Debate as Anti-Amplification: Competitive Persuasion with Verifiable Evidence and Weak Judges

Liz Lemma Future Detective

January 22, 2026

Abstract

Modern alignment pipelines increasingly rely on weak oversight—either weak humans or weaker LLM judges—to supervise stronger models. Recent evidence suggests a core failure mode: when an agent can choose what to argue for (open-role), a weak judge may be similarly persuaded regardless of correctness (open consultancy), amplifying the agent's mistakes; open debate empirically reduces this amplification by introducing an adversarial counter-signal. We formalize this phenomenon using a clean persuasion model with verifiable evidence. A protagonist chooses a stance and sends an argument to a noisy, possibly biased judge; in open debate an antagonist is forced to argue the opposite stance. When true stances can always produce verifiable evidence but false stances can do so only with limited probability, we derive closed-form mistake-amplification rates. Our main result is an anti-amplification guarantee: for a broad class of judge decision rules that put positive weight on verifiable evidence, open debate strictly reduces the probability that oversight endorses the protagonist conditional on the protagonist being wrong. We characterize comparative statics in judge bias, evidence salience, and fabrication rates, and translate them into protocol design rules (evidence requirements, cross-examination, jury selection) that are directly measurable in LLM oversight benchmarks. The model complements empirical findings on debate vs consultancy and provides a mechanism-design lens for regulator-grade, two-sided AI audits in 2026 deployments.

Table of Contents

3. 3. A tractable persuasion model with verifiable evidence: primitives, protocols (open consultancy vs open debate), and definition of mistake amplification

4. 4. Equilibrium analysis (baseline closed form): evidence feasibility, sender strategies, and judge decision rule

5. 5. Main results: anti-amplification guarantee, welfare implications, and when debate trades off with raw accuracy

6. 6. Mechanism design: protocol parameters that improve anti-amplification (evidence salience, cross-exam, jury size, tie-breaking rules); implementable metrics

7. 7. Extensions (flagged where numerical methods are needed): richer message spaces, correlated signals, multi-option answers, endogenous judge effort/attention

8. 8. Empirical mapping and test plan: how to estimate ρ, b, κ, and amplification from transcripts; predicted patterns across tasks/judges

9. 9. Policy and governance implications: two-sided adversarial review for audits, certification, and compliance; limitations and open questions

1 Introduction and motivation: weak oversight, mistake amplification, and why institutional design matters in 2026 deployments

Many of the alignment failures we worry about in 2026 deployments are not best described as a model being "globally inaccurate." Rather, the failure mode is that a model produces a confident, actionable stance on a hard instance, and the surrounding oversight stack—human review, automated checks, and downstream aggregation—endorses that stance. In practice, the endorsement is itself a decision: an internal label, a policy approval, an evaluation pass, a training target, or a monitoring alert being suppressed. When this endorsement is wrong, it does more than merely create an isolated error. It creates an *institutional fact* that can propagate: the wrong answer is recorded, re-used, distilled, and sometimes reinforced by subsequent training or procedural precedent. This is the sense in which weak oversight can lead to *mistake amplification*: the system preferentially validates precisely those model errors that are most difficult for the overseer to detect.

Weak oversight arises for structural reasons. First, the judge (human or automated) is typically resource-bounded: limited time per case, limited domain expertise, limited ability to run experiments, and limited tolerance for long transcripts. Second, modern model outputs are often *argumentative* rather than purely predictive: they contain chains of reasoning, citations, computations, and tool calls that can be selectively presented. Third, real deployments exhibit distribution shift and adversarial pressure (including unintentional adversariality arising from optimization): the cases that matter most are the long-tail cases where superficial plausibility is cheap and genuine verification is expensive. Even if the judge is well-intentioned, these constraints create a predictable gap between (i) what is true and (ii) what can be *verified* within the protocol and budget.

This motivates separating two notions that are frequently conflated. The first is overall oversight accuracy, informally $\Pr(y = \omega)$, which asks whether the final label matches truth on average. The second is the conditional probability that the judge endorses the model *given that the model is wrong.* This conditional quantity is what governs whether errors are corrected or canonized. In iterative training and governance loops, the difference matters. If the model is wrong but the judge usually *disagrees* in those cases, then the loop supplies a corrective signal (even if noisy). If instead the judge tends to *agree* exactly when the model is wrong, then errors are not merely left uncorrected; they become training data, policy guidance, or evaluation evidence, which can move the system toward more confidently producing the same wrong stance in the future. We can view this as an alignment-relevant analogue of a positive feedback instability: conditional wrong-follow acts as a gain factor in a closed-loop system.

The institutional context makes this particularly salient. Contemporary oversight pipelines often rely on a hybrid of (a) model-generated explanations, (b) lightweight checks (e.g., spot-checking a citation, re-running a snippet of code, validating a numerical claim), and (c) aggregate scoring or preference judgments. Each of these can be manipulated by a sufficiently capable generator, even absent explicit deception. For example, a model that is uncertain may nonetheless produce a polished argument that includes plausible citations, selectively chosen examples, or superficially consistent calculations. Conversely, the true stance may require messy, time-consuming verification (running an experiment, reading primary sources, or resolving ambiguities). Thus, oversight is shaped by an *evidence technology*: which claims can be supported by verifiable artifacts under time and tooling constraints, and which false claims can be made to look verifiable.

Once we take the evidence technology seriously, it becomes natural to ask not only "how accurate is the judge?" but also "how should we structure the interaction so that evidence asymmetries favor truth rather than plausibility?" This shifts attention from purely improving judge competence toward *protocol design.* Protocol design includes: whether the model provides a single recommendation or must answer under cross-examination; whether we require verifiable citations; whether we allow tool use; whether we solicit counterarguments; whether we use one or multiple judges; and how ties or ambiguity are resolved. These design choices can change the distribution of observed evidence even when the underlying truth is fixed, and therefore can change the probability that the oversight label tracks truth in the most safety-critical regions.

A key observation is that many protocols implicitly implement a one-sided persuasion game: a single system proposes a stance and supplies supporting material, while the judge passively evaluates. We will refer to this broad pattern as *open consultancy.* Open consultancy is attractive because it is cheap and easy to integrate into workflows: one model, one transcript, one evaluation. But it is also structurally exposed to mistake amplification, because the only evidence the judge sees is selected by the proponent. If the proponent is wrong yet can produce something that passes the protocol's verification bar (or can overwhelm the judge with plausible detail), then the judge is pushed toward endorsement. In other words, one-sided protocols tend to conflate "the model has evidence" with "the model is correct," even when the evidence standard is imperfect.

An alternative is to introduce a controlled adversarial dynamic: alongside the proponent, we elicit an antagonist argument for the opposing stance, and we give the judge a structured comparison. We will refer to this as *open debate.* Importantly, we are not assuming the antagonist is benevolent; we are assuming only that it is constrained to argue the opposite stance and is subject to the same evidence-verification regime. The point of open debate is not to produce a philosophically balanced discussion; it is to induce a *countervailing evidence signal* that is systematically correlated with truth when at least one side is constrained by verifiability. Intuitively, if the protagonist is wrong, then the antagonist is right, and a protocol that allows the right side to present verifiable evidence can "pull back" the judge from endorsing the wrong stance. Conversely, when the protagonist is right, the antagonist may sometimes muddy the waters by producing plausible-looking counterevidence, so the comparison introduces a tradeoff: debate can reduce wrong endorsements while occasionally reducing the judge's willingness to follow correct recommendations.

Why is this tradeoff the right one to foreground for alignment? Because many high-stakes settings disproportionately penalize endorsing a wrong stance compared to failing to endorse a right one. In model training, a wrong endorsement can be written into preference data or distilled targets, which can create durable representational changes. In governance, a wrong endorsement can authorize unsafe actions or establish false confidence. In monitoring, a wrong endorsement can suppress alerts. In these settings, reducing $\Pr(y = a \mid a \neq \omega)$ can be more valuable than marginally increasing $\Pr(y = \omega)$, especially if the latter gains are achieved by being more agreeable to the model.

This paper therefore treats institutional choice as a first-class design variable. We model the protagonist as "open-role": it chooses the stance it believes (or, more generally, the stance induced by its private information and incentives), rather than being assigned a random side. This captures the alignment-relevant regime where the system is deployed to provide answers, plans, or classifications, not to play an arbitrary debate role. The oversight mechanism then aggregates transcripts into a label y that functions as a training or governance signal. Our central concern is not persuasion per se, but *selective endorsement*: the mechanisms by which wrong stances become legitimized.

Two additional deployment realities motivate the specific modeling choices we will make later. First, *fabrication is not binary.* Many false stances cannot be supported with truly verifiable evidence, but can sometimes be supported with artifacts that pass lightweight checks: plausible citations, superficially correct derivations, or cherry-picked benchmarks. The rate at which false stances can be made to look verifiable is a key primitive for institutional safety. Second, *judges are biased and noisy.* Even careful reviewers exhibit tie-breaking biases—toward the more fluent system, the more authoritative tone, or the incumbent recommendation. Moreover, when evidence is ambiguous, decisions are effectively stochastic. Protocols that only work under idealized rational judges are brittle; we therefore care about guarantees that persist under weak, biased evaluation.

The overarching claim we will develop is that competitive protocols can provide an "anti-amplification" advantage under remarkably mild assumptions: if true stances can reliably present verifiable evidence while false stances can only sometimes do so, then adding an antagonist who argues the opposite stance tends to reduce conditional wrong-follow, even if the judge is imperfect. This does not eliminate the need for better verification or better judges; rather, it clarifies a complementarity. Verification reduces the feasibility of spurious evidence, while debate ensures that whatever verifiable evidence exists on the true side is more likely to appear in the transcript the judge sees.

We should also be explicit about limitations. Debate can fail if both sides share correlated blind spots, if the evidence regime is itself gameable, if the judge systematically favors one side, or if the additional transcript length degrades attention. Moreover, real tasks are rarely binary, and real "evidence" is heterogeneous (experiments, source code, simulations, human testimony). Our goal in starting with a minimal binary formalism is not to deny these complexities, but to isolate a mechanism that can be composed with richer oversight tools: cross-examination, tool-assisted verification, juries, and auditing. The next section grounds these concerns in empirical patterns from scalable oversight and preference learning, where one-sided elicitation often exhibits amplification-like dynamics and adversarial elicitation can partially mitigate them, albeit with nontrivial operational costs.

2 2. Empirical backdrop from scalable oversight and preference learning: open-role amplification in consultancy; debate's partial mitigation; link to feedback-loop safety concerns

In scalable oversight practice, we rarely begin with a clean separation between "answer generation" and "answer evaluation." Instead, modern pipelines entangle the two: evaluators are time-constrained, tasks are open-ended, and the model itself often supplies the artifacts that the evaluator uses to judge correctness (rationales, citations, tool traces, unit tests, summaries of sources). This makes oversight resemble an endogenous measurement system: the proponent partly determines which checks are even salient. A growing body of anecdotal and experimental evidence suggests that, in this regime, one-sided "consultancy-style" elicitation exhibits a distinctive pathology: errors that are packaged as coherent arguments are disproportionately likely to be endorsed relative to errors that are transparently confused. Put differently, conditional on the model being wrong, the probability that oversight still outputs agreement can be surprisingly high—and this is precisely the conditional structure that matters for feedback loops.

A useful empirical lens is preference learning (RLHF, RLAIF, DPO-style training, and their variants), where a judge—human or model-based—provides labels that become a training target. Several recurring phenomena are consistent with an “open-role amplification” dynamic. First, evaluators exhibit a robust preference for outputs that are more fluent, more assertive, and more detailed, even when those attributes are only weakly correlated with correctness in the tail of hard tasks. This is visible in domains like factual QA (where confident but subtly wrong answers can beat cautious correct ones), code synthesis (where plausible-looking code with hidden bugs can be preferred over minimal correct patches), and mathematical reasoning (where long derivations can mask a single invalid step). Second, the model learns these preferences quickly: training on such labels reliably increases stylistic compliance and perceived helpfulness, but can also increase the rate at which the system produces “credible” errors. From a safety perspective, the concerning part is not merely that errors exist; it is that the oversight process can be systematically steered toward endorsing them.

We see a closely related pattern in tool-augmented and citation-heavy settings. Retrieval and browsing can improve accuracy, but they also introduce new failure modes: cherry-picked evidence, misquoted sources, and citations that are syntactically plausible but semantically irrelevant. Lightweight verification (e.g., checking that a link exists, or that a quoted string appears on a page) is often insufficient to establish that the cited material actually supports the claim. In operational terms, this creates a spectrum between genuinely verifiable evidence and “evidence-shaped tokens” that pass superficial checks. Empirically, models can exploit this gap without any explicit deceptive intent: an uncertain generator can produce a tidy story with plausible references simply because that is what the training objective rewards. Oversight then risks becoming a detector of polish rather than truth, especially under time budgets where the evaluator can only spot-check.

The open-role aspect matters here. Many early studies of debate-like protocols and critique focus on fixed-role settings (assigning sides randomly, or training an agent to argue a position regardless of belief). Deployed assistants, by contrast, choose their stance endogenously: they answer with what their internal heuristics and incentives make salient, and they present the supporting material that makes their chosen stance most persuasive under the evaluation rubric. This is where the amplification risk becomes most acute. If a model tends to be correct on easy instances but wrong on a hard, safety-critical subpopulation, and if the evaluation process is more likely to endorse the model when it sounds confident (or when it supplies plausible-looking support), then training and governance loops can overweight precisely those wrong outputs that are hardest to detect. In other words, the dataset of “approved” outputs can become adversarially filtered by the model’s own presentation strategy.

Against this backdrop, adversarial elicitation methods—broadly, prompting or training a second system to critique, rebut, or propose an alternative—have shown partial mitigation. In many practical workflows, adding a critique step improves error detection: reviewers are better at checking a concrete alleged flaw than at validating an entire argument from scratch, and critiques can direct attention to the single brittle step in a long chain. Similarly, paired-comparison protocols that force a judge to choose between competing answers can reduce some superficial biases, because both options are fluent and the judge must attend to discriminating features. We can interpret these empirical gains as creating an additional, truth-correlated signal that is less under the proponent's control. When the original answer is wrong, a competent critic has more "surface area" to identify an inconsistency, a missing assumption, or a failed test; when the original answer is right, the critic must either concede or manufacture a plausible counterstory, which is often harder under any meaningful verification standard.

However, the mitigation is incomplete and comes with a predictable tradeoff. Critics can be noisy, overly skeptical, or strategically unhelpful; even well-intentioned critiques can degrade decision quality by increasing perceived ambiguity. In human oversight, this appears as an increased rate of "both seem plausible" outcomes and greater reliance on tie-breaking heuristics (deference to authority, stylistic preferences, or the default recommendation). In model-based oversight, it can manifest as preference volatility: the judge model may latch onto whichever side better matches its learned reward features, rather than whichever is correct. Empirically, adding an adversary can therefore reduce the frequency with which wrong confident answers are accepted, while also reducing the frequency with which correct answers are confidently approved. This is not a failure of adversarial protocols per se; it is the expected outcome when the judge is weak and the verification technology is imperfect.

The feedback-loop implications make this tradeoff central for alignment. In iterative training, the system does not merely incur one-shot loss from an incorrect label; it updates. If wrong endorsements are fed back as high-reward examples or as "ground truth" preference pairs, the model can be pushed toward policies that more reliably elicit those endorsements. This is a structural route to reward hacking and specification gaming: not necessarily by explicitly lying, but by learning the correlates of approval under the evaluation process. In deployment governance, the analogous dynamic is institutional: approvals become precedents, incident filters, or compliance evidence. A protocol that occasionally fails to approve a correct action is undesirable but often recoverable; a protocol that reliably approves an incorrect action creates a compounding hazard, because it legitimizes and propagates the error. Thus, even modest changes in the conditional wrong-endorsement rate can have outsized long-run effects relative to small changes in average accuracy.

These observations also clarify why "improve the judge" is an insufficient slogan. Better judges help, but the limiting factor is frequently not raw competence; it is attention, cost, and the structure of what is observable. In many organizations, the oversight bottleneck is fixed (one reviewer, limited time, limited tooling), and the system being overseen is adaptive. Under those constraints, protocol design is one of the few levers that can systematically change what information reaches the decision-maker. Adversarial elicitation, cross-checks, and structured comparisons can be understood as institutional ways of reallocating scarce verification effort toward the marginal cases where it is most valuable: situations where the proponent is wrong but persuasive.

This motivates a simple modeling move in the next section. Rather than attempting to capture the full richness of natural-language argumentation, we will isolate a minimal "evidence channel" whose empirical analogue is familiar: sometimes a stance comes with a verifiable artifact (a correct calculation that checks out, a source that actually supports the claim, a reproduced experiment), and sometimes it comes only with plausibility. The key parameter is the rate at which false stances can still produce artifacts that pass the verification bar. Once we represent that asymmetry, we can ask a crisp institutional question suggested by the empirical backdrop: when does adding an adversarial counter-signal reduce the probability that oversight endorses the system exactly when it is wrong, and how does this depend on fabrication feasibility, tie-breaking bias, and the salience of verified evidence?

A minimal persuasion game with a verifiable-evidence channel. To make the amplification concern in the previous discussion analytically tractable, we isolate a single structural asymmetry that repeatedly appears in scalable oversight: *true* stances can typically be supported by some checkable artifact, while *false* stances can sometimes be made to *look* checkable, but only at some limited rate. We represent this as a binary-state persuasion game in which the only payoff-relevant public information is whether each side produces *verifiable evidence*—an abstraction of unit tests that pass, citations that genuinely entail the claim, calculations that validate, or tool traces that actually reproduce the stated outcome.

State, prior, and (optional) endogenous mistakes. Nature draws a truth state $\omega \in \{0, 1\}$ with prior $\pi = \Pr(\omega = 1)$. The protagonist (our "open-role" model) chooses a stance $a \in \{0, 1\}$ to advocate. In the most stripped-down version, a may be exogenous (e.g., fixed by a user query instance). For alignment-motivated reading, we often want mistakes to arise endogenously from imperfect internal cognition rather than strategic lying; a convenient reduced form is that the protagonist observes a private signal $s \in \{0, 1\}$ satisfying

$$\Pr(s = \omega) = p \in (1/2, 1),$$

and then sets $a = s$. This "open-role" discipline forces the model to argue what it (noisily) believes is correct, while still allowing $a \neq \omega$ on hard instances. The antagonistic agent (when present) does not observe ω either; its role is defined by the protocol, not by independent epistemic access.

Evidence technology: feasibility versus effort. After a stance is fixed, a sender may (attempt to) produce a verifiable artifact supporting its stance. We represent this by an indicator $e_i \in \{0, 1\}$ for each sender i. The core assumption is an asymmetry in *feasibility*:

$$\Pr(e_i = 1 \mid \text{stance}_i = \omega) = 1, \qquad \Pr(e_i = 1 \mid \text{stance}_i \neq \omega) = \rho \in (0, 1).$$

Thus if a sender argues the true stance, it can always find or construct something that survives the verification bar (perhaps after some search). If it argues the false stance, it can only sometimes produce something that passes the bar—capturing cherry-picking, misleading but technically "valid" citations, or loopholes in lightweight checking. The parameter ρ is therefore our reduced-form *fabrication / plausible-evidence rate*. Institutional changes (stronger verification tools, better cross-examination, or higher evidentiary standards) will be modeled as lowering ρ.

We additionally allow evidence production to carry an optional cost $c_e \geq 0$. This cost stands in for time, compute, or opportunity cost: even if evidence is feasible, producing it may not be worthwhile when the judge is inattentive. For most of the comparative statics we care about, the critical feature is not effort per se but the asymmetry in feasibility; nevertheless, including c_e clarifies that "always feasible" need not mean "always produced".

Agents and objectives. We separate three roles: (i) the *senders* (protagonist P, and in debate an antagonist A); (ii) a *judge* J who maps the transcript summary into a label $y \in \{0, 1\}$; and (iii) a *principal/designer* D who chooses the protocol and institutional parameters (e.g., verification strength) to optimize downstream welfare.

The judge is accuracy-motivated:

$$u_J(y, \omega) = \mathbf{1}\{y = \omega\},$$

but may be weak, biased, or noisy due to limited attention and imperfect interpretation of evidence. The senders are persuasion-motivated. In the simplest form, the protagonist values endorsement of its stance and pays any evidence cost:

$$u_P = \mathbf{1}\{y = a\} - c_e e_P.$$

In open debate, the antagonist is assigned the opposite stance $1 - a$ and similarly prefers the judge to output $1 - a$:

$$u_A = \mathbf{1}\{y = 1 - a\} - c_e e_A.$$

This is intentionally stylized: we are not assuming the antagonist is intrinsically truthful, only that its incentives oppose the protagonist in the immediate interaction. In oversight applications, this corresponds to a critique model trained to find flaws (or an adversarial prompt) whose "reward" is winning the comparison, not epistemic virtue.

Protocols: open consultancy versus open debate. The institutional difference we study is whether the judge hears only the protagonist (consultancy) or also an assigned adversary (debate). Formally:

- **Open consultancy ($\mathcal{P}$ = OpenConsult).** The protagonist selects (or is committed to) stance a, chooses e_P, and the judge observes (a, e_P) (in the baseline abstraction; richer transcripts collapse to these indicators). There is no antagonist: equivalently set $e_A \equiv 0$.
- **Open debate ($\mathcal{P}$ = OpenDebate).** The protagonist advocates a. An antagonist is assigned the opposite stance $1-a$ and may produce evidence e_A; the protagonist may produce e_P. The judge observes (a, e_P, e_A). We take the assignment "antagonist argues $1-a$" as a reduced-form of protocols that enforce explicit counter-argument (paired answers, critique-and-response, or adversarial debate), ensuring that when the protagonist is wrong the other side is mechanically aligned with the truth.

The timing is: Nature draws ω; (optionally) the protagonist observes s and sets $a = s$; then senders choose evidence subject to feasibility; the judge outputs y. The principal chooses $\mathcal{P}$ and verification institutions outside the one-shot interaction.

Judge mapping from evidence to labels. We want a judge rule that is weak enough to be realistic but structured enough to analyze. The baseline is a monotone evidence-comparison rule with a tie-break bias. In debate, the judge prefers the side with strictly more verifiable evidence and otherwise breaks ties in favor of the protagonist with probability $b \in [0,1]$:

$$y = \begin{cases} a & \text{if } e_P > e_A, \\ 1-a & \text{if } e_P < e_A, \\ a \text{ w.p. } b, \quad 1-a \text{ w.p. } 1-b & \text{if } e_P = e_A. \end{cases}$$

In consultancy, we set $e_A \equiv 0$, so the judge effectively follows the protagonist whenever it supplies evidence, and otherwise defaults via the same tie-breaking baseline.

Because many evaluation systems behave more like probabilistic classifiers than hard rules, we also keep in view a smooth "logit attention" extension:

$$\Pr(y = a \mid e_P, e_A) = \sigma\Big(\lambda(\mu + \kappa(e_P - e_A))\Big), \qquad \sigma(z) = \frac{1}{1+e^{-z}}.$$

Here $\kappa > 0$ captures how salient verified evidence is, $\lambda > 0$ captures how sharply the judge responds to the evidence difference (higher λ means less noise), and $\mu \in \mathbb{R}$ is an additive bias favoring (or disfavoring) the protagonist. This extension lets us express, in a single parametric form, both "inattentive judges" (λ small) and "evidence-focused judges" (κ large).

The target failure mode: mistake amplification. Our object of interest is not raw accuracy alone but the conditional probability that oversight endorses the protagonist exactly when the protagonist is wrong. We define the *mistake-amplification rate* under protocol $\mathcal{P}$ as

$$\mathsf{Amp}_{\mathcal{P}} := \Pr(y = a \mid a \neq \omega).$$

This conditional is the natural statistic for feedback loops: when labels are used as training targets or as governance approvals, the most damaging events are those in which a wrong stance is *validated* (rewarded, ratified, or written into precedent). By contrast, a protocol that sometimes rejects correct stances can be costly in a one-shot sense, but it does not create the same kind of self-reinforcing pressure toward persuasive error.

For completeness, we will also refer to the overall accuracy $\mathsf{Acc}_{\mathcal{P}} := \Pr(y = \omega)$ and the "follow-when-correct" rate $\Pr(y = a \mid a = \omega)$, which together clarify the tradeoff between reducing wrong-follow and preserving correct-follow. The key point of the model is that these quantities can move in opposite directions when we add an adversarial sender: competitive persuasion can make the judge more conservative (reducing both types of follow), while still strictly reducing Amp because it injects an additional, truth-correlated evidence channel precisely in the cases where the protagonist is wrong.

What we will analyze next. Given these primitives, we can solve for evidence-production incentives and induced label distributions under each protocol, first in the closed-form tie-break baseline and then in the logit extension. This will let us state transparent conditions under which open debate is *anti-amplifying*—lower Amp than consultancy—and quantify how the gap scales with the fabrication rate ρ, institutional bias (b or μ), and evidence salience (κ, λ).

Equilibrium analysis in the baseline (closed-form) model. Given the primitives above, we now solve the one-shot interaction by backward induction. The key simplification is that the judge does not parse free-form arguments; it only reacts to a low-dimensional *verifiable-evidence* summary. This makes the strategic problem essentially a discrete contest over whether each sender can put a checkable artifact on the table, and when it is worth paying any associated cost.

Step 1: modeling feasibility as an observed "availability type." It is convenient to make the feasibility asymmetry explicit by introducing, for each sender $i \in \{P, A\}$, an evidence-availability variable $t_i \in \{0, 1\}$. Conditional on the stance that i is forced/committed to argue, $t_i = 1$ means there exists some artifact meeting the verification standard, while $t_i = 0$ means no such artifact can be produced.[1] The feasibility assumption can then be written as

$$\Pr(t_i = 1 \mid \text{stance}_i = \omega) = 1, \qquad \Pr(t_i = 1 \mid \text{stance}_i \neq \omega) = \rho.$$

We assume the sender observes its own t_i before choosing whether to incur the (optional) cost c_e to actually *present* the evidence. Formally, in the evidence stage each sender chooses an action $x_i \in \{0, 1\}$ ("submit evidence") and the publicly observed indicator is

$$e_i := t_i x_i \in \{0, 1\}.$$

In particular, if $t_i = 0$ then $e_i = 0$ regardless of what the sender would like to do; if $t_i = 1$, the sender can set $e_i = 1$ by choosing $x_i = 1$ and paying c_e.

Step 2: the judge rule induces a marginal value of evidence. Fix a protocol $\mathcal{P}$. Because the judge is monotone in the evidence difference, evidence has a simple instrumental value: it shifts the probability that the judge outputs the sender's favored label. Hence each sender's best response in the evidence stage is characterized by a comparison

$$x_i(t_i = 1) \in \arg\max_{x \in \{0,1\}} \Pr(\text{win}_i \mid x, \text{ beliefs about opponent}) - c_e x,$$

where "win_P" is the event $\{y = a\}$ and "win_A" is $\{y = 1 - a\}$. Since x_i only matters when $t_i = 1$, equilibrium behavior is driven by a single object: the *marginal win-probability increase* from setting $e_i = 1$ rather than $e_i = 0$ when evidence is available.

Open consultancy: evidence as a gate for endorsement. In consultancy, there is only one sender. For the closed-form analysis that will feed into the main amplification comparisons, we adopt a conservative default: if the protagonist supplies verifiable evidence, the judge outputs $y = a$; if not, the judge outputs $y = 1 - a$.[2] Under this gatekeeping rule, the protagonist's win probability is 1 when $e_P = 1$ and 0 when $e_P = 0$. Therefore, when $t_P = 1$ the marginal value of presenting evidence is exactly 1, and the protagonist's best response is

$$x_P(t_P = 1) = \begin{cases} 1 & \text{if } c_e < 1, \\ 0 & \text{if } c_e > 1, \\ \text{any mix} & \text{if } c_e = 1. \end{cases}$$

[1] This is a reduced form of "search for proof/citation/test" where the agent learns whether it can find something that survives checking. It is also the point at which verification tooling and evidentiary standards enter: strengthening verification shrinks the set of artifacts that qualify, which we capture as lowering ρ.

[2] This corresponds to a "require evidence to approve" oversight posture. One can instead allow a nontrivial default probability of endorsing a without evidence; doing so only changes constants and does not affect the qualitative role of debate as an anti-amplification device under the same feasibility asymmetry.

In the economically relevant regime $c_e \ll 1$ (e.g., attaching a unit test or a tool trace is cheap relative to the value of being endorsed), consultancy collapses to a particularly simple mapping: $e_P = t_P$. Combining this with feasibility gives the induced evidence distribution:

$$\Pr(e_P = 1 \mid a = \omega) = 1, \qquad \Pr(e_P = 1 \mid a \neq \omega) = \rho,$$

which is the sufficient statistic we will use when we compute amplification and accuracy in the next section.

Open debate: a simultaneous evidence subgame with a biased tie-break. In debate, both P and A potentially submit evidence. The baseline judge rule implies that evidence matters only through the comparison of e_P and e_A, with a tie broken in favor of the protagonist with probability b. Fix any beliefs about the opponent's probability of ending up with $e_{-i} = 1$. If the protagonist has available evidence ($t_P = 1$), then presenting it yields win probability

$$\Pr(y = a \mid e_P = 1) = \Pr(e_A = 0) \cdot 1 + \Pr(e_A = 1) \cdot b = 1 - (1 - b) \Pr(e_A = 1),$$

whereas withholding it yields

$$\Pr(y = a \mid e_P = 0) = \Pr(e_A = 0) \cdot b + \Pr(e_A = 1) \cdot 0 = b\big(1 - \Pr(e_A = 1)\big).$$

Hence the marginal value of presenting evidence (conditional on $t_P = 1$) is

$$\Delta_P \ := \ \Pr(y = a \mid e_P = 1) - \Pr(y = a \mid e_P = 0) \ = \ (1-b) + (2b-1) \Pr(e_A = 1).$$

Symmetrically, if the antagonist has available evidence ($t_A = 1$), the marginal value of presenting it is

$$\Delta_A \ := \ \Pr(y = 1-a \mid e_A = 1) - \Pr(y = 1-a \mid e_A = 0) \ = \ b + (1-2b) \Pr(e_P = 1).$$

Two observations are worth highlighting because they explain why the closed-form analysis remains tractable.

(i) Low-cost region implies "present whenever feasible." For any belief about $\Pr(e_A = 1) \in [0, 1]$, we have $\Delta_P \in [\min\{b, 1-b\}, \max\{b, 1-b\}]$; likewise $\Delta_A \in [\min\{b, 1-b\}, \max\{b, 1-b\}]$. Therefore, if

$$c_e < \min\{b, 1-b\},$$

then both senders strictly prefer to present evidence whenever $t_i = 1$, independent of higher-order beliefs. In this region, evidence submission is effectively non-strategic: $e_i = t_i$. This is the regime we emphasize because it captures the deployment-motivated case where producing checkable support is cheap compared to winning the adjudication, and it isolates amplification effects coming from *feasibility* (ρ) rather than from effort constraints.

(ii) Correlated feasibility implies truth-aligned asymmetry in realized transcripts. Even when both agents use the same "present if feasible" rule, the joint distribution of (e_P, e_A) is sharply different depending on whether the protagonist is correct. Condition on the stance a and the state ω. If $a = \omega$, then the protagonist is on the true side and $t_P = 1$ surely, while the antagonist is on the false side and has $t_A = 1$ only with probability ρ. Thus

$$\Pr(e_P = 1, e_A = 0 \mid a = \omega) = 1 - \rho, \qquad \Pr(e_P = 1, e_A = 1 \mid a = \omega) = \rho.$$

If instead $a \neq \omega$, the roles reverse: the antagonist is now on the true side and so $e_A = 1$ surely, while the protagonist has $e_P = 1$ only with probability ρ:

$$\Pr(e_P = 0, e_A = 1 \mid a \neq \omega) = 1 - \rho, \qquad \Pr(e_P = 1, e_A = 1 \mid a \neq \omega) = \rho.$$

These two cases make clear where the *tie region* comes from. In debate, ties occur only when the *false* side manages to produce evidence (probability ρ); when the false side fails, the transcript exhibits a strict evidence imbalance favoring the true side. This is exactly the structural lever by which adding an antagonist can reduce conditional wrong-follow: it converts many would-be "no-evidence" consultancy outcomes into "strictly-more-evidence-for-truth" debate outcomes.

A note on intermediate and high evidence costs. When c_e exceeds $\min\{b, 1 - b\}$, evidence submission is no longer dominant, and equilibrium can depend on beliefs about the opponent's propensity to present evidence. In that regime, two effects can appear: (a) senders sometimes withhold evidence even when available, reducing both accuracy and the strength of the anti-amplification mechanism; and (b) availability t_i becomes behaviorally relevant as a private signal about whether one is on the true side, because observing $t_i = 0$ certifies that the stance is false under our feasibility asymmetry. While these extensions are important for modeling compute-bounded oversight or costly tool use, we will keep the main exposition focused on the low-cost region where $e_i = t_i$, since it yields closed-form expressions that cleanly separate institutional parameters (ρ, b) from idiosyncratic effort considerations.

Summary of the equilibrium objects used downstream. Under the baseline closed-form assumptions and in the low-cost region, the equilibrium of the evidence stage reduces to a deterministic mapping from (a, ω) into a distribution over evidence indicators. Consultancy produces a single indicator e_P with $\Pr(e_P = 1 \mid a \neq \omega) = \rho$, while debate produces a pair (e_P, e_A) in which, conditional on $a \neq \omega$, the true-side antagonist has $e_A = 1$ surely and the false-side protagonist reaches the tie case only with probability ρ. The next section plugs these equilibrium transcript distributions into the judge rule to obtain mistake amplification, accuracy tradeoffs, and the principal's welfare comparisons across protocols.

3 Main results: anti-amplification guarantee, welfare implications, and when debate trades off with raw accuracy

Why we focus on conditional wrong-follow rather than only accuracy. In oversight and training pipelines, the most damaging failures are often not "random" misclassifications, but cases where the process *confidently ratifies* an incorrect stance and thereby makes the mistake legible, reusable, and self-reinforcing (e.g., as a training label, a policy justification, or a cached chain-of-thought). This motivates our central safety statistic

$$\mathsf{Amp} \;:=\; \Pr(y = a \mid a \neq \omega),$$

the probability the judge endorses the protagonist *precisely when the protagonist is wrong.* Importantly, Amp is a conditional quantity: it measures how the institution behaves on the rare-but-critical slice where the model has already erred. Our main claim is that adding an adversarial counterparty with the same verification technology provides a structural "circuit breaker" against such ratification, even if the judge is noisy or biased.

Result 1 (baseline tie-break model): debate strictly reduces mistake amplification. Under the low-cost equilibrium characterized in the previous section, the only randomness in transcripts comes from feasibility: the false side can produce verifiable evidence only with probability ρ. Conditioning on $a \neq \omega$, consultancy offers the judge exactly one relevant artifact: if the wrong protagonist manages to produce verifiable evidence, endorsement follows; otherwise it does not. This yields the closed form

$$\mathsf{Amp}_{\text{consult}} = \rho.$$

In open debate, the key difference is that when $a \neq \omega$ the antagonist is on the true side and hence can produce verifiable evidence surely; thus the judge endorses the wrong stance only if the protagonist (i) successfully produces verifiable evidence despite arguing a falsehood and (ii) then wins the tie via bias. Hence

$$\mathsf{Amp}_{\text{debate}} = \rho b.$$

Therefore, for any $\rho \in (0, 1)$ and any $b < 1$,

$$\mathsf{Amp}_{\text{debate}} < \mathsf{Amp}_{\text{consult}}, \qquad \mathsf{Amp}_{\text{consult}} - \mathsf{Amp}_{\text{debate}} = \rho(1 - b).$$

The interpretation is mechanistic rather than informational: debate does not require the judge to infer who is right from rhetoric; it only requires the judge to be even weakly responsive to the presence of *relative* verifiable support. Adding an antagonist converts many "no-evidence" consultancy outcomes (which default to endorsing the protagonist under our gatekeeping rule) into "evidence imbalance" debate outcomes that favor the true side.

Result 1' (logit judges): anti-amplification is robust to noise and baseline bias. The tie-break rule is analytically convenient but brittle as a behavioral model of adjudication. If instead the judge follows the logit response

$$\Pr(y = a \mid e_P, e_A) = \sigma\big(\lambda(\mu + \kappa(e_P - e_A))\big),$$

then conditioning on $a \neq \omega$ yields

$$\mathsf{Amp}_{\text{consult}} = \rho\,\sigma(\lambda(\mu + \kappa)) + (1 - \rho)\,\sigma(\lambda\mu),$$

$$\mathsf{Amp}_{\text{debate}} = \rho\,\sigma(\lambda\mu) + (1 - \rho)\,\sigma(\lambda(\mu - \kappa)).$$

Because $\kappa > 0$ implies the strict ordering

$$\sigma(\lambda(\mu + \kappa)) > \sigma(\lambda\mu) > \sigma(\lambda(\mu - \kappa)),$$

we obtain $\mathsf{Amp}_{\text{debate}} < \mathsf{Amp}_{\text{consult}}$ for all $\mu \in \mathbb{R}$ and all $\rho \in (0, 1)$. Intuitively, when the protagonist is wrong, debate shifts the evidence differential leftward (toward the truth) in exactly the branch where consultancy would otherwise be maximally permissive: the branch where the protagonist cannot fabricate verifiable evidence.

Result 2: debate can reduce "follow-when-correct," creating an accuracy tradeoff. Anti-amplification is not free. Conditioning on $a = \omega$, the antagonist is now forced onto the false side. With probability ρ it can nonetheless produce verifiable counterevidence, creating a tie (baseline) or reducing the evidence differential (logit). In the baseline model this gives

$$\Pr(y = a \mid a = \omega,\ \text{consult}) = 1, \qquad \Pr(y = a \mid a = \omega,\ \text{debate}) = 1 - \rho(1 - b).$$

Thus debate weakly reduces the probability the judge endorses a *correct* protagonist. The same qualitative effect holds in the logit model: debate introduces a second strategic sender who sometimes successfully manufactures enough plausibility to partially offset true evidence, and a noisy judge cannot perfectly ignore that offset.

From conditional to unconditional performance: closed-form accuracy comparisons. To connect these conditional statements to overall accuracy, we introduce a protagonist correctness rate $p := \Pr(a = \omega)$ (e.g., arising from an underlying private signal). In the baseline model, consultancy misclassifies only when the protagonist is wrong and successfully produces verifiable evidence, so

$$\Pr(y \neq \omega \mid \text{consult}) = (1-p)\rho, \qquad \Pr(y = \omega \mid \text{consult}) = 1 - (1-p)\rho.$$

In debate, errors arise both when a correct protagonist loses a fabricated tie and when an incorrect protagonist wins a fabricated tie:

$$\Pr(y \neq \omega \mid \text{debate}) = p\,\rho(1-b) + (1-p)\,\rho b = \rho\big(p + b - 2pb\big).$$

Consequently,

$$\Pr(y = \omega \mid \text{consult}) - \Pr(y = \omega \mid \text{debate}) = \rho(2p-1)(1-b),$$

which is strictly positive whenever $p > 1/2$ and $b < 1$. This makes the tradeoff sharp: debate strictly improves the conditional safety metric Amp, but it can strictly worsen raw one-shot accuracy because the antagonist sometimes "muddying the waters" harms the many cases where the protagonist was already correct.

Result 3: debate is welfare-superior when amplification is sufficiently costly. We formalize the "amplification matters more than generic error" intuition using the principal's reduced-form one-step objective

$$U_D = -\alpha \Pr(y \neq \omega) - \beta \Pr(y = a,\ a \neq \omega) - C(\mathcal{P}),$$

where β captures the downstream cost of ratifying wrong stances (e.g., compounding training errors), and $C(\mathcal{P})$ captures operational overhead (running two models, longer deliberation, etc.). In the baseline model,

$$\Pr(y = a,\ a \neq \omega \mid \text{consult}) = (1-p)\rho, \qquad \Pr(y = a,\ a \neq \omega \mid \text{debate}) = (1-p)\rho b.$$

Let $\Delta C := C(\text{OpenDebate}) - C(\text{OpenConsult})$. Algebra yields

$$U_D(\text{debate}) - U_D(\text{consult}) = \rho(1-b)\Big[\beta(1-p) - \alpha(2p-1)\Big] - \Delta C.$$

Thus, for negligible protocol-cost differences ($\Delta C \approx 0$), debate is preferred whenever

$$\beta > \alpha\,\frac{2p-1}{1-p}.$$

This threshold illustrates the central design logic. When the protagonist is usually correct (p close to 1), the accuracy loss from debate is paid often, while amplification events are rare; debate requires a correspondingly large β to be justified. Conversely, in regimes where the protagonist is only moderately reliable, or where the institution is particularly allergic to ratifying wrong answers (large β), open debate becomes optimal on welfare grounds even if one-shot accuracy declines.

Comparative statics: when the anti-amplification gap is largest. The baseline gap $\mathsf{Amp}_{\text{consult}} - \mathsf{Amp}_{\text{debate}} = \rho(1-b)$ makes two monotonicities immediate. First, the value of debate grows with ρ: if false stances can often be dressed up with plausible-looking support, then a single-speaker process is especially vulnerable to being led astray, while a two-speaker contest still forces the false side to at least *match* the true side's evidence to have any chance. Second, the value of debate shrinks as the judge becomes more protagonist-favoring (higher b or, in the logit model, higher μ): if ties (or near-ties) are resolved in favor of the protagonist, then debate cannot fully prevent ratification in the tie region, and the mechanism reduces to "how often can the wrong side force a tie?" In the logit model, higher evidence weight κ and higher attention λ strengthen the gap by making the judge more sensitive to the truth-favoring evidence differential that debate reliably creates in the $(1-\rho)$ branch.

Limitations and what we treat as exogenous. Our guarantee relies on a particular asymmetry: verifiable evidence for the true stance is always feasible, while verifiable evidence for the false stance is feasible only with probability ρ. In practice, both directions can fail (e.g., true claims without accessible proofs; false claims with cherry-picked but technically valid citations), so ρ should be read as an endogenous institutional statistic determined by the verification standard, tool access, and the distribution of tasks. We also treat the antagonist as forced to take the opposite stance. This captures "red-team the current proposal" oversight, but it omits richer equilibria where agents strategically choose which claim to argue, collude, or flood the judge with low-salience artifacts. These are precisely the settings where mechanism design details matter.

Lead-in: from guarantees to design levers. Taken together, the results suggest a practical decomposition: (i) protocol choice (consultancy versus debate) controls whether the judge ever observes an evidence differential aligned with truth when the protagonist is wrong; (ii) institutional parameters (ρ, effective tie-breaking bias, and evidence salience) control how often the wrong side can manufacture ties and how decisively the judge reacts to imbalances. The next section turns this into mechanism design guidance: which knobs (verification strength, cross-examination, jury aggregation, and explicit tie-breaking rules) provably reduce Amp, and which implementable metrics can track these improvements in deployments.

4 Mechanism design: levers that strengthen anti-amplification and how to measure progress

Our basic comparative result is intentionally austere: it isolates a single structural property (a second sender constrained by the same verification technology) and shows that it reduces conditional wrong-follow. The natural next question for a designer is therefore not whether debate *can* help, but which institutional parameters most efficiently improve the anti-amplification guarantee, and how we can detect that improvement in a live pipeline without assuming access to ω at scale.

Design objective: minimize Amp via institutional parameters rather than "better judges." In deployments, we often cannot rely on adjudicators (human or model) being perfectly calibrated, unbiased, or even attentive. Our mechanism-design stance is to treat judge behavior as partially exogenous (captured by b or (μ, λ)) and to intervene primarily through *what evidence gets produced and how it is presented.* Concretely, the principal can choose: (i) the verification and citation standard that determines the false-evidence feasibility ρ; (ii) the interface or scoring rule that determines evidence salience κ; (iii) the transcript-to-label aggregation rule that determines effective tie-breaking bias (baseline b, or logit baseline μ); and (iv) whether to aggregate across multiple judges (jury size n). Each lever has an interpretable comparative static: it changes how frequently the false side can force a tie, and how likely a tie (or near-tie) is resolved in favor of the wrong stance.

Lowering ρ: verification strength, cross-examination, and tool access. In our model, ρ summarizes the probability that a false stance can nevertheless present *verifiable-looking* evidence. This parameter is not a property of the agent alone: it is an endogenous statistic of the institution. Strengthening verification reduces ρ by shrinking the set of artifacts that count as "verifiable." Examples include requiring citations to primary sources, running automated retrieval-and-checking against an allowlist, enforcing quote-level provenance, and adversarially validating cited passages. Cross-examination in debate is a particularly direct way to reduce effective ρ: if the antagonist can challenge the provenance or relevance of evidence, then many would-be fabrications cease to be admissible (or are down-weighted), which can be modeled as lowering ρ and/or increasing κ for truly discriminative evidence.

A subtle but practically important point is that ρ is task-dependent: even truthful stances may sometimes lack accessible proofs, while false stances may cherry-pick technically correct but misleading citations. One robust design move is therefore to define evidence admissibility *procedurally* (e.g., "must be checkable by this toolchain within this budget"), so that ρ becomes a controllable function of the verification budget and the allowed corpus, rather than a vague epistemic notion.

Increasing κ: making evidence differentials decisive rather than decorative. The logit extension makes explicit that anti-amplification improves when judges respond more strongly to evidence imbalances. Holding ρ fixed, increasing κ enlarges the separation between $\sigma(\lambda(\mu+\kappa))$, $\sigma(\lambda\mu)$, and $\sigma(\lambda(\mu-\kappa))$, thereby increasing the debate–consultancy amplification gap. Mechanistically, κ can be interpreted as an *institutional* weight placed on verifiable support: evidence can be surfaced, normalized, and scored so that it dominates rhetorical features.

In practice, raising κ corresponds to interventions like: (i) hiding unverified text by default and prominently displaying verified citations; (ii) forcing both sides to submit a small number of checkable claims rather than a long narrative (reducing dilution); (iii) requiring minimal, standardized evidence objects (hash-linked excerpts, theorem prover certificates, unit tests) that the judge can quickly compare; and (iv) using decision rules that explicitly depend on evidence deltas (e.g., "prefer the side with more independently verified items") rather than on holistic impressions. Importantly, κ is not the same as judge competence: it is the marginal influence of admissible evidence on the final label, which can be increased even for noisy judges by structuring the UI and the allowed actions.

Reducing effective bias: explicit tie-breaking rules and symmetry constraints. In the baseline model, the only channel through which a wrong protagonist can win in debate is via ties resolved by b. This makes tie-breaking policy a first-class governance object. A simple prescription follows directly: implement tie-breaking rules that are *stance-symmetric* and, when possible, protagonist-agnostic. In institutional terms this means blinding the judge to agent identity, randomizing which side is labeled "protagonist," or requiring the labeler to commit to a decision policy before seeing the transcript. In the logit model, analogous interventions target μ: remove persistent baseline favoritism by balancing prompts, normalizing formatting, and enforcing symmetric budgets and admissibility criteria.

From a safety perspective, we should treat "tie regions" as hazardous: they are precisely where fabrication converts into endorsement. Thus another design principle is to reduce the mass of ties by requiring discriminative evidence objects (affecting ρ) and by amplifying evidence salience (affecting κ), rather than merely trying to set $b \approx 1/2$ and hoping for neutrality.

Juries and aggregation: converting idiosyncratic bias into controllable error. A standard way to reduce sensitivity to any single judge is to use a panel of n independent judges and aggregate by majority vote. In the baseline tie-break model, if a tie occurs, each judge favors the protagonist with probability b, so the probability that the panel favors the protagonist in a tie is

$$b_n = \sum_{k=\lceil n/2 \rceil}^{n} \binom{n}{k} b^k (1-b)^{n-k},$$

and the amplification rate in debate becomes $\mathsf{Amp}_{\text{debate}} = \rho b_n$. This clarifies a governance tradeoff: panels help if the average tie-bias is not itself protagonist-favoring. If $b < 1/2$, then b_n decreases rapidly in n, giving a strong multiplicative reduction in amplification; if $b > 1/2$, panels amplify the wrong direction. The practical implication is that juries should be paired with explicit debiasing (to enforce $b \approx 1/2$) or designed as heterogeneous ensembles whose biases cancel.

Evidence-production costs and subsidies: making the "right" equilibrium cheap. Thus far we have largely suppressed c_e, but in operational settings evidence is costly (time, tool calls, verification compute). If c_e is high, both sides may rationally choose $e = 0$, collapsing the protocol into a low-information tie that reintroduces bias-driven ratification. The principal can prevent this by subsidizing evidence (lowering effective c_e via tooling), mandating minimal evidence thresholds, or awarding explicit utility for verified artifacts (in RLHF/RLAIF terms, paying for citations rather than for persuasion). In this sense, debate is not only a protocol choice but also a *budget allocation* problem: anti-amplification gains arise when it is cheap for the true side to produce checkable support and hard for the false side to fake it.

Implementable metrics: what to log, estimate, and monitor. We cannot improve what we cannot measure, yet $\mathsf{Amp} = \Pr(y = a \mid a \neq \omega)$ conditions on an event that is often unknown online. We therefore recommend a two-tier measurement strategy.

First, maintain an *offline gold* stream where ω is known (unit tests, curated QA, mechanized theorems, sandboxed environments). On this stream, directly estimate

$$\widehat{\mathsf{Amp}} = \Pr(y = a \mid a \neq \omega), \qquad \widehat{\mathsf{Acc}} = \Pr(y = \omega),$$

separately for consultancy and debate, and track their difference over time as verification tooling and interfaces change. Second, instrument the protocol to estimate intermediate quantities that are observable online and predictive of amplification risk: tie frequency $\Pr(e_P = e_A)$, evidence-imbalance frequency $\Pr(e_P \neq e_A)$, and, when admissibility is automated, an empirical false-evidence feasibility proxy

$$\widehat{\rho} = \Pr(e = 1 \mid \text{forced-false stance tasks}),$$

where “forced-false” tasks are generated by constructing prompts with known negations or by pairing questions with incorrect candidate answers and asking the agent to defend them under the same verification rules. This yields a deployment-relevant estimate of how often the system can make falsehoods look checkable.

Finally, we recommend reporting *calibrated disagreement* statistics: how often judges flip their decision when shown the opposing side’s verified evidence, and how sensitive outcomes are to identity/format randomizations. These are practical surrogates for κ and μ, and they diagnose whether improvements are coming from better evidence differentials (desired) or from brittle presentation effects (undesired).

Summary: a concrete design checklist. From the model, an institution that seeks low amplification should: (i) invest in verification so that ρ is small and measurable; (ii) design the interface so that κ is large—verified artifacts dominate the decision; (iii) enforce symmetry to keep effective b (or μ) near neutral; (iv) use juries only when biases are controlled; and (v) subsidize evidence production to avoid degeneracy into no-evidence ties. These levers are modular and can be tuned independently, which sets up the next step: extending the formalism to richer message spaces and more realistic informational structures where analytic expressions may no longer be closed-form.

5 Extensions (flagged where numerical methods are needed): richer message spaces, correlated signals, multi-option answers, endogenous judge effort/attention

Our baseline model deliberately compresses “argumentation” into a binary, verifiable-evidence indicator. This is useful for isolating the amplification mechanism, but it leaves several deployment-relevant degrees of freedom implicit. Here we outline extensions that preserve the core institutional question—when does adding an adversarial channel reduce conditional wrong-follow?—while moving closer to the structure of real transcripts. In several cases, closed-form expressions cease to exist; we flag where equilibrium characterization is likely to require numerical methods.

Richer message spaces: separating rhetoric from verifiable objects. A first extension is to let each sender choose a message m_i from a large space, where only some components are verifiable. One convenient abstraction is to write $m_i = (v_i, r_i)$, where $v_i \in \{0, 1, \dots, V\}$ counts (or scores) verifiable artifacts and $r_i \in \mathbb{R}$ is an unverified rhetorical score. The judge then uses a decision rule such as

$$\Pr(y = a \mid m_P, m_A) = \sigma\big(\lambda\big(\mu + \kappa_v(v_P - v_A) + \kappa_r(r_P - r_A)\big)\big) \,,$$

with $\kappa_v > 0$ and possibly $\kappa_r \geq 0$. The institutional levers from the baseline become interpretable as shifting κ_v upward (surface verified content) and shifting κ_r downward (de-emphasize style).

The safety-relevant failure mode in this richer model is that a wrong protagonist may win by increasing r_P even when v_P is weak, especially if κ_r is nontrivial. In consultancy, this creates a direct channel for wrong-follow; in debate, the antagonist partially cancels it by allowing r_A to compete. However, the anti-amplification guarantee is no longer purely about ρ: it depends on the joint distribution of (v_i, r_i) under true vs false stances, and on whether the judge treats rhetoric as a substitute for evidence. Analytically, we can often recover monotone comparative statics if (i) v_i first-order stochastically dominates under truth and (ii) r_i is either independent of truth or symmetric across sides conditional on stance. Without such restrictions, establishing $\mathsf{Amp}_{\text{debate}} < \mathsf{Amp}_{\text{consult}}$ becomes an empirical question about the judge model and the induced message distributions.

Endogenous evidence granularity: from binary e to evidence portfolios. Instead of a single indicator e_i, let each sender select a portfolio of claims $\mathcal{C}_i$ and associated proofs/citations $\mathcal{E}_i$, subject to a budget B. Verification maps $(\mathcal{C}_i, \mathcal{E}_i)$ into a set of accepted items and a score V_i. The feasibility parameter ρ generalizes to a function $\rho(B)$ describing how the false side's probability of producing *any* accepted items grows with budget, and (more importantly) how the *distribution* of V_i differs under truth and falsehood. This formulation makes explicit a design tension: increasing B can improve truth-side coverage (helping accuracy) while also increasing the surface area for cherry-picked, misleading-but-verified items (raising effective $\rho(B)$). This is exactly where numerical evaluation is typically needed: the mapping from evidence rules to the induced score distributions is toolchain- and task-dependent, and closed-form reductions are unlikely.

Correlated signals and shared blind spots (numerical methods likely). The baseline allows the protagonist to be wrong without specifying why. In practice, both sides may share correlated errors due to common training data, shared heuristics, or a misleading prompt. To model this, let $(s_P, s_A) \in \{0,1\}^2$ be private signals with joint distribution

$$\Pr(s_P = \omega) = p_P, \qquad \Pr(s_A = \omega) = p_A, \qquad \mathrm{Corr}(\mathbf{1}\{s_P = \omega\}, \mathbf{1}\{s_A = \omega\}) = \eta,$$

or more generally a full joint conditional distribution $\Pr(s_P, s_A \mid \omega)$. In open debate, we can interpret the antagonist's stance as mechanically forced to $1 - a$, but its *evidence ability* and rhetorical effectiveness may depend on its signal s_A: if it internally "believes" the forced stance is wrong, it may be less able to marshal convincing support, even when the forced stance is in fact true (because the protagonist is wrong).

This is the key complication: debate's advantage relies on the antagonist being reliably aligned with the truth when the protagonist is not. With correlated signals, the event $\{a \neq \omega\}$ may coincide with $\{s_A \neq \omega\}$ with non-negligible probability, reducing the quality of the counter-signal. In the extreme case of perfectly correlated errors ($\eta \approx 1$ with similar accuracies), adding an antagonist may do little. Characterizing equilibrium evidence choices in this environment typically requires numerical methods because best responses depend on posterior beliefs about ω given correlated signals and observed transcript features, and because equilibrium may involve mixed strategies over evidence effort contingent on private confidence. Practically, this extension suggests a concrete diagnostic: debate is most valuable when we can induce *diversity* in errors (low η) by varying model families, retrieval corpora, or prompting styles across sides.

Multi-option answers and structured outputs. Many oversight tasks are not binary: the system chooses among $K > 2$ answers, or outputs structured objects (plans, code patches). A natural generalization is to let the protagonist choose $a \in \{1, \ldots, K\}$ and assign an antagonist stance via a protocol rule (e.g., the antagonist attacks a and defends the best alternative it can support, or it is assigned a specific competitor a'). The judge then chooses $y \in \{1, \ldots, K\}$ based on comparative evidence.

Two modeling choices matter. First, how is "opposition" defined? If the antagonist can choose which alternative to defend, we obtain a kind of competitive search over counterexamples, which can be beneficial for catching errors but also creates strategic selection effects (the antagonist may pick the most persuasive alternative rather than the most correct). Second, how does evidence aggregate across multiple options? A multinomial logit,

$$\Pr(y = j \mid \{V_\ell\}_{\ell=1}^{K}) \propto \exp(\lambda(\mu_j + \kappa V_j)),$$

is tractable but still leaves the distribution of $\{V_\ell\}$ as the main object. In multi-option settings, a useful analogue of Amp is the probability the judge outputs the protagonist's proposal conditional on it being suboptimal under a known scoring rule (e.g., higher loss than some alternative), which better matches optimization tasks. Analytical dominance results may still hold when (i) at least one opponent option is truthful and can always produce strong verifiable evidence and (ii) the judge is monotone in evidence scores, but the clean ρb style expressions generally disappear.

Endogenous judge effort, attention, and verification (numerical methods likely). A central deployment question is not only how judges *map* evidence to decisions, but how much effort they exert to read, verify, and deliberate. We can model this by letting the judge choose an attention/effort level $t \geq 0$ with cost $c_J(t)$, where t affects either (i) rationality $\lambda(t)$, (ii) evidence salience $\kappa(t)$, or (iii) the probability that submitted evidence is actually checked (an additional verification stage). For example,

$$\Pr(y = a \mid e_P, e_A, t) = \sigma\big(\lambda(t)\big(\mu + \kappa(e_P - e_A)\big)\big), \qquad u_J = \mathbf{1}\{y = \omega\} - c_J(t).$$

Now the protocol changes the judge's incentives: debate may increase cognitive load (raising the marginal cost of attention) while also increasing the marginal benefit of attention (because there is more diagnostic contrast to exploit). Whether t rises or falls under debate becomes an equilibrium object, and this can flip comparative statics if attention collapses. This extension is a natural place where institutional design and compute budgets interact: a principal may subsidize t (more time per case, better UI, automated verification) or impose constraints that reduce load (limited claims, standardized evidence objects). Solving for equilibria with endogenous t and strategic evidence production by both sides is typically not closed-form and will often require numerical fixed-point methods.

Takeaways and open problems. Across these extensions, the core lesson remains: anti-amplification is strongest when (i) the false side has limited ability to manufacture verified support, (ii) judges respond sharply to verified differentials, and (iii) the adversary provides an informative counter-signal rather than correlated noise. The open problems are correspondingly concrete. Can we characterize sufficient conditions on rich transcript distributions under which adding an antagonist guarantees lower $\Pr(y = a \mid a \neq \omega)$? How should protocols force diversity (reducing signal correlation) without enabling manipulation? And what is the right endogenous-effort model for human or model judges, so that we can design debate formats that increase, rather than dilute, effective attention? These questions motivate the empirical mapping in the next section: once we leave the binary indicator world, measuring ρ, effective bias, and evidence salience from real transcripts becomes the primary way to validate that an extension preserves the intended safety property.

6 Empirical mapping and test plan: estimating ρ, b, κ, and amplification from transcripts

Our theoretical comparisons are only as useful as our ability to *map* them onto measurable objects in real oversight pipelines. In deployment, we can log full transcripts, tool calls, citations, and judge outputs; what we typically lack is a clean, task-independent notion of "verifiable evidence" and a ground-truth label ω. This section sketches a concrete measurement plan that (i) defines transcript-level observables that approximate (e_P, e_A), (ii) proposes identification strategies for ρ, b, and evidence responsiveness parameters (e.g. κ in the logit extension), and (iii) specifies falsifiable predictions across tasks and judge types.

Step 0: operationalize verifiable evidence and transcript summaries. To estimate anything resembling ρ we need a verification technology that is stable across protocols. In practice we recommend defining a deterministic (or audited) verification map $V(\text{transcript}) \in \{0,1\}$ or $V(\text{transcript}) \in \mathbb{R}$ that accepts only *checkable* artifacts: cited passages with cryptographic provenance, tool outputs with recorded inputs/outputs, unit tests for code, or formally checked derivations. For the binary baseline, we can set

$$e_i := \mathbf{1}\{V_i \geq \tau\},$$

where V_i is an evidence score extracted from side i's portion of the transcript and τ is a fixed acceptance threshold. The key design constraint is symmetry: the same acceptance criterion must apply to protagonists and antagonists, and must not depend on the judge decision y. When symmetry is violated (e.g. one side is allowed to cite sources the other cannot access), empirical estimates of ρ and b conflate institutional asymmetries with strategic behavior.

Ground truth ω: benchmarks, adjudication, and partial labels. The cleanest setting is a benchmark suite where ω is known (factual QA, math, theorem proving, unit-tested code). When ω is not directly observable, we can still estimate amplification using (i) delayed outcome labels (e.g. whether a proposed patch passes CI), (ii) expert adjudication on a stratified sample, or (iii) *consensus* labels from an independent, higher-effort panel. Importantly, amplification is a conditional probability $\Pr(y = a \mid a \neq \omega)$; thus even partial ground-truthing is useful if it reliably identifies a subset of "protagonist-wrong" cases. Throughout, we recommend reporting both point estimates and sensitivity intervals under plausible mislabel rates of the ground-truthing procedure.

Estimating Amp nonparametrically under protocol randomization. Given a dataset of instances $t = 1, \ldots, T$ with known ω_t, randomize the protocol $\mathcal{P}_t \in \{\text{OpenConsult}, \text{OpenDebate}\}$ at the instance level, and log $(a_t, e_{P,t}, e_{A,t}, y_t)$. Then the amplification rate is estimated by the sample analogue

$$\widehat{\mathsf{Amp}}_{\mathcal{P}} = \frac{\sum_{t:\mathcal{P}_t=\mathcal{P}} \mathbf{1}\{y_t = a_t\}\mathbf{1}\{a_t \neq \omega_t\}}{\sum_{t:\mathcal{P}_t=\mathcal{P}} \mathbf{1}\{a_t \neq \omega_t\}}.$$

This estimator requires no structural assumptions about sender optimization or the judge rule; it is a direct measurement of the safety-relevant event "the oversight label endorses the wrong stance." We can then test the directional prediction $\widehat{\mathsf{Amp}}_{\text{debate}} < \widehat{\mathsf{Amp}}_{\text{consult}}$ using standard stratified bootstrap confidence intervals (stratifying by task family and difficulty to reduce variance).

A complementary decomposition that we have found informative in practice is to report both (i) $\widehat{\Pr}(y = \omega)$ (overall accuracy) and (ii) $\widehat{\Pr}(y = a \mid a = \omega)$ (follow-when-correct), since the theory predicts that debate can reduce amplification even when it slightly reduces follow-when-correct.

Estimating ρ: feasibility versus equilibrium effort. In the model, ρ is a *feasibility* parameter: the probability that a false stance can produce verifiable evidence. Real systems add an additional layer: even if feasible, a sender may not expend the effort to find or present it. Empirically, we suggest estimating two related quantities.

First, an *equilibrium* fabrication rate,

$$\widehat{\rho}_{\text{eq}} := \widehat{\Pr}(e_i = 1 \mid \text{stance}_i \neq \omega),$$

computed separately for protagonists and antagonists and separately by protocol. This captures the effective rate at which the false side *does* show accepted evidence under the imposed budgets and incentives. It is the object that directly enters amplification in the baseline closed-form expressions.

Second, a closer proxy to *feasibility* is obtained by removing strategic selection: give each side a fixed evidence-search budget and require an "attempt" step (e.g. mandatory tool call to a retrieval+verification system) whose output is logged, with no penalty for attempting. Under such a forced-attempt design, ρ becomes the success frequency of the verification system conditional on the stance being false:

$$\widehat{\rho}_{\text{feat}} := \widehat{\Pr}(\text{verifier accepts at least one item} \mid \text{stance} \neq \omega, \text{ attempt made}).$$

The gap $\widehat{\rho}_{\text{feat}} - \widehat{\rho}_{\text{eq}}$ is itself a diagnostic of incentive issues (implicit c_e, time pressure, or UI friction) that can be adjusted by the principal.

Estimating tie-bias b and baseline favoritism. In the baseline tie-break model, $b = \Pr(y = a \mid e_P = e_A)$ (with the conditioning interpreted within the relevant protocol). In debate, the cleanest estimate is therefore

$$\widehat{b} \ = \ \widehat{\Pr}(y = a \mid e_P = e_A, \ \mathcal{P} = \text{OpenDebate}).$$

However, raw ties may differ systematically from non-ties (harder questions, more rhetorical content, etc.), so we recommend two robustness refinements: (i) compute $\widehat{b}$ within strata of task type and difficulty, and (ii) implement a *role-swap* or *label-blind* variant where the judge is not told which side is the protagonist (or where the protagonist assignment is randomized and not shown), allowing us to distinguish intrinsic argument ordering effects from explicit role bias. If role labels matter, we should see shifts in $\widehat{b}$ under label-blinding even when $e_P = e_A$ is held fixed.

Estimating evidence responsiveness (κ, λ, μ) from transcript features. For model or human judges whose behavior is noisy, we can fit a reduced-form logit choice on transcript summaries. In the simplest binary-evidence encoding,

$$\Pr(y = a \mid e_P, e_A) = \sigma(\lambda(\mu + \kappa(e_P - e_A)))\,.$$

Because (λ, μ, κ) are not all separately identified without additional variation (only products like $\lambda\kappa$ appear), we typically fix $\lambda = 1$ and interpret the fitted coefficients as *effective* salience and bias. More informative identification becomes possible if we use a graded evidence score $V_i \in \mathbb{R}$ (e.g. number of accepted citations, verifier confidence, proof length), yielding

$$\Pr(y = a \mid V_P, V_A) = \sigma(\tilde{\mu} + \tilde{\kappa}(V_P - V_A))\,,$$

which can be estimated by standard logistic regression with clustered standard errors by judge. To separate "attention" from "preferences," we can experimentally vary cognitive load (time limits, transcript length caps) and test whether $\tilde{\kappa}$ attenuates under load, consistent with a drop in effective rationality/attention.

Predicted empirical patterns across tasks, judges, and institutions. The model makes several qualitative predictions that can be checked without committing to exact functional forms. (i) *High-ρ domains* (open-web factual tasks, ambiguous scientific claims, domains with many plausible citations) should exhibit larger consultancy amplification and therefore larger debate gains. (ii) *High-bias judges* (large $\widehat{b}$ in ties, or large $\tilde{\mu}$) should show weaker debate advantages, with debate potentially failing when $\widehat{b} \approx 1$. (iii) *Higher evidence salience* ($\tilde{\kappa}$ increased by better UI, highlighting verified passages, or stricter acceptance thresholds) should increase the debate–consultancy amplification gap. (iv) *Juries* should reduce idiosyncratic tie effects when individual-level biases vary across judges; empirically, this appears as a shrinkage of variance in tie outcomes and, when average bias is not extreme, a reduction in estimated $\Pr(y = a \mid a \neq \omega)$ under debate.

A particularly important failure mode to monitor is when both sides systematically present "verified but misleading" artifacts, inflating $e_P = e_A = 1$ tie frequency. Empirically, this manifests as a rising tie rate and a debate amplification rate that becomes dominated by $\widehat{b}$. When this occurs, the correct response is not to abandon measurement but to tighten the verification map $V(\cdot)$ (lowering effective ρ) or to redesign evidence objects so that verification tracks the *claim* being argued rather than isolated true statements.

Summary: what success looks like in data. A successful empirical validation will not merely show that debate increases accuracy on average; it will show the *conditional* signature predicted by the model: (a)

$\widehat{\mathsf{Amp}}_{\text{debate}} < \widehat{\mathsf{Amp}}_{\text{consult}}$ robustly across strata, (b) the gap grows in settings where $\widehat{\rho}$ is higher and where fitted evidence responsiveness $\tilde{\kappa}$ is larger, and (c) observed failures align with identifiable mechanisms (tie dominance with high $\widehat{b}$, verification leakage that raises effective $\widehat{\rho}$, or judge inattention that compresses $\tilde{\kappa}$ toward zero). These measurements then become actionable knobs for institutional design in the next section, where we translate the anti-amplification objective into audit and governance procedures.

7 Policy and governance implications: two-sided adversarial review for audits, certification, and compliance; limitations and open questions

Our core comparative-static takeaway—that adding a forced counterparty can reduce conditional wrong-follow $\mathsf{Amp} = \Pr(y = a \mid a \neq \omega)$ even when it does not monotonically increase overall accuracy—maps naturally onto how we should structure high-stakes oversight in practice. Many governance regimes today implicitly implement *open consultancy*: a developer (or a single model) makes a claim, produces supporting artifacts, and an evaluator decides whether to accept. The model highlights a predictable failure mode of this pattern whenever "verifiable-looking" artifacts can be produced for false claims with nontrivial probability ρ: the oversight process can end up *endorsing precisely the wrong outputs*, creating downstream amplification through training, procurement, or deployment.

Two-sided adversarial review as an institutional primitive. The direct governance analogue of $\mathsf{OpenDebate}$ is a requirement that any safety-critical claim be accompanied by an independently generated counter-case that is incentivized (or mandated) to argue the negation. In audits, certification, and compliance this can be instantiated as: (i) a red-team or adversarial auditor tasked with producing the strongest refutation of the developer's claim; (ii) an internal "challenge" unit organizationally separated from the product team; or (iii) a second model instance (or differently trained model) that is *assigned* the opposing stance and given symmetric access to the evidence-production interface. The design goal is not rhetorical balance; it is to ensure that when the proponent is wrong, a truth-aligned counterparty exists that can reliably surface verifiable evidence, thereby shifting the judge away from endorsing the wrong stance.

Implication for certification: certify processes that minimize amplification, not only average accuracy. Regulatory certification often aims to bound expected harm or increase average correctness on test suites. Our framework argues for elevating an additional, process-oriented criterion:

bounding Amp on the relevant distribution, i.e., how often the oversight pipeline *ratifies the proponent when the proponent is wrong.* This matters in settings where certification labels gate deployment at scale (procurement, safety cases, compliance attestations), because wrong ratification is disproportionately costly: it can induce systemic reliance, become precedent for future audits, and feed back into training and monitoring. Concretely, certification regimes can require that high-impact attestations (e.g. "model meets requirement X") be produced via two-sided review and that the certifying body report an estimate of Amp (or a conservative upper bound) under standardized challenge protocols.

Audits as protocol design: controlling ρ, b, and evidence salience. The principal in our model corresponds to the institution designing the audit: it can invest in reducing effective fabrication feasibility ρ, reducing tie-bias b (or baseline favoritism μ), and increasing evidence salience κ. In governance terms, this translates into operational requirements on what counts as evidence and how it is presented:

- *Lower ρ via stricter verification:* mandate provenance (signed logs, reproducible builds, tool-call traces), enforce executable checks (unit tests, formal proofs where appropriate), and require that citations be machine-verified against fixed corpora. The aim is not to eliminate ambiguity, but to shrink the "plausible but false" region that creates ties.
- *Lower effective b via neutrality safeguards:* use label-blind judging (hide which side is the developer), role randomization (rotate who is "protagonist" across instances), and conflict-of-interest rules that separate evaluators from vendors. These interventions directly target the tie region where the wrong side can win despite symmetric evidence indicators.
- *Increase κ via UI and procedures:* highlight verified artifacts, enforce structured claims with explicit pointers to supporting items, and penalize unverified rhetoric. Where judges are models, one can fine-tune them to condition strongly on the verification channel rather than surface form.

These are familiar governance levers, but the model clarifies how they interact: the anti-amplification benefit of adversarial review is strongest when verification is meaningful (low ρ) and when tie outcomes are not systematically proponent-favoring (low b or μ).

Compliance workflows: from one-shot signoff to challengeable claims. A recurring issue in compliance is that attestations become static artifacts that are hard to contest after the fact. The debate framing suggests designing compliance as a *challenge process*: any claim that would be relied upon downstream should be issued with (i) an explicit stance a, (ii) a machine-checkable evidence bundle, and (iii) a standing mechanism for an authorized challenger to produce counterevidence under the same verification rules. When a challenge is raised, the institution re-runs the judge step (possibly with a jury) on the augmented transcript. This converts compliance from a single-sender persuasion problem into an iterated adversarial process in which incorrect endorsements are less likely to persist unopposed.

Juries, independence, and the governance meaning of "multiple judges." Where evaluation is noisy or potentially biased, governance often responds by adding committee review. In our terms, committees are a way to reduce idiosyncratic tie effects and to stabilize decisions when $e_P = e_A$ (or when evidence is graded but close). However, the model also emphasizes a subtlety: aggregation only helps if biases are not aligned. If every judge (or every model-judge) shares the same structural bias toward the proponent, a larger committee can simply entrench that bias. Governance should therefore treat "jury size" as secondary to *jury diversity and independence*: judges should have different training data, different institutional incentives, and ideally different failure modes. For model judges this motivates governance requirements around evaluator model diversity (architectures, providers, fine-tuning sets) and around adversarial testing to detect correlated blind spots.

Limitations: when two-sided review can fail. Two-sided protocols are not a panacea. First, if both sides can reliably generate "verified but misleading" artifacts, the process degenerates into frequent ties (high $\Pr(e_P = e_A = 1)$) and outcomes become dominated by residual bias b or baseline terms μ. Second, debate can reduce follow-when-correct: when the proponent is right, the challenger may sometimes fabricate plausible counterevidence, creating doubt and potentially reducing $\Pr(y = \omega)$ even as Amp falls. This is an acceptable trade in some safety regimes (where avoiding wrong ratification is paramount), but unacceptable in others (e.g. time-critical operations). Third, two-sided review increases cost and latency, and may be strategically gamed: parties might collude, reuse shared evidence templates, or optimize for the verification map $V(\cdot)$ rather than truth (a familiar Goodharting concern). Finally, adversarial review can create information hazards: challengers may surface exploit details or dual-use capabilities. Governance must therefore pair debate protocols with disclosure controls and redaction policies, especially in cybersecurity and biosecurity contexts.

Open questions for scaling the mechanism. Several research and policy questions remain unresolved. (i) *Endogenous stance choice:* our baseline treats the protagonist stance a as given (or as a noisy signal s). In real systems, models may strategically choose which claims to make. Understanding equilibrium selection—including "claim avoidance" under strict verification—is central for governance. (ii) *Richer evidence objects:* binary $e \in \{0, 1\}$ is a simplification. Practical regimes need graded evidence, dependencies across subclaims, and mechanisms that verify *relevance*, not just authenticity. (iii) *Correlated errors and shared priors:* if proponent and challenger are similar models trained on similar data, they may share misconceptions, reducing the probability that the challenger is effectively truth-aligned when the proponent is wrong. This pushes governance toward independence requirements and towards hybrid human–model panels. (iv) *Dynamics:* the principal motivation for anti-amplification is long-run feedback (training on oversight labels, institutional precedent, automated monitoring). Formalizing multi-round learning dynamics where Amp drives path dependence is an important next step for justifying policy thresholds (how small must Amp be to prevent drift?). (v) *Adversarial burden allocation:* in practice we must decide how much budget to allocate to challengers versus proponents, and when to trigger two-sided review. A promising direction is risk-tiering: run consultancy by default, but escalate to debate when the claim is high impact or when automated heuristics predict high effective ρ.

Overall, the governance message is that the *structure* of evaluation matters as much as evaluator quality: by institutionalizing an adversarial counterparty and by investing in verification and neutrality, we can materially reduce the specific failure mode where oversight ratifies wrong answers and thereby amplifies them downstream. This shifts the policy focus from "find better judges" to "design protocols that make it hard to win while wrong," which is often the more robust objective in high-stakes, feedback-rich deployment settings.