimum point estimate 1.15 and minimum upper bound 1.23. On hard items the median strong-tier true-positive rate was 0.636 and the median strong-tier true-negative rate was 0.581. With unparseable verdicts dropped, the minimum point estimate is 1.13 and no cell falls below one. Strong generators make verifiers worse at catching errors, and verifiers find hard items harder, but neither effect is large enough to reverse the sign of acceptance. **The pre-registered prediction is falsified on the primary domain.** Table 3 lists the hard-bin cells, where an inversion was most likely.

Confounds. The protocol listed six systematic routes to $\text{LR} < 1$ that bypass the mechanism. A closed answer space allows solve-and-match, which we limited by excluding tasks with four or fewer options. Verifier failure to solve can drive TPR towards zero, which the TPR floor in M3 guards against. Correctness labels may be noisy, selection on filtered subsets can manufacture difficulty, strong and weak generators may differ in length and format, and a difficulty proxy built from the generators being compared would be circular. We addressed the last by using the verifier’s own solve rate. Label noise and format differences were not modelled. Random label noise only pulls LR towards one, so it cannot by itself hide an inversion of meaningful size. Length and format differences could bias LR in either direction, and we cannot rule out that they mask or create an effect.

Knowledge, exploratory. Thirty-three knowledge cells (11 pooled, 7 strong, 15 weak) have an LR upper bound below one, and seven of them meet the formal M3 conditions. We do not read this as support. All 33 lie in the hard bin, where the verifier solved none of its 64 attempts, and none lie in easy or medium bins, where the minimum LR is 1.12. With ten fixed options, a verifier that has settled on a wrong option accepts responses that match it and rejects the correct one. That produces $\text{LR} < 1$ without any shaping. The mechanism predicts that the inversion is driven by strong generators. In the hard knowledge bin the opposite holds. All 15 weak-tier cells have upper bound below one, with median LR 0.674, against 7 of 15

Table 4: Knowledge (exploratory), hard bin. LR and 95% bootstrap interval for strong and weak generator tiers. Weak-tier cells invert more often and more strongly, the opposite of the mechanism’s signature.

Verifier	LR strong	interval	LR weak	interval
Qwen2.5-3B	1.01	[0.91, 1.12]	0.62	[0.52, 0.72]
Qwen2.5-72B	0.85	[0.80, 0.90]	0.75	[0.62, 0.87]
Qwen2.5-7B	0.94	[0.86, 1.02]	0.78	[0.67, 0.87]
Qwen3-32B	0.48	[0.31, 0.65]	0.25	[0.13, 0.43]
Qwen3-4B	0.73	[0.61, 0.86]	0.38	[0.26, 0.51]
Qwen3-8B	0.56	[0.44, 0.70]	0.31	[0.19, 0.45]
gemma-2-27b-it	0.97	[0.90, 1.05]	0.81	[0.70, 0.92]
gemma-2-2b-it	1.06	[0.94, 1.19]	0.84	[0.75, 0.93]
gemma-2-9b-it	1.03	[0.97, 1.09]	0.81	[0.73, 0.88]
gpt-4o	0.74	[0.62, 0.83]	0.59	[0.46, 0.73]
Llama-3.1-8B	0.83	[0.59, 1.06]	0.46	[0.30, 0.62]
Llama-3.2-3B	0.93	[0.79, 1.06]	0.67	[0.54, 0.82]
Llama-3.3-70B	0.84	[0.79, 0.90]	0.77	[0.68, 0.87]
Ministral-8B	1.00	[0.95, 1.04]	0.87	[0.76, 0.97]
Mistral-Small-24B	0.70	[0.57, 0.83]	0.40	[0.29, 0.53]

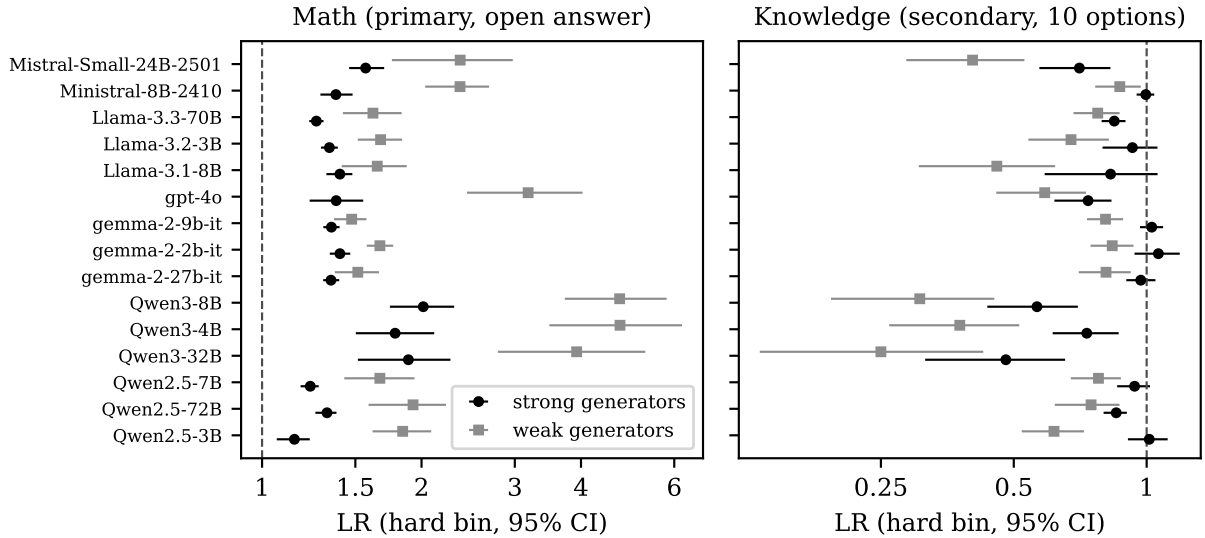


Figure 4: Hard-bin likelihood ratio by verifier, strong and weak generator tiers, with bootstrap 95% intervals (log scale). Left, open-answer mathematics, where no cell falls below one. Right, ten-option knowledge questions, where weak-tier cells invert more than strong-tier cells.

strong-tier cells with median 0.853. The inversion is better explained by verifier failure on a closed answer space than by the mechanism. Table 4 gives the cells.

Descriptive note. In mathematics, the strong-tier LR is below the weak-tier LR in 44 of the 45 (verifier, bin) pairs. This pattern was not pre-registered. We report it as description only and draw no claim from it.

Reading the result. In model terms the gap $\mu_T - \mu_F$ shrinks as generators get stronger but stays positive over the difficulty range of these benchmarks. The frontier d^* is not reached. This is consistent with Theorem 1, which says nothing about where d^* lies. It is inconsistent with our prediction that d^* would be reached in this setting. One could explain the result by noting that mathematical derivations can be checked step by step, which would mean a small λ and a low ceiling $\bar{\mu}_F$. We did not state that scope condition in advance, so we do not adopt it here.

7 Proofs

This section gives the proofs in steps. Lemma 6 was proved in Section 4.

7.1 Theorem 1 and Propositions 2, 3, 4 and 8

Throughout, $\sigma = 1$, Φ is the standard normal distribution function and $\Delta = \mu_T - \mu_F$.

Step 1, part (a). Let $G(d) = \mu_T(d) - \mu_F(c, d)$. By A3 and A4, $G'(d) = -\lambda - \partial\mu_F/\partial d \leq -\varepsilon$. So $G(d) \leq G(0) - \varepsilon d$, which is negative for $d > G(0)/\varepsilon$. Hence the set $\{d : G(d) \geq 0\}$ is bounded and $d^*(c) = \sup\{d : G(d) \geq 0\} \leq (\mu_0 - \mu_F(c, 0))/\varepsilon$. Since G is strictly decreasing, $G(d) < 0$ for all $d > d^*$. When μ_F does not depend on d , $G(d) = \mu_0 - \lambda d - \mu_F(c)$, and solving $G = 0$ gives the closed form. It is nonincreasing in c because μ_F is nondecreasing, and it is bounded below because $\mu_F \leq \bar{\mu}_F$. If $\mu_F(c) \geq \mu_0$ then $G(0) \leq 0$ and $G < 0$ on $d > 0$, so $d^*(c) \leq 0$.

Step 2, part (b), likelihood ratio. For a threshold τ , $\text{TPR} = 1 - \Phi(\tau - \mu_T)$ and $\text{FPR} = 1 - \Phi(\tau - \mu_F)$. If $\mu_T < \mu_F$ then $\tau - \mu_T > \tau - \mu_F$, so $\text{TPR} < \text{FPR}$ for every finite τ .

Step 3, part (b), area under the curve. For independent $X_T \sim N(\mu_T, 1)$ and $X_F \sim N(\mu_F, 1)$, $X_T - X_F \sim N(\Delta, 2)$, so $P(X_T > X_F) = \Phi(\Delta/\sqrt{2})$, which is below $1/2$ when $\Delta < 0$.

Step 4, part (b), posteriors. Bayes' rule gives

$$\frac{P(T = 1 \mid \text{accept})}{P(T = 0 \mid \text{accept})} = \frac{q}{1 - q} \cdot \frac{\text{TPR}}{\text{FPR}}, \quad \frac{P(T = 1 \mid \text{reject})}{P(T = 0 \mid \text{reject})} = \frac{q}{1 - q} \cdot \frac{1 - \text{TPR}}{1 - \text{FPR}}.$$

The first ratio is below $q/(1 - q)$ and the second above it.

Step 5, part (c). For the decreasing rule “accept if and only if $x \leq \tau$ ”, $\text{LR} = \Phi(\tau - \mu_T)/\Phi(\tau - \mu_F) > 1$ when $\mu_T < \mu_F$. Below d^* the increasing rule has $\text{LR} > 1$ by Step 2 with the inequality reversed. At $\Delta = 0$ the two distributions are equal, so every rule has $\text{LR} = 1$.

Step 6, part (d1) and Proposition 4. The beyond-frontier slice has $\Delta < 0$, and Step 2 applies to it unchanged.

Step 7, Proposition 2. If $\Delta \geq 0$, Step 2 gives $\text{TPR} \geq \text{FPR}$, so $\text{TPR}/\text{FPR} \geq 1 \geq (1 - \text{TPR})/(1 - \text{FPR})$. Step 4 then orders the posteriors for every q .

Step 8, Proposition 3. The posterior odds are $\frac{q}{1-q} \exp(\Delta x - (\mu_T^2 - \mu_F^2)/2)$. With $\Delta < 0$ they decrease in x , so the set $\{P(T = 1 \mid x) \geq \theta\}$ is a lower set. Every accepted claim has posterior at least θ and every rejected claim below θ . Averaging gives $P(T = 1 \mid \text{accept}) \geq \theta \geq P(T = 1 \mid \text{reject})$.

Step 9, Proposition 8. By the same computation $P(T = 0 \mid x)$ increases in x when $\Delta < 0$. Ordering audits by decreasing x therefore ranks accepted claims by decreasing probability of being false.

7.2 Proposition 5

Step 1, convexity. $g_w''(x) = wa^2s(x) + (1 - w)b^2r(x) > 0$, so g_w is strictly convex with a unique minimiser. If $\min g_w < k$, the acceptance set is $(-\infty, x_1] \cup [x_2, \infty)$ with $x_1 < x_2$. If $\min g_w \geq k$, every claim is accepted and $\text{LR} = 1$.

Step 2, the limit $w \rightarrow 1$. For fixed x , $g_w(x) \rightarrow s(x)$, and $\min_x g_w \rightarrow 0 < k$. So $x_1 \rightarrow -\infty$ and $x_2 \rightarrow x_2^* = \log(k/A)/a$. The slice likelihood ratio tends to

$$\frac{1 - \Phi(x_2^* - m_{\text{hi}})}{1 - \Phi(x_2^* - \mu_F)} < 1.$$

Step 3, the limit $w \rightarrow 0$. Symmetrically $x_2 \rightarrow \infty$ and $x_1 \rightarrow x_1^* = \log(k/B)/b$, and the ratio tends to

$$\frac{\Phi(x_1^* - m_{\text{hi}})}{\Phi(x_1^* - \mu_F)} > 1.$$

Step 4, continuity and thresholds. Off the accept-all band, x_1 and x_2 are continuous in w by the implicit function theorem, and on the band $\text{LR} = 1$. At the band's edges the boundaries merge at the minimiser and the acceptance set tends to the whole line, so LR is continuous on $(0, 1)$. Define $\bar{w} = \inf\{w : \text{LR}(w') <$

1 for all $w' > w$ and $w_{\text{low}} = \sup\{w : \text{LR}(w') > 1 \text{ for all } w' < w\}$. Steps 2 and 3 give $0 < w_{\text{low}}$ and $\bar{w} < 1$. If $w_{\text{low}} > \bar{w}$, any w strictly between them would satisfy both $\text{LR} < 1$ and $\text{LR} > 1$, so $w_{\text{low}} \leq \bar{w}$.

Step 5, $\bar{w}$ near one. In the symmetric case $a = -b$, the minimum of g_w is $2\sqrt{w(1-w)AB}$, so accept-all holds if and only if $k^2 \leq 4w(1-w)AB$. As $k \rightarrow 0$ this band covers $[\eta, 1-\eta]$ for any fixed $\eta > 0$, and $\bar{w}$ is at least the band's upper edge. So $\bar{w}$ can be made arbitrarily close to one. The symmetric boundaries are $e^{ax} = [k \pm \sqrt{k^2 - 4w(1-w)AB}]/(2wA)$.

8 Discussion

What the theory does and does not claim. The theory makes three kinds of statement. The first is about sign. Beyond d^* , reading plausibility as evidence produces acceptance that is evidence against truth. The second is about information. The inversion is not a loss of signal, and a reviewer that knows its position relative to d^* recovers an informative rule. The third is about optimality without location. A pooled Bayesian, the best a reviewer can do without location, still inverts the beyond-frontier slice once the sub-frontier share exceeds $\bar{w}$. None of these statements says how large d^* is for any real reviewer.

Self-location as a design target. Part (c) of Theorem 1 turns a pessimistic result into a design question. A reviewer needs some signal of which side of d^* a claim is on. Three kinds of signal are natural. The reviewer's own ability to solve the item is one, and it is the proxy our test used. Agreement between independent reviewers with different frontiers is another. External checks such as execution, formal proof or replication bypass x altogether, as Proposition 4 requires. Our test suggests a caution about the first signal. In mathematics, verifiers that never solved an item still accepted correct responses more often than incorrect ones, so failure to solve did not place them beyond d^* .

Corpora and audits. Propositions 7 and 8 concern what happens downstream. If a pipeline does operate beyond its frontier, improvements in the generator can worsen the accepted corpus, and suspicion-ordered audits look in the wrong place. These consequences are conditional on the same location question, and our test gives no evidence that they occur in open-answer mathematics today. They become relevant wherever claims are hard to check step by step and generators are optimised for acceptance, which are also the settings where testing is hardest.

What a confirming test would need. A test that could find the inversion needs per-instance verdicts, independent correctness labels and generator identities on items that lie far from the reviewer's frontier. It also needs an open answer space, so that solve-and-match cannot mimic the effect, and a protocol frozen before access. We are not aware of a public dataset that satisfies all of these conditions at once.

9 Related work

Likelihood ratios and false findings. The algebra by which a test with $\text{LR} < 1$ makes positive results less credible than negative ones is classical [Ioannidis, 2005]. Our contribution is to derive the sign of d' from novelty and generator capability, and to locate a frontier where it changes.

Testing strategic experts. Strategic forecasters can pass broad classes of tests [Olszewski and Sandroni, 2008], and computational limits can restore a tester's power [Fortnow and Vohra, 2009]. Those results concern whether a test can be passed. Ours concerns a receiver-side novelty index at which acceptance reverses.

Persuasion and credulity. Credulous receivers can be systematically misled [Kartik et al., 2007], a Bayesian receiver's posteriors average to its prior, so it cannot be fooled on average [Kamenica and Gentzkow, 2011], and robust signalling for boundedly rational receivers has been characterised [Feng et al., 2022]. Proposition 3 places our setting in this map. The informed Bayesian inverts its ranking rather than pooling, and the trap belongs to the reviewer that reads plausibility as evidence or cannot locate the claim.

Limits of evaluation. Mahran [2026] show that an evaluator cannot self-assess what it fails to distinguish. Theorem 1 adds a graded discriminability that changes sign at a computable $d^*(c)$ and shows that self-location is sufficient to avoid the inversion, so that a failure of self-location is necessary for it. Bowkis et al. [2026] argue that mistakes made by AI agents doing alignment research concentrate where human reviewers are least likely to catch them. NOVA models discovery by AI systems as generate, verify, accumulate and retrain,

and identifies a contamination trap in which false positives outpace genuine discoveries as easy knowledge is exhausted [Avestimehr et al., 2026]. Our model supplies an acceptance rule under which that contamination becomes anti-informative rather than merely noisy. The extremal and adversarial variants of Goodhart’s law describe the same pressure from the optimiser’s side [Manheim and Garrabrant, 2018].

Outside signals. Debate between capable agents helps non-expert judges, reaching 76% for LLM judges and 88% for humans on QuALITY [Khan et al., 2024]. The agent score difference, a log-ratio measure of how much a protocol favours truth [Sudhir et al., 2025], is the natural multi-agent analogue of our log LR. Both lie outside A1, and Proposition 4 says that some signal of this kind is required.

Evidence on the assumptions. LLM reviewer scores respond to rhetoric and are raised by adversarial rephrasing with content held fixed [Li et al., 2026b,a], and are associated with explicit novelty claims in observational data [Kousha and Thelwall, 2026], which bears on A4. Multimodal judges show a significant criterion shift towards acceptance without a clear change in mean discriminability [Cai et al., 2026], which bears on A2. Human panels penalise novel proposals [Boudreau et al., 2016], and scientists rate ideas that resemble their own more favourably [Bao et al., 2026], which bears on A3. Repeated statements feel truer even against knowledge [Fazio et al., 2015]. At low novelty, laypeople predict replication above chance [Hoozeveen et al., 2020], and a structured expert elicitation reached a classification accuracy of 84% with an area under the curve of 0.94 in its validation study [Fraser et al., 2023], consistent with $d' > 0$ below the frontier. Non-replicable papers are cited more than replicable ones [Serra-Garcia and Gneezy, 2021], which is suggestive of plausibility outrunning truth in a human corpus but is not a test of our mechanism.

10 Limitations

The theorems rely on equal variances (A6). With unequal variances the “every threshold” clauses hold only for a range of thresholds. Fidelity decay is linear in A3, which approximates at best the right tail of the evidence in Boudreau et al. [2016], where scores rise with novelty at low levels. Uniqueness of the pooled crossing is proven only for $k > g_0$.

The test covers one release, one family of verifiers and one notion of novelty. Difficulty relative to the verifier is a proxy for distance from its frontier, not a measure of it. The coherence part of M1 was not run. The knowledge result is confounded by the closed answer space and by the definition of the hard bin, and we treat it as exploratory. The same release contains further responses on the same items, but these are not independent and cannot confirm anything.

The main threat to relevance is that d^* may lie beyond the difficulty of current benchmarks. If so, the theory is true and currently idle. Our test is evidence for that reading in open-answer mathematics.

11 Conclusion

Plausibility-as-evidence review has a sign that can flip. Beyond a frontier set by the reviewer’s knowledge and the generator’s skill, accepted claims are less likely to be true than rejected ones, and the least plausible accepted claims are the safest. The flip is a failure of self-location, not a loss of information, and a pooled optimal reviewer suffers it once the sub-frontier share of claims exceeds a threshold $\bar{w}$. Our pre-registered test did not find the flip where we expected it. Strong generators do erode verifier specificity, but in open-answer mathematics acceptance remained informative in every cell. The open question is empirical. Where, if anywhere, do current review pipelines cross d^* ?

References

- Salman Avestimehr, Ken Duffy, and Muriel Médard. Nova: Fundamental limits of knowledge discovery through ai, 2026.
- Honglin Bao, Siyang Wu, Xiao Liu, Sida Li, Shiyun Cao, and James A. Evans. Contemporary ai lacks the imagination to diverge or negate in science, 2026.
- Kevin J. Boudreau, Eva C. Guinan, Karim R. Lakhani, and Christoph Riedl. Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science. *Management Science*, 62(10):2765–2783, 2016. doi:[10.1287/mnsc.2015.2285](https://doi.org/10.1287/mnsc.2015.2285).
- Aleksandr Bowkis, Marie Davidsen Buhl, Jacob Pfau, and Geoffrey Irving. Automated alignment is harder than you think, 2026.

- Anglin Cai, Zhiqi Wang, Xiaozhi Yang, Yurun Lin, Dongwon Lee, and Yuchen Yang. A plausible explanation is enough to mislead multimodal judges, 2026.
- Lisa K. Fazio, Nadia M. Brashier, B. Keith Payne, and Elizabeth J. Marsh. Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, 2015. doi:[10.1037/xge0000098](https://doi.org/10.1037/xge0000098).
- Yiding Feng, Chien-Ju Ho, and Wei Tang. Rationality-robust information design: Bayesian persuasion under quantal response, 2022.
- Lance Fortnow and Rakesh V. Vohra. The complexity of forecast testing. *Econometrica*, 77(1):93–105, 2009. doi:[10.3982/ECTA7163](https://doi.org/10.3982/ECTA7163).
- Hannah Fraser, Martin Bush, Bonnie C. Wintle, Fallon Mody, Eden T. Smith, Anca M. Hanea, Elliot Gould, Victoria Hemming, Daniel G. Hamilton, Libby Rumpff, David P. Wilkinson, Ross Pearson, Felix Singleton Thorn, Raquel Ashton, Aaron Willcox, Charles T. Gray, Andrew Head, Melissa Ross, Rebecca Groenewegen, Alexandru Marcoci, Ans Vercammen, Timothy H. Parker, Rink Hoekstra, Shinichi Nakagawa, David R. Mandel, Don van Ravenzwaaij, Marissa McBride, Richard O. Sinnott, Peter Vesk, Mark Burgman, and Fiona Fidler. Predicting reliability through structured expert elicitation with the replicats (collaborative assessments for trustworthy science) process. *PLOS ONE*, 2023. doi:[10.1371/journal.pone.0274429](https://doi.org/10.1371/journal.pone.0274429).
- Suzanne Hoogeveen, Alexandra Sarafoglou, and Eric-Jan Wagenmakers. Laypeople can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, 2020. doi:[10.1177/2515245920919667](https://doi.org/10.1177/2515245920919667).
- John P. A. Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2005. doi:[10.1371/journal.pmed.0020124](https://doi.org/10.1371/journal.pmed.0020124).
- Fengqing Jiang, Yichen Feng, Yuetai Li, Luyao Niu, Basel Alomair, and Radha Poovendran. Badscentist: Can a research agent write convincing but unsound papers that fool llm reviewers?, 2025.
- Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6): 2590–2615, 2011. doi:[10.1257/aer.101.6.2590](https://doi.org/10.1257/aer.101.6.2590).
- Navin Kartik, Marco Ottaviani, and Francesco Squintani. Credulity, lies, and costly talk. *Journal of Economic Theory*, 134(1):93–116, 2007. doi:[10.1016/j.jet.2006.04.003](https://doi.org/10.1016/j.jet.2006.04.003).
- Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers, 2024.
- Kayvan Kousha and Mike Thelwall. Are explicit novelty claims in research abstracts associated with higher chatgpt scores?, 2026.
- Lin Li, Qi Zhang, Xander Davies, Jianing Qiu, and Yarin Gal. Gaming ai-assisted peer reviews poses new risks to the scientific community, 2026a.
- Ming Li, Chenguang Wang, Xirui Li, Xinyue Zeng, Dianqi Li, Peng Shi, Dawei Zhou, and Tianyi Zhou. How can rhetoric reward-hack ai reviewers? dissecting rhetorical sensitivity in ai-based peer review, 2026b.
- Ahmed M. Mahran. What lies beyond perception: A partition theory of the limits of ai evaluation and oversight, 2026.
- David Manheim and Scott Garrabrant. Categorizing variants of goodhart’s law, 2018.
- Wojciech Olszewski and Alvaro Sandroni. Manipulability of future-independent tests. *Econometrica*, 76(6): 1437–1466, 2008. doi:[10.3982/ECTA7428](https://doi.org/10.3982/ECTA7428).
- Marta Serra-Garcia and Uri Gneezy. Nonreplicable publications are cited more than replicable ones. *Science Advances*, 2021. doi:[10.1126/sciadv.abd1705](https://doi.org/10.1126/sciadv.abd1705).
- Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas, 2025.
- Abhimanyu Pallavi Sudhir, Jackson Kaunismaa, and Arjun Panickssery. A benchmark for scalable oversight protocols, 2025.
- Yefan Zhou, Austin Xu, Yilun Zhou, Janvijay Singh, Jiang Gui, and Shafiq Joty. Variation in verification: Understanding verification dynamics in large language models, 2025.

A Numerical verification

Every step above was checked numerically. A double-precision script checks parts (a) to (d) on 200 random instances each, the limits and closed forms of Proposition 5, and Propositions 2 to 8, passing 17 of 17 checks. The steps of Lemma 6 were checked at 60 significant digits. The derivative formula in Step 2 agrees with a finite difference to relative error 4×10^{-35} on 39 instances, and Steps 3 to 5 held on every tested instance. The log-space evaluator of the uniqueness search agrees with the 60-digit one to 5×10^{-14} . Scripts and outputs are in the project repository.