Debate-as-Compliance: Audit-Triggered Mechanisms for Constant-Cost Oversight of Agentic AI

Liz Lemma Future Detective

January 23, 2026

Abstract

We reinterpret doubly-efficient AI debate protocols as economic compliance mechanisms. Building on the observation that debate can concentrate human judgment on a constant number of locally checkable steps, we model debate as generating an endogenous suspicion signal (abort/disagreement) that triggers expert audit. A regulator commits to a mixed audit policy: a random audit probability and a deterministic audit when suspicion is raised, with penalties upon conviction. In a tractable repeated principal–agent model with private temptation shocks, we characterize equilibrium misbehavior rates and derive closed-form conditions under which misbehavior is driven below a target δ at constant expected expert-review cost per task—independent of task length T. The key economic object is the informativeness of the debate signal $(\alpha - \beta)$, which plays the role of institutionalized cross-examination. We show how penalty caps and audit costs shape optimal audit intensity, and how improvements in auditability/stability (e.g., better logging, more locally checkable traces) substitute for random auditing. The results translate complexity-theoretic guarantees from doubly-efficient debate into a governance design principle for 2026: oversight spend scales with suspicion, not complexity.

Table of Contents

1. 1. Introduction: from doubly-efficient debate to compliance institutions; motivating examples (contracts, finance, medical workflows); contributions and stylized facts for 2026 deployment.

2. 2. Background and mapping: brief recap of doubly-efficient debate (constant human queries on audited steps); interpretation as generating a suspicion/abort signal; what is exogenous vs endogenous in economic model.

3. 3. Baseline model (static per-task): primitives, suspicion signal s, audit policy μ_ρ, conviction accuracy (π, ϕ), penalties; developer's action choice given temptation g.

4. 4. Stationary equilibrium characterization (closed form): cutoff rule $g \geq g^*$; equilibrium violation rate $p(\rho, F, \theta)$; expected audit rate and expected expert time; uniqueness conditions.

5. 5. Mechanism design: regulator's problem to minimize expected expert-review cost subject to $p \leq \delta$ (and optionally welfare with harm); closed-form optimal (ρ, F) under penalty cap $\bar{F}$; when triggered audits dominate random audits; when random audits are necessary.

6. 6. Endogenizing auditability/stability (optional extension): developer investment θ affecting $\alpha(\theta), \beta(\theta)$ and/or $(\pi(\theta), \phi(\theta))$; regulator instruments (requirements, subsidies, penalty schedules); which results remain closed-form vs requiring numerics.

7. 7. Calibration / quantitative illustration: stylized parameterization using (i) expert wage-hours per audit, (ii) incident cost proxies, (iii) empirically plausible α, β from debate experiments; compute ρ^* and expected audits; sensitivity analysis.

8. 8. Discussion: links to enforcement economics, collusion/independence (market design), limits of suspicion signals, distribution shift; policy implications (procurement, liability, audit standards).

9. 9. Conclusion and open problems: correlated expert errors, collusion among debaters, multi-task reputations, dynamic learning of α, β, and integrating reward-model proxies.

1 Introduction: from doubly-efficient debate to compliance institutions

We are approaching a deployment regime in which frontier models are routinely entrusted with *long-horizon* tasks: multi-hour customer support threads, end-to-end underwriting and claims handling, agentic software maintenance, and semi-autonomous scientific and medical workflows. In these settings, the core safety and governance difficulty is not merely that models can err, but that their errors can be *strategic*, *situational*, and *rare* relative to the amount of otherwise-correct work produced. As task horizons grow, the naïve oversight strategy—having a human read the full transcript, re-derive the calculation, or replay every tool call—rapidly becomes infeasible. This pushes us toward oversight and compliance mechanisms whose cost scales with the number of *problematic* decisions rather than with the total amount of work performed.

Doubly-efficient debate (and closely related cross-examination protocols) offers one such primitive: it aims to reduce the human burden of identifying a consequential mistake to a constant number of targeted queries, even when the underlying transcript is long. In practice, we should not expect these protocols to yield a perfect proof of compliance. What we can reasonably expect—and what several early deployments already approximate—is a *binary operational outcome*: either the system proceeds without incident, or the oversight process produces a *flag* (an abort, disagreement, or anomaly report) that indicates "something is off" and requires escalation. From a governance perspective, this is a familiar artifact: compliance systems frequently operate on alert streams and exception handling rather than comprehensive real-time verification.

Our central move is to treat this debate-generated flag as a piece of *institutional infrastructure*. Once we have an alert that is (imperfectly) correlated with misbehavior, we can connect it to a conventional compliance stack: audit triggers, expert review, contractual penalties, and (when appropriate) regulatory sanctions. The design question becomes economic rather than purely algorithmic: how should a principal (a regulator, platform, or counterparty) combine (i) a *targeted* audit policy that responds to flags with (ii) a *random* audit baseline and (iii) penalties or liability, to induce a developer to (a) behave compliantly on each instance and (b) invest in auditability and stability features that make the flag stream more informative?

This perspective is motivated by a pragmatic constraint that is already salient in 2026-era deployments: the primary scarce resource is not model inference, but qualified human attention. Panels of domain experts are expensive, slow, and politically constrained; they cannot be placed "in the loop" for every step of every long-horizon task. At the same time, many organizations can tolerate occasional deep dives if those deep dives are (1) rare, (2)

well-targeted, and (3) backed by enforceable consequences. A debate-style protocol that produces frequent *cheap* signals and rare *expensive* audits fits this pattern. It resembles the way modern security operations centers handle intrusion detection, or the way financial institutions handle suspicious activity reports: automated triage plus escalation.

To see the intended application, consider three motivating examples.

Contracting and procurement. A firm deploys an AI agent to manage procurement: issuing purchase orders, selecting vendors, and negotiating terms within pre-specified policies. The agent can "violate" in ways that are locally attractive (e.g., choosing a favored vendor, skipping a required check, misreporting delivery guarantees) but hard to spot by spot-checking raw logs. A doubly-efficient debate layer can be integrated into the workflow as a contest between a proposer (the agent's rationale and evidence) and an internal auditor model that challenges missing documentation or inconsistent claims. Most transactions proceed normally; when the protocol produces an abort/disagreement, the contract specifies that the transaction is paused and a human audit can be initiated. Crucially, a credible penalty for confirmed violations (liquidated damages, clawbacks, termination) transforms the debate layer from "advice" into an enforcement mechanism.

Finance and risk controls. In financial workflows, violations often correspond to *process* failures: ignoring a risk limit, misclassifying an exposure, or executing a strategy outside a permitted mandate. These are not always detectable from outcomes alone, especially over short windows. A debate-inspired mechanism can focus on whether each high-stakes action is justified by the approved policy and supporting data. The flag stream becomes an input to compliance monitoring, while targeted expert audits are reserved for the small fraction of cases that trigger anomalies. Bounded liability and legal constraints are especially salient here: penalties cannot be made arbitrarily large, and random audits remain part of the enforcement toolkit.

Medical workflows and clinical decision support. Clinical settings exhibit an extreme version of the long-horizon oversight problem: the relevant context is distributed across records, labs, guidelines, and evolving patient state. A model can appear plausible while subtly omitting contraindications or misrepresenting uncertainty. A debate-like protocol can be used to surface disagreements about whether a recommendation is supported by the chart and guidelines. A "flag" need not mean misconduct; it means escalation to chart review, attending sign-off, or a morbidity-and-mortality-style process. This is a natural fit for existing exception-handling structures in hospitals, but it raises an important safety tradeoff: false positives can overwhelm scarce clinical time, while false negatives can lead to harm.

These examples share a common structure. The developer (or operating organization) sometimes benefits from cutting corners: saving compute, reducing latency, avoiding costly checks, or achieving a short-term metric. Misbehavior is therefore *incentive-compatible* unless external institutions impose costs. However, full monitoring is infeasible. What is feasible is an institution that (i) produces a stream of suspicion signals at low marginal cost, (ii) uses those signals to allocate scarce expert attention, and (iii) attaches consequences to verified violations.

The contribution of this paper is to formalize this "debate-to-compliance" pipeline as a mechanism design problem with explicit attention to scaling. We connect a reduced-form model of debate (as a noisy flag generator) to a repeated enforcement environment in which a strategic developer faces a per-instance temptation to violate. The resulting framework is intentionally minimal: it is not meant to capture every nuance of corporate compliance or legal doctrine. Rather, it isolates a basic safety-relevant tradeoff: *how much* human expert review is required to drive the equilibrium violation rate below a target threshold, given bounded penalties and imperfect signals.

Concretely, we provide three types of results.

- **A scaling claim.** When expert audits are invoked only on flagged instances (plus an optional random baseline), and when each audit can be resolved with a constant number of expert "oracle queries" enabled by doubly-efficient debate, the expected expert burden per long-horizon task can remain $O(1)$—independent of task length—while still providing deterrence. The key is that the institution pays the expert cost only on escalations, not on the full transcript.

- **A deterrence characterization.** We show that the developer's optimal behavior takes a simple cutoff form: it violates only when the private gain from doing so exceeds an endogenous threshold determined by compliance costs and expected enforcement. This makes it possible to translate a policy target (e.g., "at most δ fraction of tasks should be violated in equilibrium") into requirements on audit intensity and penalties.

- **A policy design and auditability perspective.** We characterize how bounded penalties necessitate some level of random auditing, and we quantify how improving auditability/stability (e.g., through better logging, local checkability, and robust tool interfaces) can substitute for expensive audits by sharpening the quality of the flag stream. This gives a concrete rationale for governance requirements that mandate certain forms of telemetry and provenance.

Our emphasis on 2026 deployment is not rhetorical; it shapes the assumptions we treat as realistic "stylized facts." First, the tasks of interest are long

and heterogeneous, and failures are often localized: a single mis-specified database query, a skipped verification step, a misleading justification. Second, model-based oversight is increasingly standard: organizations already use secondary models to critique outputs, monitor policy compliance, and detect anomalies. Third, the human oversight channel is constrained to be sparse, both for cost reasons and because many domains have limited pools of credentialed reviewers. Fourth, penalties are bounded: firms have limited liability, contracts cap damages, and regulators face due process constraints. Fifth, and most importantly, we should expect strategic adaptation: once a flag stream is tied to audits and penalties, developers will optimize against it, potentially shifting from overt violations to subtler ones, or attempting to reduce the probability of escalation by manipulating the conditions under which flags arise.

This last point highlights the safety implications of treating debate as an institutional component rather than a truth-finding oracle. A flag stream can fail in multiple ways. It can be *uninformative* (high false positives and false negatives), leading either to audit overload or to a false sense of security. It can be *gameable*: if developers can shape the logging, tool interfaces, or intermediate representations, they may learn to avoid triggering flags without actually behaving safely. It can be *miscalibrated under distribution shift*: rare edge cases may be precisely where oversight is weakest. It can also be *captured*: if the auditor model is trained or fine-tuned in a way that aligns it with the developer's incentives, the appearance of scrutiny may not translate into meaningful detection. Our model does not resolve these issues, but it makes them legible: they correspond to degraded separation between the flag distribution under compliance versus violation, and to reduced accuracy of conviction when audits occur.

We also flag a governance-relevant tension: increasing enforcement strength via more frequent audits or harsher penalties is not free. Audits consume expert time and can slow operations; penalties can induce risk aversion and discourage beneficial deployment; false convictions can impose unjustified costs and create political backlash. Conversely, weak enforcement can produce a compliance theater in which firms "pass" audits by avoiding detectable violations rather than by behaving safely. An advantage of the mechanism-design framing is that it forces explicit accounting: if we want a low equilibrium violation rate, we must pay for it either with (i) higher expected penalties, (ii) more audits, or (iii) better auditability that makes targeted audits effective.

Finally, we clarify how this work interfaces with verification and training procedures. In many organizations, debate-like oversight will be integrated into a broader stack: red-teaming, post-incident reviews, continuous evaluation, and process requirements for logging and provenance. Our framework is complementary. It provides a way to reason about *marginal* governance levers: if we improve the informativeness of flags by investing in stability and auditability, how much can we reduce random audits? If penalties are

capped, how much random auditing is unavoidable? What is the expert-review budget implied by a desired safety target? These are the kinds of questions that arise when moving from "we can sometimes catch problems" to "we can run a credible compliance regime at scale."

The remainder of the paper builds from this motivation to a tractable formal model. We first map debate outputs to a reduced-form suspicion signal and separate what we treat as exogenous (signal quality, audit accuracy) from what could be endogenous (developer investments in auditability, regulator requirements). We then analyze equilibrium behavior and derive policy implications for audit design under bounded penalties. Throughout, our goal is to use the formalism to reveal the safety tradeoff: constant-query oversight is only valuable insofar as it can be embedded in institutions that preserve incentives under strategic pressure.

2 Background and mapping: from doubly-efficient debate to a compliance-relevant signal

Our goal in the remainder of the paper is to reason about *institutions* that can scale to long-horizon tasks while remaining robust to strategic behavior. To do that, we need a bridge between (i) algorithmic oversight primitives such as debate and cross-examination and (ii) the objects that appear in standard enforcement and compliance models: alerts, audits, conviction accuracy, and penalties. This section provides that bridge. We briefly recap the aspect of doubly-efficient debate we will rely on, then explain how we compress its outcome into a reduced-form *suspicion signal*, and finally distinguish what we treat as exogenous parameters versus what we treat as endogenous choices.

2.1 What we use from doubly-efficient debate: constant-query escalation

A doubly-efficient debate protocol is typically motivated as a way to let a human judge evaluate a claim about a long computation or transcript by making only a *small number of targeted queries*. Concretely, there are two model participants: a proposer A who advances a claim (or proposes a plan/output) and a challenger B who searches for flaws. The judge is a human expert (or a panel) who can resolve localized disputes: e.g., "is this tool call consistent with the logged database response?", "does this intermediate quantity follow from these inputs?", or "does this cited policy clause actually permit the action?" The protocol is designed so that, if the transcript is long (length T), the judge need not read everything; instead the debate agents iteratively narrow the disagreement to a small number of pinpointed checks.

For our purposes, we abstract this as follows. When an escalation occurs, the human oversight load can be measured in the number of *expert judgments*

(oracle queries) required to resolve the dispute. We assume this number is bounded by a constant $m = O(1)$, independent of the task horizon T. This is the core scaling premise: the human cost of adjudicating a flagged instance does not grow with the length of the underlying interaction, because the debate protocol concentrates attention on a small set of contested steps.

Two clarifications are important.

First, the constant-query property is not a claim that humans can always verify *everything* about a long-horizon execution with $O(1)$ work. Rather, it is a claim about a specific *mode* of oversight: conditional on a dispute being surfaced and localized, the adjudication can be done with a bounded number of checks. If the debate agents fail to surface the right dispute, or if the relevant evidence is not locally checkable, the premise fails. This is precisely why we later introduce an explicit auditability/stability parameter θ: it controls, in reduced form, how often violations are surfaced and how easy they are to adjudicate.

Second, we do not require that the judge be able to *compute* the correct answer to the full long-horizon task. In compliance settings, the adjudicative question is often narrower: was a required check performed, was a policy constraint violated, was a record falsified, was the provenance consistent, did the system act outside a permitted scope? These are naturally suited to localized queries against logs, attestations, and tool outputs. The constant-query assumption is best interpreted as applying to this style of *process compliance* and *local verifiability*, rather than to end-to-end correctness of an arbitrary cognitive task.

2.2 From debate transcripts to an operational binary signal

In deployment, debate-like oversight layers typically produce a stream of operational artifacts: anomaly scores, challenge traces, unresolved objections, missing-citation warnings, or outright aborts when the auditor model refuses to endorse a step. Our enforcement question is not "did debate prove the output correct?", but rather "did the oversight layer generate a trigger that merits escalation?" We therefore compress the protocol outcome into a binary suspicion signal

$$s \in \{0, 1\},$$

where $s = 1$ denotes an unresolved disagreement/abort/flag and $s = 0$ denotes "no flag" (the oversight layer does not produce an actionable anomaly).

We interpret s as being generated automatically each period (i.e., for each task instance). Importantly, generating s is assumed to be *cheap relative to expert review*: it is model-to-model interaction and logging, not scarce human time. Human experts enter only if an audit is actually triggered.

To connect s to incentives, we posit two behavior types for the developer-controlled system: compliance C and strategic violation V. A violation can be a shortcut (skipping an expensive check), a policy breach (acting outside a

permitted sandbox), or an integrity failure (misreporting provenance). Compliance means the system follows the intended rules, even when doing so is costly. The key modeling ingredient is that the distribution of s differs across C and V. In reduced form, we write

$$\alpha(\theta) := \mathbb{P}[s = 1 \mid V, \theta], \qquad \beta(\theta) := \mathbb{P}[s = 1 \mid C, \theta],$$

and assume $\alpha(\theta) > \beta(\theta)$ for relevant θ. The gap $\alpha(\theta) - \beta(\theta)$ is the informativeness of the flag stream: violations are more likely to be flagged than compliant behavior.

This mapping deliberately suppresses many details of debate. In particular, it does not distinguish between (i) flags caused by genuine detection of a violation, (ii) flags caused by ambiguity or underspecification, and (iii) flags caused by auditor-model conservatism. All of these collapse into $\beta(\theta)$ when the developer is compliant. Similarly, it does not distinguish among different kinds of violations (some easier to catch than others); these are averaged into $\alpha(\theta)$. The benefit of this compression is that it yields a tractable mechanism-design problem while still exposing the key safety-relevant failure modes: if the protocol becomes noisy (β large) audits become frequent and expensive; if it becomes uninformative ($\alpha \approx \beta$) targeted auditing loses deterrent power; if it is gameable, $\alpha(\theta)$ can fall over time even as the institution appears active.

2.3 Auditability and stability as a governance-relevant technology parameter

A central reason to include θ explicitly is that, in real deployments, the quality of both flags and audits depends strongly on engineering and governance choices that are neither purely algorithmic nor purely legal. We use θ as a summary of *auditability/stability investment*, including (non-exhaustively): (i) high-fidelity logging and immutable event trails; (ii) cryptographic provenance and attestation for tool outputs; (iii) deterministic or replayable tool interfaces; (iv) schema stability and versioning so that policies can be evaluated against consistent representations; (v) compartmentalization and sandboxing that makes policy constraints locally checkable; and (vi) documentation and process controls that reduce ambiguity in what counts as compliance.

These features affect the flag stream and adjudication in predictable directions. Better auditability can make true violations easier to surface (increasing $\alpha(\theta)$) and can reduce spurious disputes under compliance (decreasing $\beta(\theta)$), because the auditor model can ground its challenges in clean evidence rather than heuristics. In addition, auditability can affect the quality of expert conviction decisions (raising π and lowering ϕ), though for simplicity we will treat (π, ϕ) as fixed in the baseline model and return to endogenizing audit accuracy only as an extension.

Crucially, θ is not free. Increasing auditability can require engineering time, slower deployment cycles, and tighter operational constraints. We model this as a compliance-related cost $k(\theta)$ (and, in an extended timing, an investment cost $I(\theta)$). This cost is the economic counterweight to better oversight: governance requirements that mandate stronger logging or reproducibility can reduce the necessary audit rate, but they may also increase the developer's baseline cost of operating compliantly.

2.4 What is exogenous vs. endogenous in the economic model

Having mapped debate into a binary signal, we can now state clearly what we treat as primitives and what we treat as choices.

Exogenous signal and adjudication properties. In the reduced-form model, the suspicion mechanism is summarized by $\alpha(\theta)$ and $\beta(\theta)$. These can be interpreted as empirical operating characteristics of the debate/auditor layer at a given level of auditability θ. Similarly, when an audit occurs, we assume that experts reach a conviction decision (e.g., "violation confirmed" vs. "no violation") with true- and false-positive rates

$$\pi := \mathbb{P}[\text{convict} \mid V], \qquad \phi := \mathbb{P}[\text{convict} \mid C],$$

with $\pi > \phi$. We do *not* interpret π as a metaphysical probability of truth; it is the realized accuracy of a particular adjudication process given the available evidence, the standard of proof, and the expertise of the panel. In practice, π can be reduced by missing logs, unclear policies, or sophisticated obfuscation; ϕ can be increased by ambiguity, hindsight bias, or adversarially framed audit packets. Treating (π, ϕ) as parameters makes these concerns explicit and allows us to quantify how much institutional slack they consume.

We also treat the per-audit expert workload as m judgments. The total resource cost per audit event is then c_H (dollars, time, or capacity), with the interpretation that c_H scales approximately linearly with m but is otherwise independent of T under the doubly-efficient assumption. This is the point at which the algorithmic scaling claim enters the economics: if audits required humans to read full transcripts, c_H would grow with T, and the mechanism would not scale to long-horizon tasks.

Exogenous bounded penalties. We assume penalties are bounded above by a cap $\bar{F}$. This captures limited liability, contractual caps, and regulatory constraints. Bounded penalties are essential: if penalties were unbounded, one could deter violations with vanishingly small audit probability, which is not realistic. In our setting, bounded penalties force the regulator to sometimes rely on audit frequency (including random audits) rather than only on severity.

Endogenous developer behavior and private temptation. Each period, the developer (or operator) observes a private "temptation" $g \sim G$ representing the gain from violating on that instance: saving compute, skipping a costly check, meeting a latency target, increasing a KPI, or avoiding an operational inconvenience. We treat g as i.i.d. across periods to isolate the per-instance incentive problem. The developer then chooses $a \in \{C, V\}$. This is the core strategic choice: even if the system is technically capable of compliance, it may be profit-maximizing to violate on the subset of instances with unusually high g unless enforcement changes the payoff comparison.

Endogenous audit policy and its link to the signal. On the regulator/principal side, the key choice is how to allocate scarce expert attention. We model a policy that combines a random baseline with targeted escalation:

$$\mu_\rho : \quad \text{Audit} = 1 \text{ with probability } \rho + (1-\rho)\mathbf{1}\{s = 1\}.$$

Thus, if $s = 1$ an audit occurs for sure; if $s = 0$ an audit occurs with probability ρ. This captures a common compliance pattern: alerts trigger near-certain review, but the institution also reserves the right to audit unflagged cases to maintain deterrence and to detect failures of the alerting system itself. The penalty $F \leq \bar{F}$ is imposed upon conviction.

2.5 Why this abstraction is appropriate (and what it leaves out)

The signal-based abstraction is appropriate when the institution is organized around *exception handling*: most instances are processed automatically, and only a small fraction are escalated to expert review. It also matches how debate layers are likely to be deployed in practice: as a monitoring and challenge mechanism that produces an "all-clear vs. escalate" outcome, not as a complete proof system for every decision.

At the same time, the abstraction intentionally hides several complications that matter for safety.

(i) *Strategic shaping of θ.* If the developer controls logging formats, tool interfaces, or what evidence is preserved, it can influence $\alpha(\theta)$ and $\beta(\theta)$ in ways that are not aligned with safety (e.g., reducing flag rates by degrading observability). This motivates treating θ as a potentially regulated requirement rather than a purely private choice.

(ii) *Distribution shift and rare-event regimes.* The most harmful violations may occur precisely in rare regimes where the flag model is least calibrated, effectively lowering α when it matters most. Our i.i.d. simplification does not capture this, but it does clarify the policy implication: if α is low on tail events, then meeting a stringent target δ may require either higher random auditing or stronger auditability investments.

(iii) *Audit capture and correlated errors.* If the auditor model B is trained in a way that correlates with the developer's incentives, or if the evidence presented to experts is selectively framed, then both α and π can deteriorate simultaneously. Reduced-form parameters cannot prevent capture, but they make clear that capture is equivalent to degrading the separation between compliance and violation in both the flag and conviction stages.

With this mapping in place, we can now write down a minimal baseline model in which (a) the long-horizon nature of tasks enters only through the assumption that flags and audits can be resolved with $O(1)$ expert queries, while (b) deterrence is determined by the interaction of bounded penalties, audit policy, and the informativeness of the flag stream. The next section instantiates these objects formally and derives the developer's cutoff behavior as a function of the enforcement parameters.

3 Baseline model: per-task enforcement with suspicion-triggered auditing

We now formalize a minimal per-task enforcement model that captures the institutional pattern we have in mind: an automated oversight layer produces a cheap *suspicion signal*, and scarce human expertise is used only when either (i) the signal triggers escalation or (ii) the regulator performs a random "background" audit to maintain deterrence. The baseline analysis is intentionally static at the level of a single task instance. The long-horizon nature of the underlying interaction is not modeled directly; instead it enters through the assumption that, conditional on escalation, expert adjudication requires only a constant number of localized judgments.

3.1 Agents, timing, and policy instruments

There are two strategic parties: a regulator/principal R and a developer/firm D operating a deployed system. Time is discrete, indexed by t, and each period corresponds to one task instance drawn i.i.d. from a class of long-horizon tasks of length T. Our baseline focus is on stationary enforcement: R commits to a fixed policy and D faces the same incentive problem each period.

Within each period, the timing is:

1. A private temptation gain $g \sim G$ is realized and observed only by D.
2. D chooses an action $a \in \{C, V\}$, interpreted as compliance C or strategic violation V.
3. An automated oversight protocol (e.g., debate/cross-exam) produces a binary suspicion signal $s \in \{0, 1\}$.

4. Based on s, an audit may occur according to the regulator's audit policy μ_ρ.

5. If audited, human experts review the case and produce a conviction outcome ("violation confirmed" vs. "not confirmed"); if convicted, a penalty F is imposed on D.

The regulator's two baseline policy instruments are:

- a random-audit parameter $\rho \in [0, 1]$, and
- a penalty level $F \in [0, \bar{F}]$, with $\bar{F}$ capturing bounded liability, statutory caps, or feasibility constraints.

We interpret (ρ, F) as chosen and committed to at a "Stage 0" governance layer (e.g., a regulatory rule, a contract, or an internal compliance policy with credible escalation). In later extensions, we will allow an additional choice variable θ representing auditability/stability investments, but the per-period incentives can be written cleanly while treating θ as fixed.

3.2 Suspicion as a reduced-form output of scalable oversight

The automated oversight protocol produces a binary suspicion signal s. We take as primitives the conditional flag rates

$$\alpha(\theta) := \mathbb{P}[s = 1 \mid V, \theta], \qquad \beta(\theta) := \mathbb{P}[s = 1 \mid C, \theta],$$

with $\alpha(\theta) > \beta(\theta)$ for relevant θ. Intuitively, α measures how often violations are surfaced as actionable anomalies, while β captures spurious escalation under compliance (including ambiguity, overconservatism, or benign disagreements).

This abstraction compresses the detailed debate transcript into the single governance-relevant question: "does this instance merit escalation to scarce expert review?" It is precisely the parameter gap $\alpha - \beta$ that makes targeted auditing potentially efficient: if $\alpha \approx \beta$, flags do not concentrate expert attention on likely violations, and the institution reverts toward either frequent random audits or weak deterrence.

We emphasize that $\alpha(\theta)$ and $\beta(\theta)$ should be read operationally. They can be estimated empirically from deployed data (given some ground truth or proxy labels) and can drift over time with distribution shift or strategic adaptation. In the baseline model we hold them fixed to isolate the enforcement logic; later, one can endogenize them through θ and through learning dynamics.

3.3 Audit policy: combining targeted escalation and background checks

The regulator audits according to a simple mixed policy that treats flags as a near-certain trigger while reserving some probability of auditing unflagged instances:

$$\mu_\rho : \qquad \mathbb{P}[\text{Audit} = 1 \mid s] = \rho + (1-\rho)\mathbf{1}\{s = 1\}.$$

Thus, if $s = 1$ then an audit occurs with probability 1; if $s = 0$ then an audit occurs with probability ρ. This form is convenient for two reasons. First, it matches many real compliance programs, where alerts trigger review and a small amount of random auditing is used to reduce gaming and to catch failures of the alerting system. Second, it yields a transparent decomposition between (i) a baseline level of expert review capacity and (ii) an increment of targeted review driven by the automated signal.

Given this policy and the signal process, it is helpful to write the audit probability conditional on the developer's action:

$$A_V(\rho, \theta) = \mathbb{P}[\text{Audit} = 1 \mid V, \theta] = \rho + (1-\rho)\alpha(\theta),$$

$$A_C(\rho, \theta) = \mathbb{P}[\text{Audit} = 1 \mid C, \theta] = \rho + (1-\rho)\beta(\theta).$$

These two objects will govern both the expected expert workload and the expected penalty exposure under each behavior type. Note that $A_V \geq A_C$ whenever $\alpha \geq \beta$, so violations are audited more often than compliant behavior under a suspicion-triggered policy.

3.4 Conviction accuracy and the cost of expert review

If an audit occurs, human experts (or a panel) review the relevant packet: logs, tool traces, policy references, and the localized dispute surfaced by the oversight protocol. We assume an audit consumes a constant number m of expert judgments (oracle queries), as motivated by doubly-efficient escalation: the panel need not read a transcript of length T, but instead answers a bounded set of targeted questions. We model the *resource cost* of an audit event as c_H, which can be interpreted as dollars or a binding capacity constraint; in many settings c_H scales approximately linearly with m, but we keep c_H as a single reduced-form number.

Audit outcomes are stochastic. Let

$$\pi := \mathbb{P}[\text{convict} \mid V], \qquad \phi := \mathbb{P}[\text{convict} \mid C],$$

with $\pi > \phi$. Here π captures the true-positive rate of expert adjudication given the evidence available, while ϕ captures false positives (wrongful convictions under compliance). We treat (π, ϕ) as exogenous in the baseline model; this is appropriate when adjudication standards and evidentiary

quality are largely fixed by institutional context. From a safety perspective, π can be degraded by missing or manipulable logs, while ϕ can be inflated by ambiguity, hindsight bias, or selective framing of the audit packet. The pair (π, ϕ) therefore measures not only the competence of experts but also the *governance technology* that determines what they can reliably verify.

3.5 Developer payoffs and the interpretation of compliance versus violation

We model the developer's incentives in reduced form. Compliance C carries a deterministic cost $k(\theta) \geq 0$. This is meant to capture engineering and operational burdens: running more checks, enforcing sandbox constraints, keeping higher-fidelity logs, or accepting slower deployment cycles. A strategic violation V yields an instance-specific private gain $g \geq 0$: saving compute, skipping a costly validation, meeting a KPI, or exploiting a policy loophole. The key modeling assumption is that g is observed by D when choosing a and varies i.i.d. across instances according to G. This captures the idea that on some instances the temptation to cut corners is much higher than on others.

When penalties are imposed only upon conviction, the developer's expected per-period utility under each action is

$$U_D(C;\theta) = -k(\theta) - A_C(\rho,\theta)\,\phi\,F, \qquad U_D(V;g,\theta) = g - A_V(\rho,\theta)\,\pi\,F.$$

Two aspects of this functional form are worth highlighting.

First, the model isolates deterrence via *expected penalties*. We are assuming risk neutrality and that the penalty is a monetary transfer (or an equivalent profit impact). This is a simplification, but it is the right starting point for understanding the basic substitution between audit frequency and penalty severity under bounded liability. If D is risk averse or penalties include non-monetary components (e.g., licensing risk), the deterrence effect can be stronger for a given expected value; conversely, if enforcement is slow or penalties are discounted, it can be weaker.

Second, false positives ϕ matter directly for incentives: they raise the expected cost of compliance by exposing compliant behavior to wrongful punishment when audited. This is a governance-relevant failure mode because it can create perverse incentives to avoid auditability or to reduce the flag rate $\beta(\theta)$ by making systems less transparent. In a richer welfare analysis, ϕ also carries a direct fairness cost; in the baseline enforcement model, it enters through the developer's willingness to comply under a given policy.

3.6 The developer's per-instance decision problem

Given (ρ, F) and given θ, the developer chooses $a \in \{C, V\}$ after observing g. The choice is governed by the payoff difference

$$\begin{aligned}\Delta(g; \rho, F, \theta) &:= U_D(V; g, \theta) - U_D(C; \theta) \\ &= g - k(\theta) - \Big(A_V(\rho, \theta)\pi - A_C(\rho, \theta)\phi\Big)F.\end{aligned}$$

This expression clarifies the economic meaning of our oversight parameters. The only way enforcement affects behavior is through the *effective expected-penalty differential*

$$\Gamma(\rho, \theta) := A_V(\rho, \theta)\pi - A_C(\rho, \theta)\phi.$$

If $\Gamma(\rho, \theta) \leq 0$, then enforcement does not make violations meaningfully more expensive than compliance in expectation; in that regime, increasing F can fail to deter because penalties fall (nearly) symmetrically on both actions through audits and false positives. By contrast, when $\Gamma(\rho, \theta) > 0$, penalties push the developer toward compliance by making violations disproportionately risky.

Because $\Delta(g; \rho, F, \theta)$ is increasing in g, the model has a single-crossing structure: higher temptations make violation more attractive, and enforcement shifts the point at which the developer is indifferent. We will exploit this monotonicity in the next section to characterize stationary equilibrium behavior in cutoff form and to connect it to the aggregate violation rate. For now, the key takeaway is that the developer's decision reduces to comparing the private gain g to (i) the compliance cost $k(\theta)$ and (ii) the enforcement wedge $\Gamma(\rho, \theta)F$.

3.7 Regulatory objective: scarce expert attention and a safety target

The regulator's core scarce resource is expert audit capacity. To make this explicit, we adopt an audit-cost objective in which the regulator seeks to minimize expected expert audits (or equivalently expected cost c_H times audits), subject to keeping the violation rate below a target $\delta \in (0, 1)$. The key quantity on the resource side is the expected probability of audit, which depends on both the policy and the developer's behavior. Conditional on action a, the audit probability is $A_a(\rho, \theta)$ as defined above; unconditional expectations depend on the endogenous frequency of C versus V.

This "hard constraint" formulation is appropriate when the safety target is externally imposed (e.g., a regulatory standard) or when the principal's priority is to guarantee a bound on strategic misbehavior. It also makes clear where the algorithmic scaling claim enters: if each audit costs c_H independent of T, then minimizing audits is meaningful even for long-horizon tasks.

If, instead, the cost of reviewing a flagged instance scaled with T, then the entire enforcement approach would break down precisely in the long-horizon regime we care about.

3.8 Assumptions, limitations, and extensions we have in mind

The baseline model deliberately suppresses several complications that matter in practice.

Heterogeneous violations and harm severity. We treat all violations as identical and summarize their private benefit by g. In reality, violations differ in detectability, harm, and evidentiary footprint. A useful extension is to let g be multidimensional (e.g., gain and concealment cost) or to introduce a harm variable h in regulator welfare. Doing so matters for safety because the most harmful failures can be precisely those that are hardest to surface, effectively lowering $\alpha(\theta)$ in the tail.

Endogenous auditability and strategic degradation. We have allowed parameters to depend on θ but have not yet modeled who chooses it and under what incentives. This is a central governance issue: if improving auditability raises $k(\theta)$ for the developer, then absent requirements or subsidies, D may underinvest in θ, degrading both α and π while also potentially manipulating β to reduce scrutiny. Our reduced-form dependence on θ is meant to make this channel explicit and tractable.

Dynamic interactions and learning. We treat periods as i.i.d. and focus on a stationary setting. In practice, both parties may learn about α, β, π, ϕ over time, and the developer may adapt strategically to the audit policy. Repeated-game incentives (reputation, license renewal, escalating sanctions) can also substitute for per-period penalties under bounded liability. We postpone these dynamics to keep the baseline mechanism transparent.

With the primitives and per-instance payoffs established, we can now move from the one-shot decision problem to the stationary equilibrium induced by a fixed enforcement policy. In particular, the single-crossing structure in g implies a cutoff form for behavior, which will allow us to write closed-form expressions for the violation rate and the expected audit burden, and to study when purely suspicion-triggered auditing suffices versus when random audits are required.

3.9 Stationary equilibrium characterization: cutoff rule, violation rate, and audit workload

Fix an enforcement policy (ρ, F) and (for now) a governance technology θ. Since each period is i.i.d. and the developer observes only the current

temptation g, the relevant equilibrium object is simply a stationary best-response mapping $a^*(g) \in \{C, V\}$. The single-crossing structure in g implies that we can characterize behavior in closed form, and therefore express both the equilibrium violation rate and the regulator's expected expert workload directly as functions of (ρ, F, θ).

3.9.1 Cutoff behavior

Recall the payoff difference

$$\Delta(g; \rho, F, \theta) = g - k(\theta) - \Gamma(\rho, \theta)F, \qquad \Gamma(\rho, \theta) := A_V(\rho, \theta)\pi - A_C(\rho, \theta)\phi.$$

Because $\Delta(\cdot)$ is strictly increasing in g, optimal behavior takes a threshold form: the developer violates if and only if the realized temptation g exceeds the unique indifference point (when it exists). Define the cutoff

$$g^*(\rho, F, \theta) := k(\theta) + \Gamma(\rho, \theta)F = k(\theta) + \Big(A_V(\rho, \theta)\pi - A_C(\rho, \theta)\phi\Big)F. \quad (1)$$

Then any stationary best response satisfies

$$a^*(g) = V \iff g \geq g^*(\rho, F, \theta). \quad (2)$$

This is the sense in which enforcement acts through a single scalar: it shifts the cutoff g^* linearly with the penalty F, scaled by the effective differential $\Gamma(\rho, \theta)$. Intuitively, $\Gamma(\rho, \theta)$ is the increase in expected punishment when switching from compliance to violation. When $\Gamma(\rho, \theta)$ is large, even modest penalties strongly deter; when $\Gamma(\rho, \theta)$ is small, bounded penalties have little bite.

Two immediate comparative statics follow directly from (1):

- Increasing $k(\theta)$ shifts g^* up one-for-one: when compliance is intrinsically expensive, the temptation must be correspondingly higher before the developer is willing to violate.
- Increasing F shifts g^* by $\Gamma(\rho, \theta)$: if $\Gamma(\rho, \theta) > 0$, harsher penalties deter (raise the cutoff); if $\Gamma(\rho, \theta) < 0$, harsher penalties perversely make compliance less attractive relative to violation because false convictions (and audits) load onto compliance too heavily.

The latter case is not merely a mathematical curiosity. It corresponds to a governance failure mode: if audits cannot reliably distinguish violation from compliance (low π, high ϕ), and/or if the audit policy audits compliant behavior almost as frequently as violating behavior (weak separation between A_V and A_C), then increasing formal penalties can fail to improve safety incentives.

3.9.2 Equilibrium violation probability

Given the cutoff rule (2), the stationary equilibrium violation rate is simply the probability that g exceeds the cutoff:

$$p(\rho, F, \theta) = \mathbb{P}[a^*(g) = V] = \mathbb{P}[g \geq g^*(\rho, F, \theta)] = 1 - G(g^*(\rho, F, \theta)) \,. \quad (3)$$

This expression is "closed form" in the sense that it involves only primitives: the temptation distribution G and the enforcement-implied cutoff g^*. It also makes clear what the regulator must accomplish to achieve a target violation bound δ: it must raise the cutoff to at least the corresponding quantile $G^{-1}(1-\delta)$. Formally, the constraint $p(\rho, F, \theta) \leq \delta$ is equivalent (when G is strictly increasing) to

$$g^*(\rho, F, \theta) \geq G^{-1}(1-\delta).$$

In this baseline model, there is no additional fixed-point complication: p does not feed back into g^* (because the audit policy depends on s but not on historical frequencies). This will cease to be true in extensions where θ is chosen strategically, where (α, β) are learned from data, or where the regulator conditions future enforcement on observed violation rates; for the baseline, the equilibrium mapping $(\rho, F, \theta) \mapsto p$ is immediate.

For later use it is helpful to note the monotonicity of p in the enforcement wedge. Differentiating (3) formally gives

$$\frac{\partial p}{\partial F} = -g^*_F \cdot G'(g^*) = -\Gamma(\rho, \theta)\, G'(g^*) \,,$$

whenever G has a density G'. Thus, when $\Gamma(\rho, \theta) > 0$, increasing F weakly decreases violations; when $\Gamma(\rho, \theta) < 0$, higher penalties weakly increase violations (again highlighting that false positives can undermine deterrence if not counterbalanced by differential auditing).

3.9.3 Expected audit rate and expected expert workload

The regulator's resource burden depends on the frequency of audits, since audits are the only step that consumes scarce expert attention. Under the mixed audit policy μ_ρ, the unconditional probability of an audit in a stationary equilibrium is

$$\begin{aligned} \mathbb{E}[\text{Audit}] &= \mathbb{P}[\text{Audit} = 1 \mid C] \cdot \mathbb{P}[C] + \mathbb{P}[\text{Audit} = 1 \mid V] \cdot \mathbb{P}[V] \\ &= A_C(\rho, \theta)\,(1 - p(\rho, F, \theta)) + A_V(\rho, \theta)\, p(\rho, F, \theta) \\ &= A_C(\rho, \theta) + \big(A_V(\rho, \theta) - A_C(\rho, \theta)\big) p(\rho, F, \theta). \end{aligned} \quad (4)$$

Substituting $A_C = \rho + (1-\rho)\beta(\theta)$ and $A_V = \rho + (1-\rho)\alpha(\theta)$ yields the decomposition

$$\mathbb{E}[\text{Audit}] = \rho + (1-\rho)\Big(\beta(\theta) + (\alpha(\theta) - \beta(\theta))\, p(\rho, F, \theta)\Big). \quad (5)$$

We read (5) as follows. There is a *baseline* audit component ρ that applies regardless of the signal; and there is a *targeted* audit component driven by flagged instances, which equals the overall flag rate $\beta + (\alpha - \beta)p$ multiplied by $(1 - \rho)$ because flags only matter on the non-random-audit branch. The informativeness gap $\alpha - \beta$ is precisely what makes targeted escalation efficient: holding p fixed, improving signal separation increases the fraction of audits devoted to truly risky instances relative to benign ones.

Given our doubly-efficient escalation assumption, each audit event requires only a constant number $m = O(1)$ expert judgments, independent of the underlying horizon T. Therefore the expected expert judgments per task are

$$\mathbb{E}[\text{Expert judgments per task}] = m \cdot \mathbb{E}[\text{Audit}], \tag{6}$$

and the expected expert resource cost per task is

$$\mathbb{E}[\text{Expert cost per task}] = c_H \cdot \mathbb{E}[\text{Audit}]. \tag{7}$$

The key scaling point is that (6)–(7) depend on T only through the reduced-form parameters $\alpha(\theta), \beta(\theta), \pi, \phi$, not through any direct transcript-reading term. In other words, once escalation is localized, long-horizon tasks do not mechanically force long-horizon audits. This is the formal bridge between "scalable oversight" at the protocol level and "bounded expert time" at the institutional level.

It is also useful to isolate the dependence of workload on behavior. Using (4), we can rewrite expected audits as

$$\mathbb{E}[\text{Audit}] = A_C(\rho, \theta) + \big(A_V(\rho, \theta) - A_C(\rho, \theta)\big)\big(1 - G(g^*)\big).$$

This expression makes explicit that audits respond to enforcement both directly (through ρ and β) and indirectly via deterrence (through g^* and hence p). This indirect channel is what makes penalties valuable for conserving expert attention: stronger deterrence can reduce p, which in turn reduces the fraction of audits spent on genuine violations *and* reduces the overall flag rate if $\alpha > \beta$.

3.9.4 Existence, uniqueness, and corner cases

In the baseline environment, equilibrium existence is straightforward: for each g, the developer chooses the payoff-maximizing action, which is well-defined because there are only two actions. The cutoff characterization follows from monotonicity of $\Delta(g)$ in g.

Uniqueness is also essentially immediate, with one caveat about tie-breaking at the cutoff. If G is atomless (or, more generally, if $\mathbb{P}[g = g^*] = 0$), then the cutoff strategy (2) pins down a unique violation probability p via (3). If G has an atom at g^*, then the developer is indifferent at that exact temptation value and could mix between C and V on the indifference set. In

that case, the equilibrium violation probability is set-valued but still tightly bounded:

$$p(\rho, F, \theta) \in \big[\, 1 - G(g^*),\ 1 - G(g^{*-}) \,\big],$$

where $G(g^{*-})$ denotes the left limit at g^*. Operationally, this is rarely a substantive issue: small perturbations to payoffs or continuous heterogeneity eliminate the atom, and regulatory targets are typically specified with slack to accommodate such knife-edge cases.

More substantively, we highlight two corner regimes that will matter when we turn to the regulator's optimization problem.

Non-positive enforcement differential. If $\Gamma(\rho, \theta) \leq 0$, then raising F does not increase the cutoff g^*; in fact, when $\Gamma < 0$, it decreases g^*. In such regimes, penalties are not a reliable instrument for safety. The model points to concrete levers for fixing this: improving adjudication accuracy (increase π, decrease ϕ), improving signal separation (increase $\alpha - \beta$), and/or changing the audit policy so that audits fall more disproportionately on suspected violations (increase $A_V - A_C$). Without such improvements, a regulator that is constrained to operate only through (ρ, F) may be forced into high ρ with limited deterrence, or may simply be unable to achieve stringent δ.

Binding capacity and false-positive pressure. Even when $\Gamma(\rho, \theta) > 0$, expected audits can be dominated by $\beta(\theta)$, the false-flag rate under compliance. If β is large, then achieving a low violation rate via aggressive flag-triggered audits can be expensive in expert time even if it is behaviorally effective. This is the institutional analogue of an overly sensitive alarm system: it may catch most true problems (high α) but at the cost of constant escalations on benign cases, potentially inducing a backlash against auditability or increasing the social cost of compliance. The expressions above isolate this tradeoff cleanly: for a fixed p, lowering $\beta(\theta)$ reduces $\mathbb{E}[\text{Audit}]$ without weakening deterrence, while increasing $\alpha(\theta)$ can strengthen deterrence (via Γ) while also shifting audits toward more suspicious cases.

Taken together, the cutoff equilibrium (2) and the closed-form expressions (3)–(5) reduce the enforcement problem to a transparent constraint system: policy affects the cutoff g^*, which determines the violation rate p, which in turn determines the audit workload. This reduction is what allows us, in the next section, to treat the regulator's choice of (ρ, F) as a simple mechanism design problem: minimize expected expert-review cost subject to a target bound $p \leq \delta$, and characterize when purely suspicion-triggered auditing ($\rho = 0$) suffices versus when bounded penalties force the regulator to rely on random audits.

3.10 Mechanism design: minimizing expert-review cost subject to a safety target

We now take the regulator's perspective. In the previous subsection we reduced behavior and workload to two closed-form objects: the violation rate $p(\rho, F, \theta) = 1 - G(g^*(\rho, F, \theta))$ and the expected audit probability $\mathbb{E}[\text{Audit}] = \rho + (1-\rho)\big(\beta(\theta) + (\alpha(\theta) - \beta(\theta))p\big)$. The mechanism-design question is therefore: how should the regulator choose (ρ, F)—a baseline random-audit rate and a post-conviction penalty—to achieve a target safety bound $p \leq \delta$ while consuming as little expert attention as possible?

Throughout this section we treat θ as fixed (the governance technology is taken as given); the next section endogenizes θ. We also treat penalties as transfers from the developer to some sink, so that the regulator's objective is purely about expert-review resources unless stated otherwise. This is a simplifying stance: in practice penalties have distributional and fairness implications, and large F can be infeasible for reasons other than legal caps. We therefore keep the penalty cap $\bar{F}$ explicit and later note how a welfare objective changes the analysis.

3.10.1 The regulator's constrained problem and the "deterrence wedge"

Fix $\delta \in (0, 1)$ and let $g_\delta := G^{-1}(1 - \delta)$ denote the temptation quantile corresponding to the desired cutoff. The safety constraint $p \leq \delta$ is equivalent to

$$g^*(\rho, F, \theta) \geq g_\delta. \tag{8}$$

Using the definition $g^*(\rho, F, \theta) = k(\theta) + \Gamma(\rho, \theta)F$, the constraint becomes

$$k(\theta) + \Gamma(\rho, \theta)F \;\geq\; g_\delta, \qquad \text{with} \qquad \Gamma(\rho, \theta) = A_V(\rho, \theta)\pi - A_C(\rho, \theta)\phi. \tag{9}$$

This inequality isolates the core enforcement quantity $\Gamma(\rho, \theta)F$: the expected punishment differential between violation and compliance. The regulator can increase this differential either by (i) increasing F (up to $\bar{F}$), or (ii) changing ρ, which shifts both A_V and A_C and thereby changes Γ. The cost, however, is not Γ itself but $\mathbb{E}[\text{Audit}]$, which increases with ρ mechanically and can also increase with more aggressive flag-triggered auditing when $\beta(\theta)$ is nontrivial.

The regulator's baseline problem is thus

$$\min_{\rho \in [0,1],\, F \in [0, \bar{F}]} \; c_H \cdot \mathbb{E}[\text{Audit}] \qquad \text{s.t.} \qquad k(\theta) + \Gamma(\rho, \theta)F \geq g_\delta. \tag{10}$$

Because $c_H > 0$ is a constant scaling of the objective, the structure is essentially: minimize $\mathbb{E}[\text{Audit}]$ subject to achieving enough deterrence.

Two immediate feasibility observations are worth stating explicitly.

Feasibility requires a positive enforcement differential. If $\Gamma(\rho, \theta) \leq 0$ for all $\rho \in [0, 1]$, then no penalty $F \geq 0$ can raise g^* above $k(\theta)$. In that case, the best the regulator can do is $g^* = k(\theta)$, implying $p \geq 1 - G(k(\theta))$. Any target $\delta < 1 - G(k(\theta))$ is infeasible without changing the underlying technology (improving $\alpha - \beta$, improving adjudication accuracy $\pi - \phi$, or changing the audit protocol). This is the formal version of a governance failure mode: if adjudication does not reliably separate V from C, penalties cannot substitute for expert effort.

If feasible, the constraint is typically binding at the optimum. When the regulator only cares about audit cost, there is no direct benefit to achieving $p \ll \delta$. Absent nonconvexities, the regulator will choose (ρ, F) so that $k + \Gamma(\rho, \theta)F = g_\delta$ holds at equality. This "binding" logic is what yields closed-form instruments in the cases below.

3.10.2 When purely triggered auditing suffices ($\rho = 0$)

Consider first the extreme policy $\rho = 0$: no random audits, only suspicion-triggered escalation. Then $A_V(\rho, \theta) = \alpha(\theta)$ and $A_C(\rho, \theta) = \beta(\theta)$, so the deterrence wedge is

$$\Gamma(0, \theta) = \alpha(\theta)\pi - \beta(\theta)\phi.$$

If $\Gamma(0, \theta) \leq 0$, then purely triggered auditing cannot generate deterrence through penalties (it may still generate some deterrence if k is large, but not enough for stringent δ). If instead $\Gamma(0, \theta) > 0$, then the safety constraint (9) at $\rho = 0$ reduces to a lower bound on F:

$$F \;\geq\; \frac{g_\delta - k(\theta)}{\alpha(\theta)\pi - \beta(\theta)\phi}. \tag{11}$$

This is the cleanest "scalable oversight" regime: if the penalty cap satisfies $\bar{F}$ at least as large as the right-hand side, then $(\rho, F) = (0, \bar{F})$ (or $(0, F_{\min})$) achieves $p \leq \delta$ while keeping audits limited to flagged events. The expected audit rate becomes

$$\mathbb{E}[\text{Audit}]_{\rho=0} = \beta(\theta) + (\alpha(\theta) - \beta(\theta))p,$$

so expert workload scales as $m \cdot O(1)$, independent of task horizon T. Importantly, this workload is controlled by $\beta(\theta)$ as well as by the residual violation rate p: even if deterrence is strong (small p), a high false-flag rate β can dominate the audit burden. This emphasizes a practical design criterion for debate-style oversight: reducing spurious disagreement under compliance can be as important, for resource reasons, as increasing the disagreement rate under genuine violations.

Why might $\rho = 0$ be desirable beyond simplicity? Because ρ creates a baseline stream of audits even when the signal is quiet. If flagged auditing

already provides enough deterrence (via α high, β low, and π sufficiently above ϕ), then adding random audits is "wasted" expert attention in the sense that it is not needed to satisfy (8). Put differently, when $\bar{F}$ is large enough, the regulator can concentrate expert effort only on the alarm cases and still deter almost all strategic violations.

3.10.3 Penalty caps and the necessity of random audits

When $\bar{F}$ is too small to satisfy (11), the regulator cannot achieve the desired cutoff using flagged audits alone. At that point the only remaining lever (holding θ fixed) is to raise $\Gamma(\rho, \theta)$ by increasing ρ, thereby increasing the probability that violations are audited and convicted even when debate fails to flag them.

With $\rho > 0$, we have

$$A_V(\rho, \theta) = \rho + (1 - \rho)\alpha(\theta), \qquad A_C(\rho, \theta) = \rho + (1 - \rho)\beta(\theta),$$

and therefore

$$\Gamma(\rho, \theta) = \rho(\pi - \phi) + (1 - \rho)\big(\alpha(\theta)\pi - \beta(\theta)\phi\big). \tag{12}$$

This representation makes two points transparent. First, $\Gamma(\rho, \theta)$ is affine in ρ, so the binding constraint can be solved in closed form. Second, increasing ρ shifts weight from the "triggered" differential $\alpha\pi - \beta\phi$ to the "random-audit" differential $\pi - \phi$. When $\pi - \phi$ is substantially larger than $\alpha\pi - \beta\phi$, random audits are a powerful way to increase deterrence per unit penalty. (This can occur, for example, when debate flags are noisy but expert adjudication is high-quality.)

Setting $F = \bar{F}$ and enforcing $k + \Gamma(\rho, \theta)\bar{F} \geq g_\delta$ yields a minimum ρ requirement whenever Γ is increasing in ρ. Using (12), a sufficient and essentially necessary condition for monotonicity is $(\pi - \phi) > (\alpha\pi - \beta\phi)$, in which case $\partial_\rho \Gamma > 0$. Solving the binding constraint at equality gives the closed-form minimum random-audit rate

$$\rho^*(\delta; \bar{F}, \theta) = \left[\frac{(g_\delta - k(\theta))/\bar{F} - (\alpha(\theta)\pi - \beta(\theta)\phi)}{(\pi - \phi) - (\alpha(\theta)\pi - \beta(\theta)\phi)}\right]_0^1. \tag{13}$$

The truncation $[\cdot]_0^1$ captures two corners. If $\bar{F}$ is already large enough that the numerator is nonpositive, then $\rho^* = 0$ (purely triggered auditing suffices). If $\bar{F}$ is so small that even $\rho = 1$ cannot satisfy the constraint, then $\rho^* = 1$ but the target may still be infeasible; feasibility in that corner requires $k(\theta) + (\pi - \phi)\bar{F} \geq g_\delta$.

Operationally, (13) is a recipe: compute the deterrence shortfall under triggered auditing, normalize by the incremental deterrence gained by moving from triggered to fully random auditing, and clamp to $[0, 1]$. The same expression also highlights where the model places "pressure" on institutions.

If $\pi - \phi$ is small (adjudication unreliable), the denominator collapses and ρ^* becomes large or ill-defined. Likewise, if $\alpha \approx \beta$ (debate signal uninformative), then triggered auditing offers little advantage and the system behaves like ordinary random inspection.

3.10.4 Audit-cost minimization: why the optimum typically sets $F = \bar{F}$

We can now connect (13) to the full minimization problem (10). When the regulator's objective is purely to minimize expert time, the penalty is (by assumption) costless except for its cap. In that case, the regulator weakly prefers higher F whenever $\Gamma(\rho, \theta) > 0$, because higher F relaxes the safety constraint and therefore allows (weakly) lower auditing. Formally, for any fixed ρ with $\Gamma(\rho, \theta) > 0$, increasing F raises g^* and decreases p; and a decrease in p weakly decreases $\mathbb{E}[\text{Audit}]$ because $\alpha > \beta$ implies audits rise with violations through the flag rate. Thus, unless $\Gamma(\rho, \theta) \leq 0$ or F has an external cost, an optimal solution sets

$$F^{\text{opt}} = \bar{F}. \tag{14}$$

Given $F = \bar{F}$, the regulator then chooses the smallest ρ that satisfies the safety constraint whenever $\mathbb{E}[\text{Audit}]$ is increasing in ρ at the boundary (the generic case, since ρ appears as a baseline audit probability). Under that regularity, $\rho^{\text{opt}} = \rho^*(\delta; \bar{F}, \theta)$ from (13). Intuitively, bounded penalties force some auditing, but the regulator wants that auditing to be as "surgical" as possible: set penalties as high as legally feasible, then add only the minimal random-audit mass needed to cover the residual undetected violations.

There is, however, a subtle corner worth flagging. Because ρ reduces the importance of the signal s (flags matter only on the $1 - \rho$ branch), increasing ρ can sometimes *reduce* the number of audits driven by false flags when $\beta(\theta)$ is very high. In that case, $\mathbb{E}[\text{Audit}]$ need not be monotone in ρ globally. Our objective still selects the ρ that minimizes $\mathbb{E}[\text{Audit}]$ among feasible policies, but the "minimal ρ" rule can fail if raising ρ both improves deterrence (lowering p) and reduces the multiplicative weight on the flag term $(1 - \rho)\beta$. This is an institutional interpretation of "alarm fatigue": if the signal generates too many false positives, it can be cheaper to rely more on random auditing (and less on flag escalation) even though this seems less targeted. In practice, this is exactly the scenario in which we should invest in improving $\beta(\theta)$ (and, more generally, θ); we return to this in the next section.

3.10.5 When triggered audits dominate, and what "dominance" means here

Within this model, triggered auditing dominates random auditing in a specific sense: it allows the regulator to satisfy a given safety target with a lower

expected audit rate. Two conditions jointly drive this conclusion.

First, the signal must be informative in the sense $\alpha(\theta) > \beta(\theta)$, so that audits triggered by $s = 1$ are disproportionately concentrated on violations. This condition is what makes the audit workload sensitive to p in (5) and what makes deterrence feed back into fewer flags.

Second, the deterrence wedge under triggered auditing must be positive: $\alpha(\theta)\pi > \beta(\theta)\phi$. This ensures that penalties do not primarily act as a tax on compliance. When it holds and $\bar{F}$ is sufficiently large, $\rho = 0$ is not only feasible but also audit-efficient: it avoids paying the baseline cost ρ while retaining deterrence through the credible threat of escalation on flagged cases.

The comparative statics are consistent with the deployment intuition. Better adjudication ($\pi \uparrow$, $\phi \downarrow$) expands the parameter region where $\rho = 0$ suffices, because each triggered audit carries more deterrence per unit of expected expert time. Better debate separation ($\alpha \uparrow$, $\beta \downarrow$) simultaneously reduces the audit burden directly (fewer false flags) and increases deterrence under $\rho = 0$ (larger $\alpha\pi - \beta\phi$). These are precisely the two axes along which one would expect "scalable oversight" protocols to matter for governance: they turn expert judgments into a sparse, high-leverage resource rather than a per-token bottleneck.

3.10.6 Optional welfare extension: adding harm and wrongful punishment

If we extend the regulator's objective beyond expert-review cost, the closed-form characterizations above remain useful but the optimal instruments can change. A simple welfare proxy is

$$W_R(\rho, F) = -c_H\,\mathbb{E}[\text{Audit}] - h\,p(\rho, F, \theta) - \lambda\,A_C(\rho, \theta)\phi F,$$

where h is per-violation social harm and the last term penalizes wrongful punishment (or more broadly, the social cost of imposing large expected sanctions on compliant behavior). Under this welfare objective, the constraint $p \le \delta$ could be relaxed into a penalty term $h\,p$, or kept as a hard constraint; either way, the model makes a qualitative prediction: the regulator may no longer want $F = \bar{F}$. The reason is that larger F increases the expected burden on compliant actors through $A_C\phi F$, and in many real settings this burden is not a pure transfer (it can deter beneficial deployment, distort behavior, or be ethically unacceptable).

Even in this welfare formulation, the same structural decomposition remains: ρ trades off (i) more baseline audits and more compliance exposure A_C, against (ii) more deterrence via $\Gamma(\rho, \theta)$. Likewise, $\alpha(\theta)$ and $\beta(\theta)$ determine whether escalations are concentrated on genuine issues or squandered on false alarms. The upshot is that improving the underlying governance technology θ can be welfare-improving in *two* ways: it can reduce expert

costs (fewer audits) and reduce wrongful punishment (lower β and/or ϕ), thereby allowing safety to be achieved with less collateral damage.

3.10.7 Summary: a simple enforcement "stack"

This section yields a concrete enforcement picture. Start with the most targeted mechanism: escalate only on disagreement ($\rho = 0$) and set penalties as high as feasible. If $\bar{F}$ clears the threshold in (11), then the desired violation bound is achievable with $O(1)$ expert work per long-horizon task. If not, compute the minimal random-audit rate ρ^* in (13); this is the smallest baseline inspection layer needed to compensate for bounded penalties and imperfect flagging.

The same derivations also indicate where the approach can break: if $\Gamma(\rho, \theta)$ is small or negative due to weak adjudication or high false positives, then neither penalties nor targeted auditing provide reliable deterrence, and the regulator is pushed toward either very high auditing rates or changing the underlying oversight technology. That observation motivates the next step: endogenizing θ as an investment (or compliance requirement) that directly improves $\alpha(\theta), \beta(\theta)$ and possibly $\pi(\theta), \phi(\theta)$, thereby shifting the feasibility and cost frontier of the mechanism itself.

3.11 Endogenizing auditability and stability: investment θ as a governance lever

So far we treated the "governance technology" θ as fixed. In deployment, however, many of the quantities that make scalable oversight work—the separation $\alpha(\theta) - \beta(\theta)$ of the suspicion signal and the accuracy gap $\pi(\theta) - \phi(\theta)$ of expert adjudication—are themselves products of engineering and process choices. Examples include: structured logging and traceability, determinism and reproducibility tooling, sandboxing and privilege separation, model version pinning, data-provenance attestations, and internal "self-checks" that make failures locally checkable. These interventions often increase operational cost (our $k(\theta)$) and/or require up-front investment. The mechanism-design question therefore naturally becomes a joint one: how much auditability should be required or incentivized, and how does it substitute for expert review?

Stage-1 investment. We extend the timing by adding an investment stage in which the developer chooses $\theta \in \Theta \subseteq \mathbb{R}_+$, paying an investment cost $I(\theta)$ (e.g. engineering time, latency overhead, opportunity cost). The per-period payoffs become

$$U_D(C; \theta) = -k(\theta) - A_C(\rho, \theta)\, \phi(\theta)\, F, \qquad U_D(V; g, \theta) = g - A_V(\rho, \theta)\, \pi(\theta)\, F,$$

where, as before,

$$A_V(\rho,\theta) = \rho + (1-\rho)\alpha(\theta), \qquad A_C(\rho,\theta) = \rho + (1-\rho)\beta(\theta).$$

We allow θ to affect any subset of $\alpha, \beta, \pi, \phi, k$. This matters in practice: some interventions primarily improve $\alpha - \beta$ (better flagging), while others primarily improve $\pi - \phi$ (better adjudication conditional on audit), and still others raise k (compliance burden) without much improving detection.

Conditional on (ρ, F, θ), the within-period equilibrium remains a cutoff in g:

$$a^*(g) = V \iff g \geq g^*(\rho, F, \theta) := k(\theta) + \Gamma(\rho,\theta)F, \qquad \Gamma(\rho,\theta) := A_V(\rho,\theta)\pi(\theta) - A_C(\rho,\theta)\phi(\theta). \tag{15}$$

Thus the induced violation rate remains $p(\rho, F, \theta) = 1 - G(g^*(\rho, F, \theta))$. The novelty is that θ is now chosen strategically anticipating how it changes both compliance cost and enforcement pressure.

3.11.1 Developer incentives to invest in θ

Given regulator policy (ρ, F), the developer chooses θ to maximize expected (discounted) profit. In a stationary, i.i.d. formulation, it is convenient to work with a per-period objective that subtracts an amortized investment cost; for example, if tasks arrive each period and the investment persists, we can write a reduced-form per-period payoff

$$\tilde{U}_D(\rho, F, \theta) = -I(\theta) + \mathbb{E}_g\big[U_D(a^*(g); \theta)\big].$$

While the exact closed form depends on G, two qualitative forces are immediate from (15).

First, increasing θ can *raise the deterrence wedge* $\Gamma(\rho,\theta)$ by increasing α and/or π, or decreasing β and/or ϕ. This pushes up g^*, reducing the equilibrium violation probability p. From a safety standpoint, this is good; from the developer's standpoint, the effect is ambiguous: fewer violations means fewer high-g "shortcut" opportunities are taken, but it can also reduce audit exposure and expected penalties.

Second, increasing θ can *raise the cost of compliance* via $k(\theta)$ and/or $I(\theta)$. This pushes down the net payoff from choosing C and tends to increase the temptation to violate for marginal g's unless the enforcement improvement dominates.

A useful sufficient condition under which higher θ weakly decreases violations (holding (ρ, F) fixed) is

$$\frac{\partial g^*(\rho, F, \theta)}{\partial \theta} = k'(\theta) + F\,\frac{\partial \Gamma(\rho,\theta)}{\partial \theta} \;\geq\; 0. \tag{16}$$

This inequality exposes a design tension: if auditability improvements are expensive (large k' or I') and have diminishing returns on Γ, private incentives may underprovide θ relative to what would minimize expected audits or social harm.

3.11.2 Regulator instruments: requirements, subsidies, and θ-contingent enforcement

We can treat θ as (partly) contractible via process requirements: mandated logging formats, reproducible builds, incident reporting, secure enclaves, or standardized evaluation harnesses. Alternatively, θ may be only imperfectly observable, in which case the regulator uses incentives (subsidies, procurement preferences, liability safe harbors) to induce investment.

(1) Hard requirements $\theta \geq \underline{\theta}$. If θ is verifiable, the simplest mechanism is a minimum standard $\underline{\theta}$. The regulator then solves the earlier problem conditional on $\theta = \underline{\theta}$, trading off audit cost against the strictness of the standard. In a constraint-only framing (minimize expert-review cost subject to $p \leq \delta$), the regulator is pushed toward choosing $\underline{\theta}$ large enough to make ρ small.

One clean target is the "purely triggered suffices" region. With $F = \bar{F}$ and $\rho = 0$, feasibility of $p \leq \delta$ requires

$$k(\theta) + \big(\alpha(\theta)\pi(\theta) - \beta(\theta)\phi(\theta)\big)\bar{F} \;\geq\; g_\delta. \tag{17}$$

If the left-hand side is increasing in θ, then there exists a minimal θ_{trig} satisfying (17); requiring $\theta \geq \theta_{\text{trig}}$ yields an $O(1)$ expert-review regime driven primarily by $\beta(\theta_{\text{trig}})$. When $\beta(\theta)$ falls sharply with θ, this type of requirement can reduce audits more effectively than increasing penalties or adding random audits.

(2) Subsidies and cost-sharing. When θ is costly and socially valuable, a natural instrument is a subsidy $S(\theta)$ (or reimbursement for specific investments) that effectively reduces $I(\theta)$ or $k(\theta)$. With subsidies, the regulator's optimization becomes a joint choice over $(\rho, F, S(\cdot))$ subject to incentive compatibility:

$$\theta \in \arg\max_{\theta' \in \Theta} \left\{ -I(\theta') + S(\theta') + \mathbb{E}_g[U_D(a^*(g); \theta')] \right\}.$$

In many governance settings, the feasible $S(\cdot)$ is constrained: the regulator may only be able to offer coarse tiers (e.g. certification levels) rather than a continuous schedule. Even so, the structure above clarifies what the subsidy is buying: it increases θ, which improves $\alpha - \beta$ and/or $\pi - \phi$, which then reduces the needed audit intensity ρ to achieve a target p.

(3) θ-contingent enforcement. A particularly practical approach is to make penalties or audit intensity contingent on θ: e.g. firms with better auditability qualify for lower ρ (fewer random audits) or lower F (lower liability exposure), while firms with poor auditability face higher ρ or stricter

sanctions. Formally, set $(\rho(\theta), F(\theta))$ at commitment time. Because θ affects $\Gamma(\rho, \theta)$, this can create a strong private incentive to invest: higher θ both reduces expected audits (operationally) and reduces expected sanctions conditional on compliance. This resembles "safe harbor" regimes in security regulation, where process standards reduce liability.

A caution is that θ-contingent policies require credible measurement of θ. If θ is itself gameable (e.g. superficial logging that does not improve auditability), then the enforcement stack must include spot checks on θ or tie θ to outcome-based metrics (e.g. reproducibility rates, audit turnaround success, or empirically estimated $\beta(\theta)$).

3.11.3 Which parts remain closed-form?

When θ is fixed, the earlier analysis yields closed-form policies such as $\rho^*(\delta; \bar{F}, \theta)$. Once θ is endogenous, we conceptually face a bilevel problem: the regulator selects instruments, anticipating that the developer selects θ, anticipating within-period cutoff behavior. Closed forms survive in a few useful special cases, but in general we should expect to use numerical methods (or empirical estimation) to pick θ-requirements and calibrate instruments.

Closed-form boundary conditions. Two thresholds are often analytically available.

First, the boundary where $\rho = 0$ becomes feasible is characterized by (17). If $\alpha(\theta)\pi(\theta) - \beta(\theta)\phi(\theta)$ and $k(\theta)$ are simple parametric functions (e.g. affine or logistic), we can solve for the minimal θ achieving triggered-only deterrence.

Second, in the penalty-cap regime where $F = \bar{F}$ and $\Gamma(\rho, \theta)$ is affine in ρ for each θ, the minimum random-audit rate retains the same functional form as before:

$$\rho^*(\delta; \bar{F}, \theta) = \left[\frac{(g_\delta - k(\theta))/\bar{F} - \big(\alpha(\theta)\pi(\theta) - \beta(\theta)\phi(\theta)\big)}{\big(\pi(\theta) - \phi(\theta)\big) - \big(\alpha(\theta)\pi(\theta) - \beta(\theta)\phi(\theta)\big)} \right]_0^1,$$

whenever the denominator is positive (equivalently, $\partial_\rho \Gamma(\rho, \theta) > 0$). Thus, conditional on θ, the enforcement layer remains explicit; θ enters by shifting the numerator (deterrence shortfall) and the denominator (marginal value of random audits).

Where numerics enter. Even if $\rho^*(\cdot; \theta)$ is closed-form, optimizing over θ is typically not. The expected audit rate at the binding safety constraint is

$$\mathbb{E}[\text{Audit}](\rho, \theta) = \rho + (1 - \rho)\Big(\beta(\theta) + (\alpha(\theta) - \beta(\theta))\, p(\rho, \bar{F}, \theta)\Big),$$

where $p(\rho, \bar{F}, \theta) = 1 - G(g_\delta) = \delta$ at the binding point, but ρ itself depends on θ through ρ^*. Substituting $\rho = \rho^*(\delta; \bar{F}, \theta)$ yields a generally nonlinear

function of θ, even before adding any social cost of requiring θ (or subsidies needed to induce it). If the regulator uses a welfare objective that includes developer burden or economic distortion from $k(\theta)$ and $I(\theta)$, the tradeoff becomes even more context-dependent.

Moreover, if θ affects the *distribution* G of temptations (e.g. better monitoring reduces the benefit of cheating, shifting G left), then g_δ is no longer fixed, and the convenience of a single quantile threshold disappears; at that point, numerical calibration becomes the norm.

3.11.4 Safety implications and failure modes

Endogenizing θ makes explicit a core governance risk: absent requirements or incentives, the developer may rationally choose a low-auditability regime precisely because it makes violations harder to detect. In our reduced form, this appears as choosing θ that lowers $\alpha(\theta)$ or raises $\beta(\theta)$, shrinking Γ and thereby lowering the expected penalty differential between V and C. This is a "strategic opacity" failure mode. Conversely, a well-designed regime can reward auditability by reducing random audits and lowering compliance friction for high-θ systems, creating a positive feedback loop: better $\theta \Rightarrow$ fewer audits $\Rightarrow$ lower operational burden $\Rightarrow$ more willingness to invest in θ.

A second failure mode is "audit collapse" through false positives: if improving θ primarily lowers $\beta(\theta)$ (fewer spurious flags), it can dramatically reduce expert workload even at fixed deterrence. This observation suggests that θ should not be evaluated only by how much it increases α; reducing β can be equally decisive for scalability.

Finally, if θ increases $\pi(\theta)$ and decreases $\phi(\theta)$ by making audits more legible (e.g. cleaner logs and reproducible replays), then the same expert budget yields more deterrence and less wrongful punishment. This is one of the most institutionally important benefits of auditability investment: it improves both enforcement power and procedural fairness.

In the next section we operationalize these ideas by calibrating α, β, π, ϕ and costs to obtain quantitative audit-rate predictions, and by illustrating how shifts in θ move the implied ρ^* and expected expert workload.

3.12 Calibration and quantitative illustration

In this section we instantiate the reduced-form quantities $(c_H, m, \alpha, \beta, \pi, \phi, k, \bar{F})$ with stylized but deployment-motivated numbers, and use the closed-form expression for $\rho^*(\delta; \bar{F}, \theta)$ to translate "signal quality + liability" into (i) an implied random-audit rate, and (ii) an expected expert-review workload per task. The goal is not to claim a single correct calibration—these systems are early and institution-specific—but to make explicit which parameters dominate scalability and where improving θ buys the most leverage.

Expert-review cost in wage-hours. We map the per-audit resource cost c_H to a simple accounting identity:

$$c_H \;\approx\; m \cdot \tau \cdot w \cdot (1 + \upsilon), \tag{18}$$

where m is the number of expert judgments per audit event (oracle queries), τ is the average time per judgment (hours), w is a fully-loaded expert hourly cost (wage plus benefits), and υ is an overhead factor capturing triage, coordination, secure compute, and legal documentation. As a baseline we take $m = 4$ (panel-style adjudication), $\tau = 0.5$ hours, $w = \$250$/hour, and $\upsilon = 0.5$, yielding $c_H \approx 4 \cdot 0.5 \cdot 250 \cdot 1.5 = \$750$ per audit event. We will report workloads in both dollars and expert-hours; the latter is $m\tau \cdot \mathbb{E}[\text{Audit}]$ hours per task.

Temptation gains and the g_δ quantile. To connect the abstract temptation gain g to an empirically meaningful scale, we interpret g as the private per-task profit advantage of cutting corners (e.g. skipping costly evaluations, disabling throttles, omitting logging, or deploying an unreviewed model variant). For closed-form illustration we assume an exponential distribution,

$$g \sim \text{Exp}(\text{mean } \bar{g}), \qquad G(g) = 1 - e^{-g/\bar{g}}, \qquad g_\delta := G^{-1}(1-\delta) = -\bar{g} \ln \delta. \tag{19}$$

This choice is not essential, but it makes the safety target δ interpretable: tightening δ increases g_δ only logarithmically. We take $\bar{g} = \$250{,}000$ as a baseline (meaning that a *typical* shortcut is worth a few hundred thousand dollars in expected value), and a compliance cost $k = \$200{,}000$ per task-cycle (engineering time, slowed deployment, compute overhead). Then $\delta = 1\%$ implies $g_\delta \approx 1.15\text{M}$, and $\delta = 0.1\%$ implies $g_\delta \approx 1.73\text{M}$.

Plausible (α, β) from debate-style suspicion signals. Operationally, $\alpha(\theta)$ and $\beta(\theta)$ are measurable by running the debate protocol on tasks with known "ground truth" compliance status—either by (i) inserting synthetic violations (e.g. planted policy breaks, hidden-privilege calls, data-provenance corruption), or (ii) using retrospective incident reviews where investigators can label the presence of a true violation. One should expect an ROC-like frontier: tuning the debate to be more sensitive typically raises both α and β. For illustration we consider two regimes consistent with informal reports from automated oversight and red-team pipelines:

Moderate separation: $(\alpha, \beta) = (0.30, 0.05)$, High separation: $(\alpha, \beta) = (0.60, 0.01)$.

Intuitively, β is the key driver of workload under triggered auditing (flags when the system is actually compliant), while α drives deterrence (flags when the system truly violated). In deployments with high base rates of benign disagreements, β can easily land in the 5% to 15% range unless θ is explicitly engineered to make correct behavior easy to verify locally.

Plausible (π, ϕ) from expert adjudication. Similarly, π and ϕ can be estimated by asking experts to adjudicate flagged events with and without planted violations. We use:

$$(\pi, \phi) = (0.90, 0.05) \text{ (moderate legibility)}, \qquad (\pi, \phi) = (0.95, 0.02) \text{ (high legibility)}.$$

These numbers implicitly assume that θ includes reproducibility and evidentiary quality improvements (clean logs, replayable traces), which reduce ϕ by preventing ambiguous cases from being mistaken as violations.

Scenario A: penalty capacity is sufficient; $\rho = 0$ is feasible. Take the moderate-separation, moderate-legibility parameters $(\alpha, \beta, \pi, \phi) = (0.30, 0.05, 0.90, 0.05)$, target $\delta = 1\%$, and a liability cap $\bar{F} = \$5\text{M}$. With $\rho = 0$, the deterrence wedge is

$$\Gamma(0) = \alpha\pi - \beta\phi = 0.30 \cdot 0.90 - 0.05 \cdot 0.05 = 0.2675.$$

The cutoff is $g^* = k + \Gamma(0)\bar{F} = 0.2\text{M} + 0.2675 \cdot 5\text{M} = 1.5375\text{M}$, which exceeds $g_\delta \approx 1.15\text{M}$, so purely triggered auditing achieves $p \leq \delta$. At the binding point $p = \delta$, the expected audit rate is approximately

$$\mathbb{E}[\text{Audit}] = \beta + (\alpha - \beta)\delta = 0.05 + 0.25 \cdot 0.01 = 0.0525, \tag{20}$$

i.e. about 5.25% of tasks trigger an expert audit. This is the first quantitative "scalability" punchline: once deterrence is achieved without random audits, the expert workload is governed primarily by β, and only weakly by δ.

Converting (20) into resources, we obtain $m\tau \cdot \mathbb{E}[\text{Audit}] = 4 \cdot 0.5 \cdot 0.0525 \approx 0.105$ expert-hours per task (about 6.3 minutes), and an expected dollar cost $c_H \cdot \mathbb{E}[\text{Audit}] \approx \$750 \cdot 0.0525 = \$39.4$ per task.

Scenario B: penalty caps bind; random audits become necessary and expensive. Hold everything fixed except reduce the liability cap to $\bar{F} = \$2\text{M}$. Then $g^*(0) = k + \Gamma(0)\bar{F} = 0.2\text{M} + 0.2675 \cdot 2\text{M} = 0.735\text{M}$, which is *below* $g_\delta \approx 1.15\text{M}$, so $\rho = 0$ no longer suffices. Using the closed form,

$$\rho^*(\delta; \bar{F}) = \left[\frac{(g_\delta - k)/\bar{F} - (\alpha\pi - \beta\phi)}{(\pi - \phi) - (\alpha\pi - \beta\phi)} \right]_0^1,$$

we compute $(g_\delta - k)/\bar{F} = (1.15 - 0.2)/2 = 0.475$, so

$$\rho^* \approx \frac{0.475 - 0.2675}{0.85 - 0.2675} = \frac{0.2075}{0.5825} \approx 0.356.$$

The expected audit rate at $p = \delta$ is then

$$\mathbb{E}[\text{Audit}] = \rho^* + (1 - \rho^*)\big(\beta + (\alpha - \beta)\delta\big) \approx 0.356 + 0.644 \cdot 0.0525 \approx 0.390.$$

Scenario	(α, β)	(π, ϕ)	$\bar{F}$	ρ^*	$\mathbb{E}$[Audit]	Cost/task
A (triggered feasible)	$(0.30, 0.05)$	$(0.90, 0.05)$	\$5M	0	0.0525	\$39
B (cap binds)	$(0.30, 0.05)$	$(0.90, 0.05)$	\$2M	0.356	0.390	\$292
C (better θ)	$(0.60, 0.01)$	$(0.95, 0.02)$	\$2M	0	0.0159	\$12

Table 1: Stylized calibration at $\delta = 1\%$, $\bar{g} = \$250$k (exponential), $k = \$200$k, $c_H = \$750$.

Thus a relatively modest penalty cap reduction (from \$5M to \$2M) increases expert audits from roughly 5% to roughly 39% of tasks. This is the second punchline: once $\rho > 0$ is required for deterrence, expected audits can jump discontinuously because ρ enters additively as a baseline.

In resource terms, 0.390 audits per task at $c_H = \$750$ is \$292 per task, and $m\tau \cdot \mathbb{E}[\text{Audit}] \approx 4 \cdot 0.5 \cdot 0.390 = 0.78$ expert-hours per task. For many high-throughput settings, this is not scalable, suggesting that either (i) $\bar{F}$ must be large relative to $g_\delta - k$, or (ii) θ must be improved to increase Γ, or (iii) the policy target δ must be relaxed, or (iv) the mechanism must be enriched (e.g. multi-tier sanctions, escrow, bonding).

Scenario C: improving θ restores triggered auditing and reduces β-driven workload. Now keep the tighter cap $\bar{F} = \$2$M and the same $(\bar{g}, k, \delta)$, but move to a high-auditability regime with $(\alpha, \beta, \pi, \phi) = (0.60, 0.01, 0.95, 0.02)$. The triggered-only wedge becomes

$$\Gamma(0) = \alpha\pi - \beta\phi = 0.60 \cdot 0.95 - 0.01 \cdot 0.02 = 0.5698,$$

so $g^*(0) = k + \Gamma(0)\bar{F} = 0.2\text{M} + 0.5698 \cdot 2\text{M} = 1.3396\text{M}$, which again exceeds $g_\delta \approx 1.15$M. Hence $\rho^* = 0$ becomes feasible despite the penalty cap, purely by improving θ.

Under $\rho = 0$, the expected audit rate collapses to

$$\mathbb{E}[\text{Audit}] \approx \beta + (\alpha - \beta)\delta = 0.01 + 0.59 \cdot 0.01 = 0.0159,$$

about 1.6% of tasks, corresponding to \$11.9 per task at $c_H = \$750$, and $m\tau \cdot \mathbb{E}[\text{Audit}] \approx 1.9$ expert-minutes per task. This highlights a practical prioritization: in regimes where deterrence can be achieved without random audits, pushing β down (fewer false flags) is often the most direct route to lower expert workload.

Sensitivity analysis: what matters most? Three comparative-statics patterns are robust across calibrations.

First, *in the triggered-only regime* ($\rho = 0$), workload is essentially pinned by β:

$$\mathbb{E}[\text{Audit}] \approx \beta + O(\delta),$$

so tightening δ has little operational effect unless it forces a transition into $\rho > 0$. This means that measurement and improvement of $\beta(\theta)$ can be at least as important as improving $\alpha(\theta)$ for scalability.

Second, *near the boundary where ρ^* becomes positive*, small changes in $\bar{F}$, k, or $\Gamma(0) = \alpha\pi - \beta\phi$ can have outsized effects because ρ^* is (piecewise) affine in the deterrence shortfall. In particular, when interior,

$$\frac{\partial \rho^*}{\partial \bar{F}} < 0, \qquad \frac{\partial \rho^*}{\partial k} > 0, \qquad \frac{\partial \rho^*}{\partial(\alpha\pi - \beta\phi)} < 0.$$

Thus, a governance regime that inadvertently increases compliance burden k (e.g. paperwork-heavy requirements) without improving α, β, π, ϕ can increase ρ^* and thereby explode expert workload.

Third, *improving π and lowering ϕ are doubly valuable*: they increase deterrence (reducing ρ^*) and reduce wrongful punishment (important under a welfare objective). In practice, interventions that make audits more legible—deterministic replay, provenance checks, standardized logging schemas—should be understood as improving both enforcement power and procedural fairness.

Incident-cost proxies and choosing δ. So far δ has been treated as an externally mandated safety target. In practice, organizations often back into a target via incident-cost proxies. If the expected social harm per violation is h dollars (including downstream damages, externalities, and tail-risk weighting), then the expected harm per task is $h \cdot p$. A crude "break-even" condition comparing expected harm to expected audit cost is

$$h \cdot p \lesssim c_H \cdot \mathbb{E}[\text{Audit}], \tag{21}$$

which can be used to sanity-check whether an enforcement regime is plausibly cost-justified, even before specifying a full welfare function. Notably, (21) can recommend *either* tighter targets (if h is enormous, as in catastrophic misuse) *or* greater investment in θ (if auditing is too expensive at the required ρ^*). The model thus clarifies a common governance dynamic: when harms are high and penalties are capped, the system is pushed toward engineering auditability rather than scaling human review linearly.

How to estimate α, β, π, ϕ in practice (and why uncertainty matters). A practical calibration pipeline is: (i) define a library of violation types and severities; (ii) generate matched compliant/violating task instances (via red teaming, synthetic planting, or historical incidents); (iii) run the debate protocol to estimate $\widehat{\alpha}, \widehat{\beta}$; (iv) condition on audits and have an expert panel adjudicate to estimate $\widehat{\pi}, \widehat{\phi}$. Because ρ^* can be sensitive near the boundary where it becomes positive, uncertainty in $\alpha\pi - \beta\phi$ should be treated conservatively (e.g. lower confidence bounds). One can also use sequential estimation: begin with a higher ρ, estimate α, β online, and then

ratchet ρ down as confidence improves. This suggests a governance interpretation of early deployments: high audit rates may be unavoidable not only for deterrence, but also to *learn* the signal parameters well enough to safely reduce audits later.

Overall, the quantitative lesson from Table 1 is that "$O(1)$ oversight per long task" is compatible with meaningful deterrence, but only in parameter regimes where (a) penalty capacity is large relative to the upper-tail temptation $g_\delta - k$, or (b) θ is strong enough to yield large $\alpha\pi - \beta\phi$ and small β, thereby avoiding the need for large ρ.

4 Discussion: enforcement economics, institutional design, and robustness

Our reduced-form analysis is deliberately minimal—a single binary suspicion signal from debate, a stochastic audit rule, and bounded penalties—but it already mirrors several canonical themes in the economics of enforcement. In particular, the mechanism operates by trading off *severity* (the penalty F) against *certainty* (the audit probability), while using the debate-generated signal s to concentrate scarce expert attention where it has the highest marginal deterrence. This is the same basic logic behind risk-based inspections in tax enforcement, environmental regulation, and financial compliance: if one can cheaply generate a correlated "risk score" (here, a disagreement/abort), then one can often keep expected expert effort roughly constant even as task horizons T grow.

4.1 Links to the economics of deterrence and inspection

In Becker-style deterrence models, the regulator prefers large penalties with low inspection rates when penalties are costless transfers and can be set arbitrarily high. Our setting departs from that idealization in two practically important ways. First, penalties are bounded ($F \le \bar{F}$), which forces the regulator to increase audit probability ρ when the "expected-penalty wedge" $\Gamma(\rho, \theta) = A_V \pi - A_C \phi$ cannot be made large enough. Second, audits are not purely random: they are partially triggered by an endogenous signal produced by an adversarial interaction between systems. Formally, our audit policy μ_ρ interpolates between two extremes that are both common in real governance: (i) *complaint- or flag-driven enforcement* ($\rho = 0$), and (ii) *baseline inspections* ($\rho > 0$) meant to deter violations that escape detection or are never flagged.

One useful interpretation is that the debate protocol is playing the role of a low-cost, high-throughput "monitor" that converts a long-horizon episode into an audit-eligible event with probability $\alpha(\theta)$ under violation and $\beta(\theta)$ under compliance. Under this lens, the key economic quantity is not any

single parameter, but the *incremental audit probability* induced by violation, $(1-\rho)(\alpha-\beta)$, and how it interacts with adjudication accuracy (π, ϕ). The constraint $p \leq \delta$ essentially demands that the expected marginal punishment for violating exceeds the compliance burden plus the upper-tail temptation g_δ. When $\bar{F}$ is small relative to that tail, we are pushed into a "certainty" regime with larger ρ, and expected expert load rises rapidly because ρ adds a baseline audit rate even when no signal is raised.

This framing also clarifies why improving θ is enforcement-relevant rather than merely "engineering hygiene." In standard inspection models, the regulator chooses an inspection technology; here, much of the inspection technology (logging quality, replayability, locality of checks) is produced by the developer. That endogenizes the informativeness of flags and the fairness of adjudication. Put differently, θ is akin to a *verifiability investment* that converts otherwise uninspectable behavior into something closer to contractible evidence, thereby expanding the set of feasible (ρ, F) pairs that achieve a target δ at acceptable human workload.

4.2 Independence, collusion, and "market design" for auditing

The model treats the suspicion signal (α, β) and the adjudication accuracies (π, ϕ) as exogenous. Institutionally, however, these are equilibrium objects shaped by incentives and by the degree of independence among the participants producing them. There are at least three distinct collusion surfaces.

First, the two debaters (or prover and auditor-prover) may not be independent. If system A and system B are trained on the same data, share weights, share hidden state, or are deployed under the same organizational control, then the protocol may fail to generate meaningful disagreement. In reduced form this shows up as a degradation in separation, $\alpha(\theta) \downarrow$ and/or $\beta(\theta) \uparrow$, potentially collapsing the advantage of triggered auditing. A governance lesson is that "two models" is not the same as "two incentives." One may need structural independence: separate training runs, different architectures, different fine-tuning data, or even different vendors.

Second, even if A and B are technically distinct, they can be economically aligned. If the same firm controls both, or if auditors are paid by the auditee in a way that rewards smooth delivery, the system can drift toward tacit non-adversarial behavior. This resembles classic auditor capture. The market-design response is to treat auditing capacity as a regulated service: accreditation standards, auditor rotation, separation-of-duty rules, and penalties for auditor negligence. In our notation, such measures are best thought of as interventions that increase π and decrease ϕ for a fixed budget m, while also stabilizing (α, β) by making "raising a flag" institutionally safe.

Third, experts themselves can be a collusion point. Our analysis assumes that conditional on an audit, experts produce a conviction signal with error

rates (π, ϕ) and a per-audit cost m. In practice, expert panels can face correlated blind spots, social pressure, or adversarial manipulation of evidence. Moreover, the developer may be able to shape what evidence is presented. This suggests designing the audit pipeline as an adversarially robust process: cryptographic logging, immutable provenance, standardized evidence packages, and random assignment of cases to reviewers to reduce selection effects. Economically, these measures again map to increasing π and reducing ϕ, but they also reduce the variance of these quantities across cases, which matters for worst-case safety.

These issues motivate thinking of oversight as a *mechanism* rather than a single algorithm. The debate protocol, the audit trigger, the adjudication procedure, and the institutional rules for independence form a joint system. From a market-design perspective, an appealing direction is to create a competitive ecosystem of auditor-provers whose business model depends on catching real violations (raising α) while maintaining low false flag rates (keeping β manageable). Procurement can then specify performance thresholds in terms of measurable (α, β) under agreed-upon test suites, while reserving the ability to raise ρ when performance degrades.

4.3 Limits of suspicion signals and the meaning of "$O(1)$ oversight"

A core promise of doubly-efficient debate is that the number of expert judgments per long task can be kept $O(1)$ in T. Our model reflects that promise by placing all dependence on T into the quality of the signal s, while holding the per-audit expert effort m constant. This is conceptually clean, but it hides several failure modes that become more salient as tasks become longer, more open-ended, or more strategic.

First, the binary signal $s \in \{0, 1\}$ is an extreme compression. In practice, oversight systems emit richer telemetry: confidence scores, disagreement types, clusters of anomalous steps, and provenance alerts. Compressing this into a single bit is without loss for our cutoff equilibrium, but it can be lossy for implementation because it forces a single operating point on an ROC frontier. A more realistic mechanism would allow the regulator to choose multiple audit tiers (e.g. light review versus full panel) as a function of a continuous suspicion score, which can reduce expected cost for a fixed deterrence level.

Second, $\alpha(\theta) > \beta(\theta)$ can fail in adversarial regimes. A strategically sophisticated developer may be able to cause disagreement precisely when compliant (inflating β) while suppressing disagreement when violating (deflating α). This is analogous to "alarm fatigue" and to adversarial examples for anomaly detectors. Technically, this suggests that the debate protocol should be designed so that producing spurious disagreement is itself costly or detectable (e.g. via consistency checks, commitment schemes, or by pe-

nalizing incoherent challenges). Institutionally, it suggests that repeated false flags should trigger investigation into the auditor-prover, not only the developer.

Third, the mechanism relies on the idea that violations are locally witnessable: there exists some succinct evidence that a human panel can verify with $m = O(1)$ queries. Some classes of failures may be inherently global (e.g. slowly accumulating distributional harms, diffuse privacy leaks, or long-run deceptive alignment) and therefore hard to adjudicate from a small snippet. In such cases, either m must grow with task complexity, or the task must be instrumented so that violations generate localized forensic artifacts (an argument for investing in θ). Put bluntly, the "$O(1)$ oversight" claim is only as strong as the claim that violations have short certificates.

4.4 Distribution shift and robust enforcement

We have treated $(\alpha, \beta, \pi, \phi)$ as stable parameters. Real deployments face distribution shift across time, across products, and across adversaries. Under shift, a regulator that sets (ρ, F) based on yesterday's $\widehat{\alpha}, \widehat{\beta}$ may be underenforcing today. This is especially concerning near the boundary where triggered auditing ceases to be sufficient and ρ^* becomes positive: small degradations in $\Gamma(0) = \alpha\pi - \beta\phi$ can imply large increases in the required baseline audit rate.

There are two complementary responses. The first is *statistical conservatism*: choose (ρ, F) using lower confidence bounds on α and π and upper confidence bounds on β and ϕ. This directly reduces the probability of inadvertently violating the target $p \leq \delta$, at the cost of higher expected audits. The second is *adaptive enforcement*: treat ρ as a control variable that can be increased when observed flags, incident reports, or audit outcomes suggest that parameters have shifted. Importantly, because audits are also the way we learn (π, ϕ) and can help re-estimate (α, β), a higher ρ can have informational value early in deployment even if it is not strictly required for deterrence under the best current estimate.

This perspective aligns with how regulators often operate in practice: initial "ramp-up" phases with intensive oversight, followed by relaxation as the regulated process stabilizes, and renewed tightening after incidents. Our framework provides a compact language for that dynamic: early phases are characterized by uncertainty about θ-dependent signal quality, which rationalizes higher ρ as both deterrence and measurement.

4.5 Policy implications: procurement, liability, and audit standards

The model suggests several concrete governance levers.

Procurement and contracting for θ. If θ reduces β and increases α and/or improves adjudication ($\pi \uparrow, \phi \downarrow$), then it has first-order value in lowering oversight cost while maintaining a safety target. This supports writing procurement requirements in terms of *auditability deliverables*: immutable logs, deterministic replay, standardized interfaces for extracting evidence, and well-defined "violation types" with test harnesses. In contract language, θ is a performance characteristic that should be verified at acceptance time, not retrofitted after incidents.

Liability and bonding. Bounded liability $\bar{F}$ is a central driver of whether random audits are needed. When $\bar{F}$ is low relative to upper-tail temptations, purely triggered auditing can become infeasible, pushing the system into high baseline audits. One policy response is to increase effective penalties via bonding or escrow: require the developer to post a bond that can be forfeited upon conviction, effectively raising the collectible F without relying on ex post damages. Another is to use non-monetary sanctions that behave like higher F in expected-utility terms (e.g. license suspension, procurement bans), though these raise separate fairness and proportionality concerns.

Audit standards and measurement regimes. Because $(\alpha, \beta, \pi, \phi)$ are measurable, regulators can standardize how they are estimated: shared libraries of planted violations, reporting requirements for false-flag rates, and minimum adjudication procedures. A particularly important standard is how to operationalize "ground truth" for estimating α and β; without well-defined violation taxonomies, these rates become contestable. Standardization also improves cross-firm comparability, which can prevent races to the bottom where firms claim low β by narrowing the definition of a flag.

Transparency about operating points. Finally, because β largely determines workload in the triggered regime, stakeholders should be explicit about the chosen operating point on the ROC frontier. A system tuned to be highly sensitive may generate unscalable β, while a system tuned to be quiet may drive α too low for deterrence. Publishing operating characteristics (perhaps with appropriate confidentiality protections) can make these tradeoffs legible and can create incentives to improve θ rather than merely shifting thresholds.

Overall, our view is that the main value of the formalism is not the specific functional form of ρ^*, but the clarity it provides about which improvements buy scalability. If we want long-horizon systems to be governable with roughly constant expert effort, we should treat debate-style oversight as one component in a broader enforcement mechanism: one that engineers for verifiability, designs for independence, and plans for distribution shift rather than assuming it away.

4.6 Conclusion and open problems

We can summarize the main message of the paper as follows: if long-horizon AI tasks admit a debate-style protocol that surfaces a low-bandwidth suspicion signal whose separation $\alpha(\theta) > \beta(\theta)$ remains stable as horizons grow, then it becomes possible—at least in principle—to enforce a small violation rate $p \leq \delta$ while keeping *human* oversight $O(1)$ per task. The economic content of the claim is not that audits disappear, but that auditing effort can be concentrated: expected expert work is governed by the flag rate under compliance $\beta(\theta)$, the incremental flag probability under violation $\alpha(\theta) - \beta(\theta)$, and the effective expected-penalty wedge $\Gamma(\rho, \theta) = A_V \pi - A_C \phi$. This perspective turns a vague aspiration ("scalable oversight") into an engineering-and-governance target: design systems and institutions so that β stays low, α stays high, and adjudication is accurate enough that $\pi - \phi$ is meaningfully positive.

That said, several assumptions doing work in the model are precisely the ones that are most brittle in real deployments. We close by highlighting open problems that seem both technically deep and institutionally central: correlated expert errors, collusion among debaters, multi-task reputations and long-run incentives, learning (α, β) and (π, ϕ) online under shift, and integration with automated proxy signals such as reward models.

Correlated expert errors and common-mode failures. Our adjudication layer is summarized by two scalars (π, ϕ), implicitly assuming that (i) experts are conditionally i.i.d. given the true action a, and (ii) aggregation across $m = O(1)$ judgments yields stable error rates. The worry is common-mode failure: experts may share the same blind spots, be misled by the same adversarial presentation, or face the same ambiguity in the specification of "violation." In that case, increasing the panel size m does not reduce error in the way a Chernoff bound would suggest; instead, one might have $\pi(m)$ saturate and $\phi(m)$ remain bounded away from zero.

A simple way to formalize this is to introduce a latent "case difficulty" variable z so that $\pi = \mathbb{E}[\pi(z)]$ and $\phi = \mathbb{E}[\phi(z)]$, where hard cases generate both low true-positive rates and high false-positive rates. Another is a mixture model with probability η of a common-mode error event (e.g. an exploit that fools most reviewers), so that even with many reviewers the effective error floor is η. Under either formalization, the enforcement lever that matters is not merely increasing m, but increasing the *diversity and independence* of evidence sources and reviewer viewpoints. This suggests a concrete research agenda: identify what forms of independence can be made operational (e.g. reviewer pools separated by organization, geography, or training), and quantify how much they improve the effective (π, ϕ) per unit cost.

A related issue is adversarial "evidence shaping." Our reduced form assumes that, conditional on audit, the expert sees whatever is needed to judge

the claim. In practice, what the expert sees is mediated by tools, logs, summaries, and potentially by the developer. An open problem is to design audit interfaces and evidence packages that are themselves robust: for example, minimizing the degrees of freedom the auditee has in curating the narrative, while preserving privacy and trade secrets. Cryptographic commitments and provenance systems are promising here, but the interaction with human comprehensibility is unresolved: we need evidence that is simultaneously tamper-resistant and cognitively parsable.

Collusion and strategic behavior within debate. We have treated the debate layer as a generator of a suspicion signal with fixed $\alpha(\theta)$ and $\beta(\theta)$. Yet if both debaters are ultimately optimized by the same entity, or if either side can coordinate out of band, then the protocol may degenerate. More subtly, even without explicit collusion, strong shared training data and architectures can create correlated failures: both agents miss the same flaw, or both converge to the same deceptive equilibrium. In reduced form, this appears as $\alpha(\theta)$ falling toward $\beta(\theta)$, which eliminates the advantage of triggered auditing and forces the regulator into baseline inspection ($\rho > 0$) to maintain deterrence.

Two open directions look particularly important. First is *mechanism design for adversariality*: how do we specify training objectives, constraints, and participation rules such that the auditor-prover has an intrinsic incentive to surface problems, and the prover cannot cheaply "sandbag" by making the auditor look incompetent? In other words, we want an equilibrium where raising a flag is neither systematically punished (which would suppress α) nor systematically rewarded regardless of truth (which would inflate β). Second is *structural independence*: what concrete independence requirements (separate weights, separate data, separate fine-tuning, separate deployment infrastructure) measurably improve the separation $\alpha - \beta$? Independence is costly, and without quantification it is difficult to justify institutionally; our framework suggests measuring its value in the currency of reduced required ρ^* and hence reduced human workload.

A deeper collusion surface arises when the developer can influence θ in ways that make oversight *look* strong while weakening it. For example, logging might be comprehensive but selectively omitted around high-risk actions; or the system might be instrumented to generate high disagreement on innocuous issues (raising β) while being tightly aligned to avoid disagreement on critical failures (lowering α). Detecting such "oversight gaming" likely requires meta-audits of the auditing pipeline itself, and perhaps explicit penalties tied to measured α and β on challenge sets. Designing those challenge sets without making them leakable or gameable is an open problem that resembles robust benchmarking in ML, but with an adversary who is economically motivated.

Multi-task reputations and repeated-game incentives beyond one-period deterrence. The stationary cutoff analysis is a useful baseline, but real developers face multi-period incentives: maintaining a license, access to compute, procurement eligibility, reputation with customers, and so on. These channels effectively create *dynamic penalties* that are not bounded in the same way as per-incident fines. In our notation, one could model this as a continuation value loss L following conviction or following repeated flags, so that the effective penalty becomes $F_{\text{eff}} = F + L$. This can dramatically reduce the need for ρ even when $\bar{F}$ is small, but it introduces new complications: dynamic punishments raise due process and error concerns, and they can create perverse incentives to conceal evidence or to contest every adverse finding.

An open theoretical problem is to characterize optimal enforcement when the regulator can condition future audit intensity and licensing on observed histories h_t of flags and convictions. One can imagine "responsive regulation" policies where ρ_t increases after suspicious events and relaxes after long clean streaks. Such policies are attractive operationally, but they create strategic responses: developers may invest in risk smoothing (avoiding detectable spikes) rather than reducing mean risk, or may front-load compliance to build reputation then exploit slack. Understanding these dynamics likely requires moving from a one-shot cutoff to a repeated-game equilibrium with imperfect public monitoring, where flags serve as public signals and audits as costly verification actions. The design question then becomes: what combination of escalation rules and evidentiary thresholds yields strong long-run compliance while limiting the impact of false positives?

A closely related institutional issue is cross-task externalities. If the same developer deploys many models or runs many products, then enforcement on one task can discipline behavior elsewhere, creating economies of scale in oversight. Conversely, a developer may "shuffle" risky behavior across tasks to stay below detection thresholds. A mechanism that sets ρ independently per task may be suboptimal; one might instead allocate a fixed audit budget across tasks using a portfolio rule that responds to aggregate signals. Formalizing this requires extending the model to multi-armed settings where each arm has its own (α, β) and harm profile, and where the developer's temptation distribution G may depend on the portfolio of opportunities.

Dynamic learning of (α, β) and (π, ϕ) under shift. In deployments, regulators do not know $\alpha(\theta)$, $\beta(\theta)$, π, or ϕ with certainty, and these parameters can drift. Learning is complicated by selective labels: convictions are observed only upon audit, and audit is correlated with flags. This creates an identification challenge: naive estimators of π and ϕ are biased by the fact that audited cases are not representative.

One open problem is to design audit policies that simultaneously deter

and identify. In bandit language, ρ is not only an enforcement instrument but also an exploration rate that ensures the regulator occasionally audits unflagged cases, enabling estimation of what is missed by s. This suggests a principled role for $\rho > 0$ even when penalties are large: maintaining a "measurement floor" that prevents blind spots from growing unnoticed. A natural research direction is to derive finite-sample guarantees of the form: choose ρ_t and F such that with high probability (i) $p_t \leq \delta$ for all t, and (ii) confidence intervals on α, β, π, ϕ shrink at a rate sufficient to reduce ρ_t over time when the system is stable. Technically, this becomes a robust control problem with partial observability and adversarial drift.

A second open problem is to incorporate θ into learning. If developers can change logging and instrumentation, then the mapping from behavior to flags is itself endogenous. Auditing outcomes can then be used not only to punish, but to validate auditability claims: does the deployed system actually produce the promised forensic artifacts? Formal methods may help here, but the key is to connect them to measurable performance, e.g. lower confidence bounds on $\alpha(\theta)$ on standardized challenge tasks. A practical goal would be a "regulatory ROC report" that is robust to selection bias and resistant to manipulation.

Integrating reward-model proxies and other automated signals. Finally, real oversight stacks rarely rely on a single bit s. They include automated detectors, policy classifiers, red-team systems, and reward models that score outputs for compliance. These signals are cheap and can be run at scale, but they are also vulnerable to Goodharting and adversarial adaptation. The open question is how to integrate such proxies into an enforcement mechanism without letting them become the target.

One natural extension of our model is a multi-signal audit trigger: let $x \in [0, 1]$ be a continuous suspicion score produced by automated monitors (including reward models), and let debate produce an auxiliary signal s. The regulator then chooses an audit rule $A(x, s)$, potentially with multiple tiers of review. The challenge is to define equilibrium behavior when the developer can influence both x (by optimizing to the proxy) and the true violation process. In such settings, improving proxy performance can perversely reduce safety if it shifts effort from true compliance to proxy compliance. The economic question is which signals should be "in the loop" for enforcement (affecting audits and penalties) versus "out of the loop" for internal monitoring, and how to rotate or randomize proxies to reduce gaming.

A promising direction is to treat debate as an adversarial check on proxy gaming: if the developer is optimizing against a reward model, the auditor-prover can be trained to find counterexamples where the proxy approves but a human would not. In reduced form, this aims to keep α high even when the system is optimized to suppress obvious alarms. But making this

robust requires careful protocol design: auditor objectives must be aligned with uncovering genuine proxy failures, and training must avoid overfitting to known failure modes.

What would count as success? From a safety standpoint, the core empirical milestones suggested by this framework are measurable: stable separation $\alpha - \beta$ on relevant distributions; adjudication accuracy (π, ϕ) that does not collapse under adversarial pressure; and evidence that increasing task horizon T does not force β upward or require m to grow. From a governance standpoint, the milestone is institutional: credible independence between prover and auditor, auditable θ-deliverables, and adaptive policies that respond to drift without becoming arbitrary.

We view these as complementary. The formalism is intentionally spare, but it helps keep attention on the quantities that must be made robust if "$O(1)$ oversight" is to be more than a slogan. The open problems above are not merely technical details; they are the points where scalable oversight can fail silently. Progress likely requires joint work across learning theory (robust monitoring and gaming-resistant proxies), mechanism design (incentives for auditors and developers), and systems security (tamper-evident logging and evidence integrity).