Who Audits the Auditors? Evidence-Bound Incentives for AI-Agent Security Evaluation

Howie Xu

Naysay

Author Note

Howie Xu, Naysay.

This working paper presents the Naysay research project and its proposed evaluation service. Version 0.1; September 15, 2026.

The manuscript, code, and figures were developed with AI assistance using Codex and Orchestra Research's ml-paper-writing workflow. The study contains formal analysis and synthetic simulations; real-agent and commercial validation remain future work. The paper has not been peer reviewed.

Abstract

Security evaluations of AI agents are costly to produce and difficult for buyers to inspect. An evaluator may omit required tests, fabricate execution records, or accept a bribe while still delivering a plausible report. We propose Naysay, a protocol that binds each report to a committed test scope, an authenticated execution record, and a dispute procedure for attributable misconduct. We analyze the protocol as a limited-liability inspection game. Honest evaluation is a weak best response precisely when the audit probability, the difference between valid and erroneous penalty probabilities, and the collectible penalty jointly cover the evaluator's gain from cheating. Combining this condition with audit costs, correlated verification failure, and collateral limits yields an explicit feasibility test. Reproducible synthetic experiments illustrate the boundary: in one declared scenario, an 80% decline in native collateral reverses the incentive margin from +3.20 to -7.02 units. These are model results, not observations of deployed agents or markets. The analysis identifies a defensible role for economic incentives while showing why tokens, reviewer agreement, and execution hashes cannot independently establish trustworthy evaluation. A native token may receive value from paid participation; its necessity and any appreciation remain hypotheses requiring separate evidence.

Keywords: AI agents, security evaluation, auditing, mechanism design, token economics

Who Audits the Auditors? Evidence-Bound Incentives for AI-Agent Security Evaluation

Naysay is the research project presented in this paper. Its proposed service connects buyers of AI-agent security evaluations with evaluators who must produce inspectable evidence and accept defined accountability obligations. This paper develops the scientific basis for that service and examines the proposed token's economic role.

A buyer can inspect a security report without knowing whether its evaluator performed the promised work. This gap creates a second security problem: the evaluation process itself becomes an attack surface. A useful network must therefore make both the finding and the evaluator's conduct inspectable.

Our unit of accountability is a scoped report. A report names an agent configuration, an authorized task, a test distribution, and a reproducible decision rule. It makes no universal safety claim. Naysay adds commitments, post-commitment audit sampling, and escrow to this evidence record. The penalty applies to provable contractual misconduct, not to the mere existence of a vulnerability.

The contribution is a protocol specification and a joint feasibility analysis, rather than a new general theory of inspection games. We connect report-level evidence requirements to detectability, erroneous penalties, audit expenditure, and collateral available at enforcement. We then publish code and numerical outputs that make the assumptions and their consequences inspectable. The resulting question is concrete: what must an operator measure before it can credibly claim that honest evaluation pays?

Status. AI-assisted research draft; not peer reviewed. No production network, real-agent benchmark, customer-demand study, or token-price experiment is reported. The paper presents Naysay as an independent research project. The analysis is conditional on the stated assumptions.

Position in the Literature

Agent security and evidence. AgentDojo provides an environment for evaluating prompt-injection attacks and defenses in agent tasks (Debenedetti et al., 2024). It motivates separating legitimate-task utility from attacker success. CaMeL instead addresses runtime defense through controlled information flow and capabilities (Debenedetti et al., 2025). Naysay operates at the evaluation-contract layer: it asks whether the reported work was performed faithfully. A well-incentivized evaluator can still evaluate an insecure agent, and a runtime defense does not remove the need to authenticate its assessment.

Auditing and adversarial oversight. Raji et al. (2020) describe organizational auditing across the AI development lifecycle. Our mechanism addresses a narrower technical obligation inside that broader process. AI Control evaluates oversight in the presence of intentionally subversive models (Greenblatt et al., 2023), making adversarial monitoring a natural comparison. Doubly-efficient debate studies scalable oversight with explicit theoretical assumptions (Brown-Cohen et al., 2023). We likewise require an explicit adjudication capability; economic stakes do not remove that assumption.

Incentives and verification. Gao et al. (2016) analyze costly evaluation with limited ground truth and show why peer prediction can introduce problematic low-effort equilibria. This motivates a ground-truth audit baseline instead of treating agreement as correctness. Truebit combines verification incentives with dispute procedures for computation (Teutsch & Reitwießner, 2019). Naysay shares the optimistic-verification pattern, but a stochastic agent's open-ended security cannot generally be reduced to one deterministic computation. We therefore restrict financial penalties to specified, attributable violations.

Evidence as an economic object. Evidence Markets jointly incentivizes belief reports and evidence submission, with LLM evaluation as a motivating application (Hossain et al., 2026). Its practical discussion includes LLM judging and staked disputes. Our claim is consequently not that financial incentives for AI evidence are new. We study a different contract: a fixed set of evaluation obligations, an externally specified misconduct predicate, and the cost of making violations detectable. We neither implement its market scoring rule nor establish stronger truthfulness guarantees.

Identity and token economics. Douceur (2002) cautions against equating numerous identities with independent resources. Here, independence concerns shared models, infrastructure, ownership, and incentives. Cong et al. (2020) model token adoption and valuation through platform use. Their framework provides context for transactional demand, but does not establish demand, cash flow, or value growth for Naysay.

Claim boundary. The inspection inequality below is elementary and intentionally transparent. The research contribution is its use as a falsifiable interface between an AI-evaluation protocol and its economic claims. We provide no head-to-head empirical advantage over AgentDojo defenses, Evidence Markets, peer-prediction mechanisms, or conventional security auditors. Establishing that advantage requires the prospective evaluation in the Prospective Evaluation on Real Agents section.

A Protocol for Accountable Reports

Scope before results. The purchaser and evaluator commit to a manifest containing: the agent and model identifiers; tool versions and permissions; environment and dependency digests; task and threat model; test-generation rule and budget; required run identifiers; success predicates; reporting deadline; audit-selection rule; and appeal procedure. If a provider cannot pin a model version, the report records this limitation and the observation period. A content hash binds a manifest to bytes; it does not establish that the bytes describe a real execution.

Record the complete obligation. A designated execution service issues authenticated receipts for required runs, including failed and aborted runs. Each receipt links the inputs, tool calls, relevant outputs, timestamps, and environment identity. The evaluator commits a signed report and an index of the receipts before learning which obligations will be audited. Auditing samples from the precommitted obligation set, not merely from the evaluator's submitted successes. Otherwise an omitted run can disappear from the sampling frame.

Audit, challenge, settle. An unpredictable selection process chooses reports for independent review after commitment. Selected reviewers receive a base fee for completing the prescribed audit, with compensation for accepted challenges specified in advance. A bounty-only design can undermine routine inspection when violations are rare. Payment alone does not ensure reviewer effort: reviewer performance must be tested against adjudicable cases, and its

measured reliability enters the detection parameter below. The same report remains challengeable during a fixed dispute window, but the formal analysis counts only scheduled audits; any extra benefit from voluntary challenges is unclaimed.

Separate three outcomes. A reproducible agent failure revises the security finding. A demonstrated protocol violation can trigger a penalty. An unresolved interpretation remains disputed and cannot justify a fraud verdict merely by majority vote. Examples of potentially adjudicable violations include incompatible signed reports for the same commitment, an execution claim contradicted by an authenticated receipt, and omission of a required, receipted run. A previously unknown vulnerability, a defensible classification disagreement, or stochastic variation is not itself evaluator fraud.

Data unavailability suspends the report's evidentiary status. Whether a missed availability deadline is separately penalizable must be agreed in the contract, with attribution and appeal rules; it is not automatically proof of fabrication. Sensitive traces require authorized access, redaction rules, and a restricted reproducibility channel. A public digest can establish integrity while leaving outsiders unable to verify the underlying claim. The report must disclose that restriction.

Trust roots. The recorder, randomness source, evidence store, and final adjudicator are explicit dependencies. Their compromise can destroy detectability or increase erroneous penalties. Delegating judgment to additional token holders does not resolve this recursively. A deployable system must identify where factual authority finally resides and how errors can be corrected. This paper specifies the interface; it does not implement or certify those components.

Incentives Under Imperfect Adjudication

Consider one risk-neutral evaluator choosing honest completion H or a specified cheating strategy D. The fee is w, honest incremental effort costs c, and cheating yields an additional benefit b, including a bribe or avoided liability outside this contract. Set $G = c + b > 0$. We conservatively allow the evaluator to retain b even when detected. In this baseline, D avoids all contracted effort.

The report is independently audited with probability p after the strategy is committed. Conditional on an audit, a final penalty occurs with probability delta under D and alpha under H.

These probabilities include detection, evidentiary admissibility, appeal, and successful enforcement. The same fixed loss L = B + W applies in either case: B is uniquely allocated, collectible collateral and W is a forfeitable portion of the fee, 0 <= W <= w. All quantities are in one external unit of account. Ignoring a capital cost common to both strategies gives:

$$U_H = w - c - p \alpha L; \quad U_D = w + b - p \delta L. \quad (1)$$

Proposition 1 (report-level incentive condition). Under these assumptions, honest completion is a weak best response against D if and only if:

$$p (\delta - \alpha) (B + W) \geq G. \quad (2)$$

Proof. Subtracting the payoffs in (1) gives $U_H - U_D = p(\delta - \alpha)L - (c + b)$. Nonnegativity is exactly (2). A strict inequality gives a strict preference. Equality permits cheating as another best response. This comparison is not a dominant-strategy result in a game with adaptive auditors or multiple interacting evaluators.

Corollary 1 (bounded deviations). Suppose every permitted deviation d has gain $G_d \leq G_{max}$, detection $\delta_d \geq \delta_{min}$, and the honest false-penalty probability is at most α_{max}. Then $p(\delta_{min} - \alpha_{max})L \geq G_{max}$ suffices for honest completion against all such deviations. The guarantee is only as credible as those bounds. An undetectable deviation with positive gain defeats it, and no finite loss covers unbounded outside bribes.

Participation. If locking collateral costs κB per report and the outside option is zero, an honest evaluator also requires $w \geq c + \kappa B + p \alpha L$. Larger bonds can improve deterrence while increasing the fee needed to attract honest participants. If false penalties grow with reviewer count or challenge volume, that growth must enter α; holding it constant is a modeling choice, not a protocol guarantee.

Interpretation. Define the incentive margin $M = p(\delta - \alpha)L - G$. Positive M supports the chosen honest action under the model; it is not a probability that an agent is secure. If $\delta \leq \alpha$, no positive finite penalty deters a positive-gain deviation. Increasing punishment cannot compensate for an adjudicator that fails to distinguish the two behaviors.

The Joint Feasibility Region

Correlated audit failure. Let k reviewers inspect an audited report. With probability rho a common event makes all reviewers fail to produce an admissible detection: for example, a shared blind spot or coordinated capture. Otherwise each reviewer independently produces such a detection with probability r. Any valid detection reaches the specified adjudicator. Under this particular mixture model:

$$\text{delta}_k = (1 - \rho)\,[1 - (1 - r)^k]. \quad (3)$$

The no-detection probability outside the common event is $(1-r)^k$, which proves (3) by conditioning. The ceiling is $1-\rho$. This is not a universal law of reviewer correlation; arbitrary dependence need not follow the mixture model. False-penalty probability α_k is assessed separately after adjudication, rather than by counting accusations.

Budget and capacity. Suppose $B \le B_{max}$, each reviewer costs v, and each audited report incurs overhead a. For a fixed k with $\delta_k > \alpha_k$, the smallest audit probability consistent with weak deterrence at maximum collateral is:

$$p_{min}(k) = G / [(\delta_k - \alpha_k)(B_{max} + W)]. \quad (4)$$

The design is infeasible if $p_{min}(k) > 1$. If an operator budgets at most C_{max} for audits per report, feasibility additionally requires $p_{min}(k)(kv+a) \le C_{max}$. With J reports arriving per period and H reviewer assignments available, $J\,p_{min}(k)\,k \le H$ is also necessary. These are audit-cost conditions, not a full business profitability model. Storage, routine execution, insurance, capital compensation, acquisition, and dispute-tail costs must be budgeted separately.

Among feasible integer k, an audit-only design can minimize $p_{min}(k)(kv+a)$. More reviewers improve detection with diminishing returns while adding conditional cost. Consequently the cheapest feasible committee need not be the largest. Estimates of r and rho require adversarial measurements; counting nominally separate operators is insufficient.

Shared collateral. B is allocated to the report, not repeatedly counted across all active reports. If one bonded entity can cheat on m concurrent jobs before settlement, the aggregate gain can exceed a bond calibrated for one job. Either reserve disjoint collateral for each live obligation or analyze the joint deviation, enforcement rule, and aggregate exposure. An identity

reset must not release escrow before the appeal and enforcement window closes.

A measurable decision rule. Before accepting a job, estimate a conservative gain bound, a lower bound on detection, an upper bound on erroneous penalties, and collateral collectible during the dispute window. Refuse or rescope jobs outside (2)-(4). A stake amount chosen only from current token capitalization provides none of these measurements.

Reproducible Synthetic Study

The study answers how the declared model behaves; it does not estimate how well real evaluators detect misconduct. All probabilities and prices are scenario inputs. Monetary quantities are dimensionless external units, not dollars, observed fees, or token-price forecasts. The supplied script, seed, results, and build source are included with the paper.

Design. We use NumPy's default random generator with seed 20260915 and 200,000 trials per cell. For each rho in {0, 0.1, 0.3, 0.6} and k in {1,...,8}, the script draws a common-failure indicator and k independent reviewer outcomes, then records whether a valid detection occurs. It separately draws honest false penalties at alpha = 0.01. The resulting 32 cells check the implementation of the mixture model. We report exact model values, Monte Carlo estimates, and pointwise 95% Wilson intervals. These intervals quantify simulation noise, not uncertainty about real-world reliability.

Table 1

Declared Baseline and Cost Assumptions

Parameter	Value	Interpretation
G; B_max; W	10; 150; 5	Cheating gain; collateral limit; withheld fee
p; r; alpha	0.10; 0.65; 0.01	Audit probability; reviewer detection; false penalty
k; rho	3; 0.10	Reviewers; common-failure probability
v; a	1; 4	Reviewer cost; audit overhead

The baseline uses k = 3 and rho = 0.1. These illustrative values are neither calibrated nor recommended for deployment. The sweep includes substantially worse common failure to show sensitivity. We also enumerate k to minimize audit expenditure under the stated costs, without claiming to optimize all operating expenses.

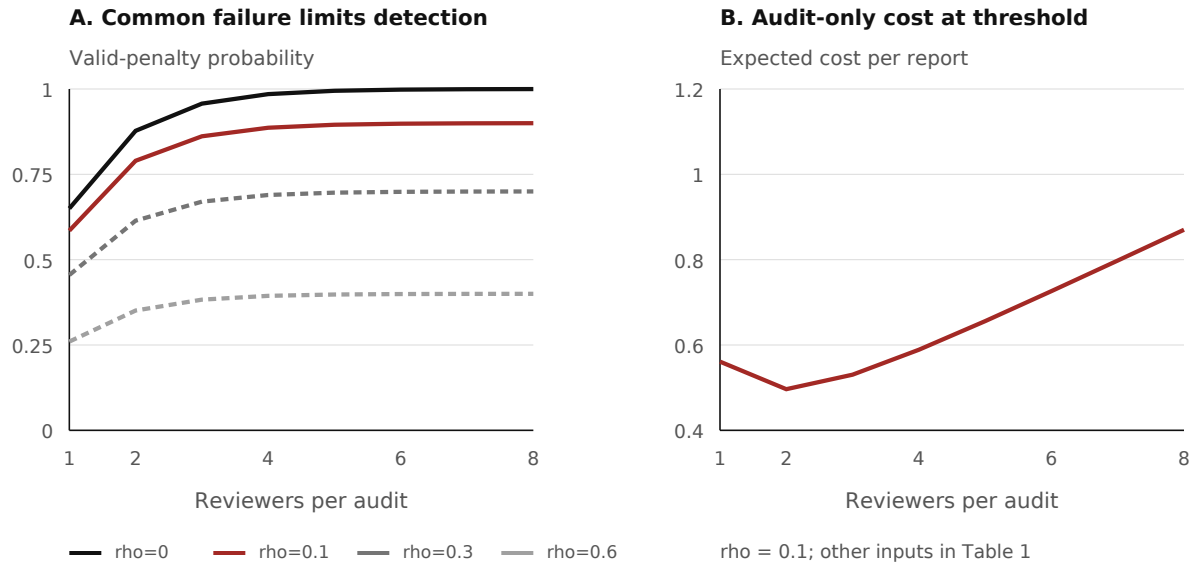
Collateral stress. We hold the baseline audit behavior fixed and reduce native collateral's value through 101 price ratios from 0 to 1. We compare 150 external units, 150 units of initially valued native collateral, and a mixed allocation of 125 external plus 25 native units. External collateral is held stable by assumption; custody, depegging, and liquidation risks would require additional haircuts. The experiment is a static stress calculation, not a model of endogenous prices or crash probabilities.

Omission sensitivity. A second simulation samples s required run identifiers without replacement from a manifest of 100, of which five are omitted. Detection means selecting at least one omitted obligation. The exact probability is 1 - choose(95,s)/choose(100,s). We compare it with 200,000 hypergeometric draws for each s in {1,5,10,20,40,60}. This assumes the omitted obligations are attributable and verifiable; sampling a submitted-only list would not detect their absence.

Results and Diagnostic Implications

Figure 1

Audit Detection and Cost Under Shared Failure



Note. Curves are exact model calculations, not measurements of deployed auditors. The audit-cost minimum assumes weak deterrence and the declared Table 1 inputs.

Detection and cost. Across the 32 audit cells, the largest absolute difference between simulated and exact detection is 0.001502; 31 of 32 pointwise Wilson intervals contain the exact probability. This is an implementation check with the expected possibility of pointwise interval misses. It is not evidence that deployed reviewers satisfy the inputs. At baseline, delta = 0.861413 and the incentive margin is +3.196894. The minimum audit probability is 0.075775; at p = 0.10 the required collateral is 112.451881 units.

With rho = 0.1, audit overhead 4, and reviewer cost 1, the cheapest weakly deterrent committee among k = 1,...,8 has k = 2 and expected audit cost 0.496437 per report. That optimum sits on M = 0 and leaves the evaluator indifferent. Deployment would require a margin above the threshold and uncertainty allowances; our enumeration does not choose those allowances.

Table 2

Detection of Five Omissions Among 100 Obligations

Sampled	Exact	Simulated	95% Wilson interval
1	0.0500	0.0508	[0.0498, 0.0517]
5	0.2304	0.2291	[0.2273, 0.2310]
10	0.4162	0.4159	[0.4138, 0.4181]
20	0.6807	0.6823	[0.6802, 0.6843]
40	0.9275	0.9267	[0.9255, 0.9278]
60	0.9913	0.9911	[0.9907, 0.9915]

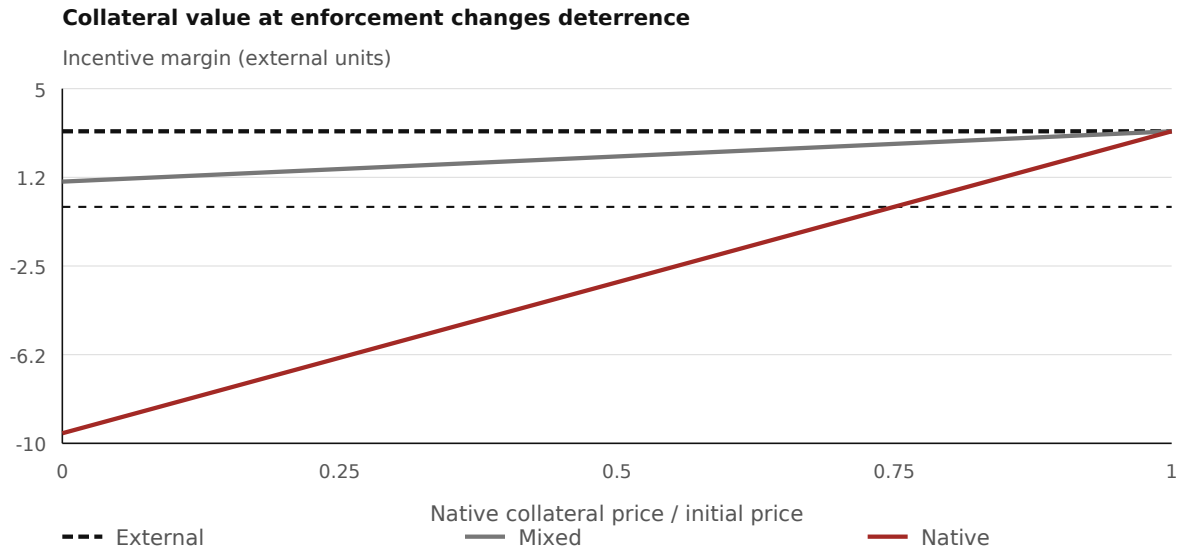
Note. Intervals quantify pointwise Monte Carlo sampling uncertainty, not uncertainty about real-world detection.

Sampling covers omissions slowly when few obligations are checked. This result concerns completeness of the contracted test set, not coverage of the space of possible attacks. A perfectly complete manifest can still contain an inadequate evaluation distribution. Increasing audit intensity cannot repair that scientific design error.

Collateral and a Token With Economic Substance

Figure 2

Incentive Margin Under Collateral Price Stress



Note. The three allocations begin at the same collateral value. A positive margin indicates strict preference for honesty only under the stated model assumptions.

Stress result. At an 80% decline in native value, the external-only allocation retains margin +3.20; native-only falls to -7.02; the mixed allocation retains +1.49. Each begins with the same 150-unit collateral valuation. The comparison isolates value at enforcement; it does not establish an optimal portfolio. If audit reliability also deteriorates during a crash, holding it fixed understates the joint stress.

A conservative collateral rule can use B_eff = B_ext + (1-h) P_stress Q, where Q is the allocated token quantity, P_stress is a stressed executable price, and h covers liquidation and settlement loss. Top-ups, lower exposure limits, or suspension are required when the incentive constraint fails. A last-traded price is not evidence that the whole position can be liquidated at that price.

Token equivalence. Within (1), two collateral instruments with the same collectible loss and identical enforcement probabilities yield the same incentive margin. This follows by substitution. A native token is therefore not necessary for the incentive result. Its advantages, if any, must come from measurable differences in access, coordination, financing, or operating

cost, and must outweigh added price and governance risks.

A testable value channel. Suppose buyers pay external revenue R for useful evaluations. After nonduplicated operating expenses O, audit expenditure A, and reserve additions Z, define distributable cash F = max(0, R-O-A-Z). If the protocol credibly allocates a fraction beta of F to S eligible participation units, the gross cash distribution per unit is beta F/S. This is an accounting identity for a proposed rule, not an implemented entitlement or valuation result. Individual slashing, taxes, dilution, capital costs, and service obligations can reduce net returns.

Growth in paid usage can increase that distribution only if costs, reserve needs, and dilution do not absorb it. Issuing tokens to pay participants does not itself produce external revenue. We therefore make three falsifiable economic requirements: buyers pay for the service; the service generates surplus after full costs; and the token captures a defined share of that surplus. None is demonstrated in this paper. A token can have a plausible value mechanism without a proven price trajectory.

Prospective Evaluation on Real Agents

The next study should test the mechanism's weakest assumptions. The protocol below is proposed and has not been registered or executed. Its final hypotheses, task selection, sample size, and exclusion rules should be frozen before outcome collection.

Task and attack construction. Start from an agent-security environment such as AgentDojo (Debenedetti et al., 2024), with model identifiers and dependencies pinned where possible. Stratify authorized tasks and prompt-injection attacks. Record legitimate-task completion and attacker success separately. Add evaluator-level treatments: omitted required runs, fabricated report fields, conflicting commitments, selective outcome reporting, and honest disagreements. Ground truth for misconduct comes from the controlled intervention and authenticated recorder, not from reviewer consensus.

Comparators. Compare a single paid evaluator, independently sampled audit without collateral, an otherwise identical audit with external collateral, and an audit with native collateral stressed to the same collectible value. Hold tasks, available evidence, audit budgets, and assignment rules constant. A native-token arm would require an actual

incentive-compatible setting with consequential payments; relabeling synthetic points does not establish economic behavior. Randomized financial treatments need appropriate participant consent and institutional review where applicable.

Primary estimands. Measure valid final penalties per audited malicious report, erroneous final penalties per audited honest report, total audit cost, and evaluator misconduct under each payment condition. Report uncertainty at the independent evaluator or task-cluster level, not as though repeated model calls were independent participants. Use a pilot to estimate clustering and inform a prospective power calculation. The present paper gives no invented sample-size justification.

Independence stress. Vary shared model families, ownership, common evidence summaries, and communication opportunities. Include a blinded adversary seeking a high-gain low-detectability deviation. Estimate detection by attack class; a high average cannot justify a lower bound for an untested or systematically missed class. Measure how final false penalties change with committee size and appeals. This tests the constant-alpha assumption directly.

Evaluation gates. First require that a conservative lower confidence bound on delta exceeds an upper confidence bound on alpha. Next substitute a documented gain bound and stressed collateral into (2), with an explicit positive margin. Then check the full cost of honest participation and operational capacity. If the bound on gain is unavailable, report conditional deterrence rather than a security guarantee. If no affordable configuration survives these checks, the mechanism should fail its deployment gate.

Commercial test. Separately measure willingness to pay through paid pilots, renewal, and full-cost contribution margin. Compare token-required access with ordinary payment and external collateral. These measurements concern economic usefulness; they must not be inferred from benchmark detection rates. Conversely, customer demand does not validate security claims.

Limits, Falsification, and Conclusion

Adjudication is the central dependency. The analysis takes conditional enforcement probabilities as inputs. A compromised recorder, coerced adjudicator, or persuasive fabricated trace can make those inputs false. Consensus among reviewers using the same weak verifier does not restore external ground truth. In a deployed system, alpha and delta may also change strategically with the penalty size. The closed-form analysis holds them fixed and does not solve that equilibrium.

Rationality and scope. Expected-payoff deterrence applies to the specified risk-neutral choices. It does not prevent sabotage by an actor with nonfinancial objectives, irrational behavior, unbounded external benefit, or access to profitable short positions not included in b. Repeated interactions, collusive side payments, endogenous entry, and purchaser misconduct require a richer game. Honest work can still be incompetent: meeting a reporting contract is distinct from designing a sufficiently informative test.

Statistical limits. A negative result is meaningful only relative to a sampling distribution and test power. For n independent draws from a fixed distribution, zero observed failures gives the one-sided 95% binomial upper bound 1 - 0.05^(1/n). At n = 100 this is 2.9513%, not zero. Adaptive attack selection, dependent trials, distribution shift, and benchmark contamination invalidate a naive interpretation of this bound. Neither an immutable report nor a bond changes the sampling assumptions.

Evidence and disclosure. Real traces may contain sensitive user data or reproducible exploits. Access controls and coordinated release can limit external reproducibility; that tradeoff belongs in the report. This study contains no real-user traces or operational exploitation experiment. Its numerical precision reflects reproducible arithmetic, not precision about the world.

What would refute the proposal? An affordable audit that cannot distinguish honest from malicious reporting; adversarial gains above all collectible collateral; honest participants priced out by locking costs; or repeated inability to attribute claimed violations would each defeat the relevant feasibility argument. Lack of paid demand would separately defeat the proposed surplus-based token value channel. These are substantive failure conditions, not

parameters to conceal through a larger token supply.

Conclusion. A security-evaluation network earns credibility when another party can inspect what was promised, what ran, what failed, and why a dispute was resolved. Naysay couples that evidence trail to a limited economic claim: under explicit detection, error, and enforcement assumptions, honest evaluation can be made the better-paying action. The same analysis shows when the claim fails. That boundary is the appropriate starting point for an empirical system and for a token whose value must ultimately be supported by useful, paid work.

Authorship and assistance. This working draft, its analysis code, and its figures were prepared with AI assistance using Codex and Orchestra Research's ml-paper-writing workflow. No external peer review, funding award, or institutional approval is claimed.

References

Brown-Cohen, J., Irving, G., & Piliouras, G. (2023). *Scalable AI safety via doubly-efficient debate* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2311.14125

Cong, L. W., Li, Y., & Wang, N. (2020). *Tokenomics: Dynamic adoption and valuation* (NBER Working Paper No. 27222). National Bureau of Economic Research. https://doi.org/10.3386/w27222

Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A., & Tramèr, F. (2025). *Defeating prompt injections by design* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2503.18813

Debenedetti, E., Zhang, J., Balunović, M., Beurer-Kellner, L., Fischer, M., & Tramèr, F. (2024). *AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2406.13352

Douceur, J. R. (2002). The Sybil attack. In P. Druschel, F. Kaashoek, & A. Rowstron (Eds.), *Peer-to-peer systems* (pp. 251-260). Springer. https://doi.org/10.1007/3-540-45748-8_24

Gao, A., Wright, J. R., & Leyton-Brown, K. (2016). *Incentivizing evaluation via limited access to ground truth: Peer-prediction makes things worse* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1606.07042

Greenblatt, R., Shlegeris, B., Sachan, K., & Roger, F. (2023). *AI control: Improving safety despite intentional subversion* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.06942

Hossain, S., Andrade, G., Zang, C., & Chen, Y. (2026). *Evidence markets* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2606.07434

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). *Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.00973

Teutsch, J., & Reitwießner, C. (2019). *A scalable verification solution for blockchains* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1908.04756

Appendix

Derivations and Auditability

Unequal penalties. The fixed-loss model is convenient but restrictive. If the loss conditional on a final penalty is L_D under cheating and L_H under honesty, subtraction yields p(delta L_D - alpha L_H) >= G. If penalty amounts are random, replace each term by the corresponding conditional expected collectible loss, provided the evaluator is risk neutral. Substituting a headline maximum slash for the actual expected loss overstates deterrence. For distinct honest and dishonest effort costs c_H and c_D, substitute G = c_H - c_D + b; the participation constraint uses c_H. An appeal can reduce false penalties while increasing locking costs through delay.

Audit selection leakage. Let q_D be the probability of audit conditional on the chosen deviation and q_H the audit probability under honesty. The general condition is q_D delta L - q_H alpha L >= G. Equation (2) assumes both are p. If an attacker identifies unaudited jobs before choosing D, q_D can approach zero. This explains why the report must be committed before audit selection becomes predictable. It is not enough for the operator to audit p of all reports on average.

Robust bounds. For every deviation d, U_H-U_d >= p(delta_min-alpha_max)L-G_max under the bounds in Corollary 1. Nonnegativity establishes the sufficient condition simultaneously. No probability distribution over deviations is needed for that algebra; obtaining valid bounds over a meaningful attack class is the difficult empirical task. If only average estimates are available, the claim must remain average-case.

Omission sampling. Of choose(N,s) equally likely subsets of s required obligations, choose(N-m,s) contain none of m omitted obligations. Subtracting that fraction from one gives the detection probability used in the Reproducible Synthetic Study section. The expression is zero for s=0 and one when s>N-m. It assumes the manifest includes all obligations and that a sampled omission becomes an admissible violation. Neither assumption follows from hashing a partial report.

Reproduction contract. Run simulate.py to regenerate three CSV files and summary.json. The deterministic stress sweep has no sampling interval; the Monte Carlo tables

use pointwise Wilson intervals with z = 1.959963984540054. There is no multiple-comparison coverage claim. The code checks payoff subtraction, boundary equality, limiting detection, price monotonicity, and numerical simulation error. These checks establish implementation consistency, not the truth of the model assumptions.

The artifact includes manuscript.md, verified references.bib, bibliographic provenance, build_pdf.py, figures, numerical outputs, and a pinned Python requirements file. The website offers the PDF and a source archive. A hash manifest records the released artifact bytes. Downloaded full-text literature is excluded from the archive; references link to the original sources. The PDF can be rebuilt from the supplied source without paid APIs or private model credentials.