Reputation is Amortized Verification:

Inspection Games for Agent Economies

Paper 3 of the Harbor research program — B2/R7, executed

Prepared for Erich Owens

August 2026

Abstract

A marketplace where AI judges grade AI work is an inspection game. We give three results. *Stage:* a bonded judge is deterred iff $\rho dB \geq G$; the critical audit rate is $\rho^\star = G/(dB)$ under the keep-gain convention and $\rho_c^\star = G/(d(G+B))$ under confiscation — the settlement convention alone moves the headline audit budget by roughly 17%. *Tower:* when level $k{+}1$ audits level-k auditors sampled *sealed* from a pool spanning C disjoint cliques, rational bribery is all-or-nothing and profitable only while corrupt value exceeds $C\,B$; below that threshold corruption decays geometrically at $(1-\rho d)$ per level, so *finite* bond capital certifies a tower deep enough to drive corrupt value below any fixed unit — a depth logarithmic in that value. *Amortization:* as a judge's verified history grows, incentive-compatible audit schedules decline — lifetime audit spend falls from $\Theta(T)$ (flat) to $\Theta(\log T)$ (losses only if audited) to $O(1)$, with the constant $aG^2/(2dBrv)$ (cheats also surface at an independent rate r). Reputation, on this account, is not a soft social layer bolted onto verification: it is the mechanism by which verification cost is amortized over a history. The results convert the source volume's central open conjecture (III.11.1) into a parameterized theorem with a named implementation, and every number reproduces from a bundled script.

Express lane: deterrence holds iff audit-probability × detection × bond ≥ corrupt gain; stacking auditors sampled sealed from C independent cliques shuts off the bribery-profitable phase, leaving a geometric collapse that finite total bond certifies at logarithmic depth; and history-indexed audit schedules make lifetime verification spend $\Theta(\log T)$ or $O(1)$ instead of $\Theta(T)$ — formal statements in the boxes of §4, §5, §6.

1 Introduction

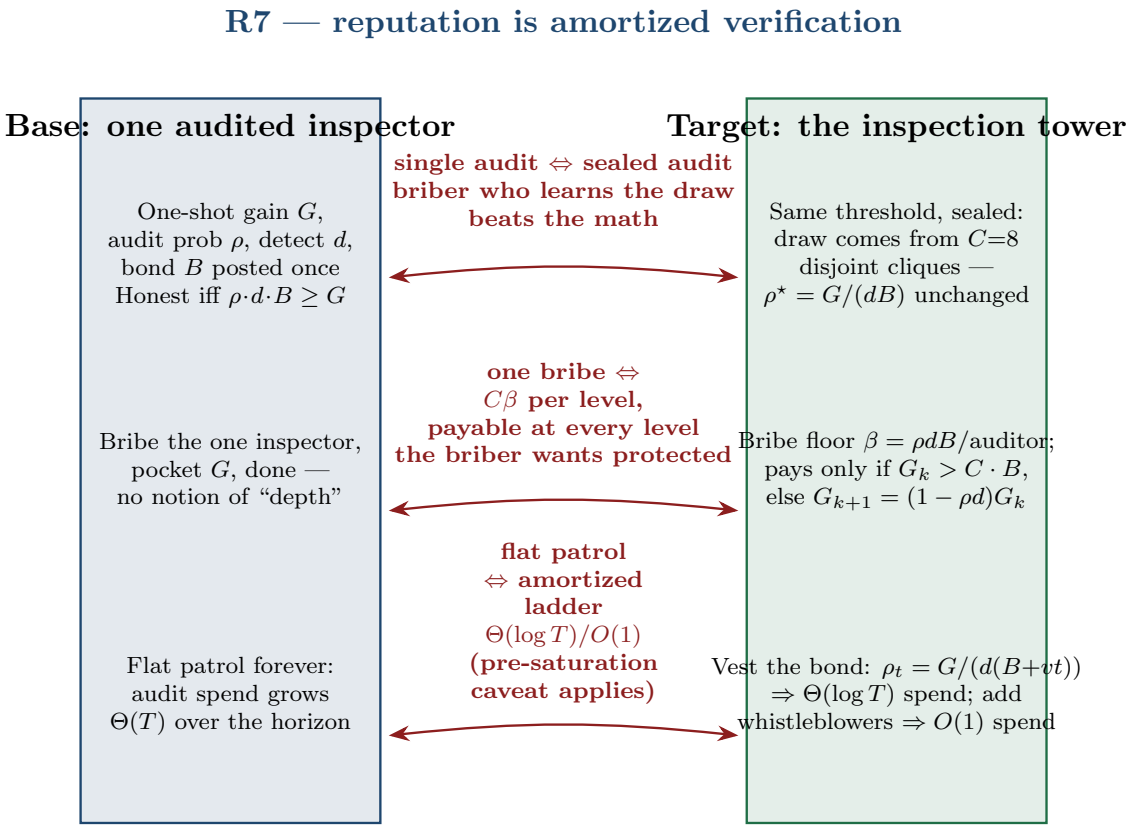
The scene. A marketplace where AI agents do work and AI judges grade it. A judge posts a bond that is slashed if it is caught cheating. A judge can take a bribe: pass shoddy work, pocket G. So you audit the judges — but auditors can be bribed too, and now it looks like watchers all the way up, each level dearer than the last. Meanwhile the platform's accountant asks a different question: a judge with ten thousand verified grades is still being audited at the same rate as a newcomer — is that spend buying anything?

The idea in one breath.

Honesty holds when expected punishment beats the bribe; the tower of auditors converges on finite bond capital because each level has geometrically less corruption left to protect, and drawing judges sealed from C independent camps buys down the depth by shutting off the phase where bribery still pays; and the audit rate an old judge needs falls with the reputation it would forfeit — so reputation is amortized verification.

The idea, before the formalism. Traffic enforcement does not post an officer on every corner; it makes expected fine × odds of being caught exceed the minutes saved by speeding.

That is the stage game: gain G, audit probability ρ, detection probability d, slashable bond B. The tower adds jury selection from disjoint pools: if each auditing layer is drawn sealed from C rival benches, a briber must pay all C benches to be safe — worthwhile only for enormous corruption, which then bleeds value paying for its own protection. And the amortization result is the actuarial half of the same logic: a judge with a long verified history has a large future income stream at stake, which is itself a punishment the platform did not have to fund — so the platform can audit it less, and the incentive constraint stays exactly binding. Figure 1 draws the correspondence. *Misread to preempt:* the tower result does *not* require infinitely many honest auditors, and the amortization result does *not* assume old judges are intrinsically nicer. The former needs only finite bonds and a sealed draw, because what each level protects shrinks geometrically once bribery stops paying; clique diversity ($C > 1$) buys depth and capital, not convergence. The latter needs only that accumulated reputation is genuinely forfeit on capture — deterrence is re-derived, not assumed, at every t.



The tower prices depth, not detection: finite bond, sealed against the draw, certifies a depth logarithmic in the corrupt value — 27 levels at C=8, 53 at C=1.

Figure 1: The relation map: the stage game (left) and the tower (right) are the same deterrence arithmetic, applied once and applied recursively. The arrows carry the load — a single committed audit maps to a *sealed* draw from C benches; one bribe maps to a bribe bill of $C\beta$ that must be paid *at every level* the briber wants protected; and a flat patrol budget maps to a history-indexed schedule. The relation preserved across all three rows is *expected forfeiture* $\geq$ *corrupt gain*; what changes is only who pays it and over what horizon. What C buys is depth and capital (53 levels and 2650 at C=1 against 27 and 1350 at C=8), not the difference between a tower that terminates and one that does not.

Contributions. (i) The stage-game threshold with its convention fork made explicit (§4) — the same game yields $\rho^\star = 0.25$ or $\rho_c^\star = 0.2083$ at our running parameters depending on whether capture confiscates the gain, and quoting one without pinning the settlement rule is a repro-

ducible way to be 17% wrong. (ii) The tower-contraction theorem (§5): sealed sampling from C disjoint cliques makes bribery uneconomical below corrupt value $C\,B$, after which corruption decays geometrically — so finite bond capital certifies an audit depth logarithmic in the corrupt value, with sealing identified as critical rather than hygienic and the clique multiplier priced honestly (it removes the linear bribery phase; it is not what makes the tower converge). (iii) The amortization theorems (§6): flat auditing spends $\Theta(T)$; the loss-only-if-audited schedule spends $\Theta(\log T)$; independent revelation at rate r gives $O(1)$ lifetime spend, with the constant $aG^2/(2dBrv)$ — and which of the two holds is an empirical fork, not a modeling taste. (iv) The connection to the program's repeated-game mechanization (§7): the discount-factor thresholds $\delta \geq 1/3$ (grim) and $\delta^\star = 0.3425080314$ (graduated trigger), now CI-checked by Z3 and TLC/Apalache in the repository's `proofs/economics/` directory, are the same deterrence arithmetic with continuation value in place of bonds. (v) The conversion of the source volume's Conjecture III.11.1 into a parameterized theorem with a named implementation (§8). Every number is tagged [verified] (externally recomputable) or [internal] (regenerates from the bundled script `b2_tower.py`).

2 Related work, and what is actually new

The stage game is classical. Deterrence-by-expected-penalty is Becker [2]; the leader–follower and simultaneous inspection formulations, including the committed-inspector solution we use, are surveyed by Avenhaus, von Stengel and Zamir [1], whose treatment of inspector commitment and inspectee indifference we follow. Reputation-as-equilibrium-behavior is Kreps–Wilson [3] and Fudenberg–Levine [4]: a long-lived player's continuation value disciplines short-run temptation. Batched verification of many items with few tests goes back to Dorfman's group-screening design [5], which is the right lineage for pooled audit sampling, though we do not use pooling here.

New, honestly stated: the amortized-verification framing — deriving the *declining audit schedule* that keeps the incentive constraint exactly binding as history accumulates, and costing it end-to-end ($\Theta(T)$ vs. $\Theta(\log T)$ vs. $O(1)$, with the closed form $aG^2/(2dBrv)$); the contraction theorem for a *stackable* audit tower whose auditors are themselves corruptible LLM judges, with model heterogeneity identified as the mechanism that supplies the disjoint cliques the contraction needs; and the observation that sealing the sampling draw is critical (§5). *Imported:* everything in the previous paragraph. The contribution is closing a conjecture the source volume posed, with a named implementation and a measurement obligation — not a new solution concept.

3 The vocabulary, defined

Express lane: skip to §4 — nothing in this section is needed to follow the theorems, only to read them comfortably; the two terms that are hypotheses rather than vocabulary, sealed *and* clique*, are defined where they first do work, at the head of §5.* This paper is economics, and its terms are cheap to define but expensive to guess at. A software reader who has never taken a mechanism-design course loses nothing but the dictionary; here it is, in the order the terms are needed.

An *inspection game* is a two-player game between someone who may cheat and someone who may check, where checking costs something and catching pays something. Its defining feature — and the reason it is not simply a matter of "audit more" — is that it typically has no equilibrium in which either player behaves predictably: if the inspector always checks, cheating never pays and the checks are wasted; if the inspector never checks, cheating always pays. The stable outcome is that both randomise, which is why every audit rate in this paper is a probability rather than a schedule.

A *mixed strategy* is exactly that randomisation — a probability distribution over what to do rather than a choice. Two players are in a *mixed equilibrium* when each is *indifferent* between its options given what the other is doing, since a player who strictly preferred one option would simply take it. Solving for indifference is how the critical rates below are derived, and it is the whole technique.

A player's behaviour is *incentive-compatible* when honesty is at least as profitable as the alternative — not when it is enforced or morally preferred. Everything this paper deters, it deters by making cheating cost more than it pays.

A *bond* is capital the judge posts before working and forfeits if caught cheating. It is the paper's central instrument, and it differs from a wage in a way that matters: a wage is paid for output and withheld for non-performance automatically, whereas a bond is forfeited only when a *detection event* occurs. That gap — between misbehaving and being caught misbehaving — is what the audit probability ρ and detection probability d measure, and it is why ρdB rather than B is the quantity a rational judge weighs. To *slash* a bond is to confiscate it on detection; the term is borrowed from proof-of-stake systems, where it means the same thing.

The two *settlement conventions* of §4 are the two natural answers to what happens to the corrupt gain itself when a cheat is caught. Under *keep-gain* the cheater keeps G and loses the bond; under *confiscation* the cheater loses both. They are not a modelling detail — they move the required audit budget by roughly 17% at the worked parameters, which is the paper's first practical finding.

A *bribe floor* is the least a briber must offer a judge to make corruption worth its while: exactly the judge's expected forfeiture $\beta = \rho dB$. Below that, no rational judge accepts.

Finally, *amortization* is the accounting sense of the word, not the algorithmic one: spreading a fixed cost over many uses so the per-use share falls. The paper's title claim is that a reputation *is* this operation applied to verification cost — the reason a long-trusted judge can be audited less is that its history has already paid for the confidence.

4 The stage game and the convention fork

A judge grades one outcome. Cheating yields a one-shot gain G. The platform audits with probability ρ at cost a per audit; an audit detects a cheat with probability d; detection slashes the judge's posted bond B. Under the *keep-gain* convention the caught cheat retains G and forfeits B; under the *confiscation* convention capture forfeits the gain as well.

Theorem 1 (Stage deterrence)**.** *A judge with one-shot corrupt gain G, facing audit probability ρ, detection probability d, and slashable bond B, is deterred iff $\rho dB \geq G$; the critical committed audit rate is*

$$\rho^\star \;=\; \frac{G}{dB} \qquad \textit{(keep-gain; Becker's condition).}$$

If capture also confiscates the gain, the cheat payoff is $G - \rho d(G+B)$ and the critical rate falls to

$$\rho_c^\star \;=\; \frac{G}{d(G+B)} \;<\; \rho^\star.$$

In the simultaneous (uncommitted) game, with a the audit cost and L the auditor's loss from an uncaught cheat, the mixed Nash equilibrium is $\rho^\star = G/(dB)$, $q^\star = a/(dL)$: the judge's indifference pins the audit rate, the auditor's indifference pins the cheat rate.

Why the same $\rho^\star$ twice. The formula is unchanged across the two game forms not by coincidence but because both solve the same judge-indifference equation: commitment picks the cheapest rate that deters, and simultaneous play lands an equilibrium mix exactly there.

Proof. Keep-gain: the cheat payoff is $G - \rho dB$, nonpositive iff $\rho \geq G/(dB)$. Confiscation: replace B by $G + B$ in the forfeiture. Mixed equilibrium: the judge mixes only if indifferent, $G - \rho dB = 0$; the auditor mixes only if indifferent between auditing $(-a - qL(1-d))$ and not $(-qL)$, which solves to $q^\star = a/(dL)$. Direct best-response checks confirm both indifferences numerically (§9). □

Numbers by hand. At the program's running parameters G=10, d=0.8, B=50: $\rho^\star = 10/(0.8{\cdot}50) = 0.25$ [verified] — audit one grade in four, and the cheat payoff is exactly +0.000000 at $\rho^\star$ [internal, `b2_tower.py`]. Under confiscation, $\rho_c^\star = 10/(0.8 \cdot 60) = 0.2083$ [verified]. The convention moves the required audit budget from 0.25 to 0.2083 — a 16.7% cut for free, *if and only if* your settlement rail actually claws back the gain. With a=1, L=40: $q^\star = 1/(0.8 \cdot 40) = 0.03125$ [verified]. Figure 2 plots the frontier $\rho dB = G$ in the (B, ρ) plane with this point marked. *Now you try:* double the bond to B=100 — what is $\rho^\star$? (0.125: a bigger bond lets you audit half as often; capital and vigilance are substitutes along $\rho dB = G$.)

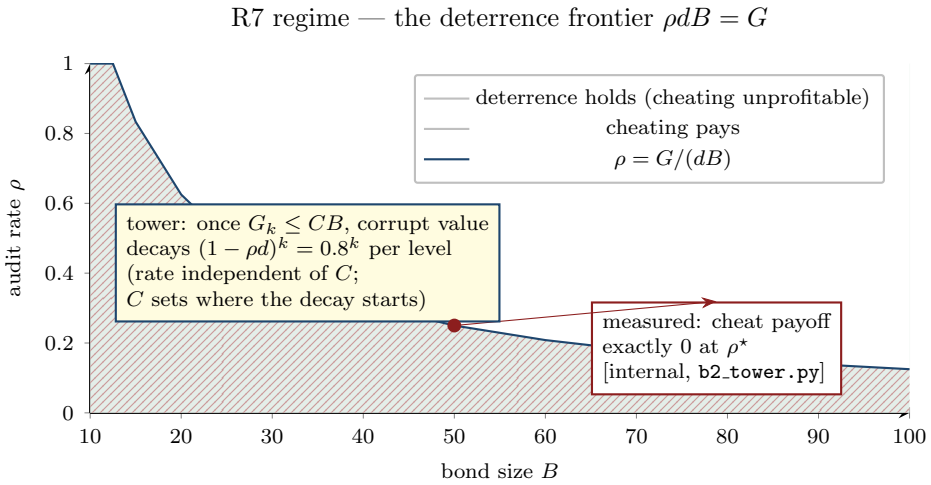


Figure 2: The regime diagram: the deterrence frontier $\rho dB = G$ in the (bond B, audit rate ρ) plane at G=10, d=0.8. Above the curve $\rho = G/(dB)$, cheating is unprofitable; below it, cheating pays. The marked point is the running example $(B, \rho^\star) = (50, 0.25)$, where the measured cheat payoff is exactly zero [internal, `b2_tower.py`]. Below the curve, the region is hatched rather than tinted so the two half-planes stay distinguishable in greyscale. The inset records the tower consequence of §5: below the bribery threshold $G_k \leq CB$, surviving corrupt value decays as $(1-\rho d)^k$ — a $\times$0.8-per-level collapse at these parameters, at every C. What C controls is not the rate but where the decay begins.

5 The tower: bribing sealed juries from C cliques

Two terms first, because both are hypotheses of the theorem below rather than hygiene recommendations. *Sealed* sampling means the briber must commit its money before learning which judge will handle the case; the entire clique multiplier vanishes without it, as the sealing discussion after the box shows. A *clique* means a pool of judges whose errors and loyalties are statistically and economically independent of another pool's — distinct vendors running distinct model families. The word is borrowed from social structure, not from graph theory, and it carries no adjacency-matrix meaning in this paper.

Who audits the auditors? Let level $k{+}1$ audit level-k auditors. An auditor at any level is itself a bonded judge, so Theorem 1 prices its honesty: the minimum bribe an auditor rationally accepts is its expected forfeiture $\beta = \rho dB$. The platform's one structural lever is *where auditors come from*: each audit is sampled, sealed, from a pool spanning C disjoint cliques — distinct

principals, distinct model families — such that a corrupt agreement inside one clique does not reach the others. A briber protecting corrupt value G_k at level k faces a choice: bribe, or absorb the audit.

Theorem 2 (Tower contraction)**.** *Let level $k{+}1$ audit level-k auditors, each audit sampled sealed (the draw unknown to the briber) from a pool spanning C disjoint cliques, each auditor bonded at B with audit parameters (ρ, d) and bribe floor $\beta = \rho d B$. Then the briber's expected net value is affine in the number of cliques bought, $c \in \{0, \dots, C\}$: buying c still leaves the un-bought $(C{-}c)/C$ share of the sampling weight exposed to the same per-draw audit-and-detect risk ρd, so the payoff has no interior optimum and rational bribery is all-or-nothing — partial bribery is (weakly) dominated by one of the two corners. Bribing all C is profitable iff*

$$G_k\, \rho d \;>\; C\beta \quad\Longleftrightarrow\quad G_k \;>\; C\, B.$$

Below that threshold bribery stops and corrupt value decays geometrically:

$$G_{k+1} \;=\; (1 - \rho d)\, G_k.$$

Consequently finite *bond capital certifies a tower deep enough to drive surviving corrupt value below any fixed unit, at B per level. If $G_0 \le CB$ the decay is geometric throughout and the certified depth is $\lceil \log G_0 / \log \frac{1}{1-\rho d} \rceil$. If $G_0 > CB$, a linear phase of $\lceil (G_0 - CB)/(C\rho d B) \rceil$ levels precedes it — during which the briber bleeds $C\beta$ per level — followed by $\lceil \log(CB) / \log \frac{1}{1-\rho d} \rceil$ geometric levels. Either way the depth is finite: logarithmic in the corrupt value that remains once bribery has stopped paying, plus that linear prefix if it was still paying at level 0. Raising C shortens or deletes the linear phase; it does not change the geometric rate $(1 - \rho d)$, which is what makes the tower converge.*

Proof sketch. Buying c of the C cliques costs $c\beta$ up front (the seal forces payment before the draw is known) and retains value G_k if the draw lands on a bought clique (probability c/C) or $(1 - \rho d)G_k$ if it lands on an honest one (probability $(C{-}c)/C$). Expected net value is

$$V(c) \;=\; G_k\Big(1 - \tfrac{C-c}{C}\rho d\Big) - c\beta \;=\; V(0) + c\Big(\tfrac{G_k \rho d}{C} - \beta\Big),$$

affine in c, so the maximum over $c \in \{0, \dots, C\}$ sits at a corner: bribery is all-or-nothing. Bribing beats not bribing iff $V(C) > V(0)$, i.e. $G_k \rho d > C\beta = C\rho d B$, i.e. iff $G_k > CB$; the audit bite ρd cancels, leaving a pure capital comparison between the value protected and the bonds that must be compensated. Once $G_k \le CB$ the geometric recursion takes over, and the level count to drive the surviving value below one unit from there is $\lceil \log \min(G_0, CB) / \log \frac{1}{1-\rho d} \rceil$, each level requiring one bond B; if $G_0 > CB$, the linear phase adds $\lceil (G_0 - CB)/(C\rho d B) \rceil$ levels ahead of it. □

Why sealing is critical, not hygienic. The factor C in the profitability threshold exists *only because the briber cannot see the draw.* If the sampling draw leaks — the briber learns *which* auditor will handle its case before the bribe is placed — it need buy exactly one auditor at price β, and bribery is profitable iff $G_k \rho d > \beta$, i.e. iff $G_k > B$: the clique multiplier vanishes and heterogeneity buys nothing. A leaked draw converts an eight-bench jury system into a single corruptible clerk. Sealed randomized selection (in deployment: VRF-selected honeypot audits, §8) is therefore part of the theorem's hypothesis, and auditing the *sampler* is part of the security boundary.

Numbers by hand. $G_0 = 400$, parameters as above, $\rho = \rho^\star = 0.25$, so $\beta = \rho d B = 10$ and $\rho d = 0.2$. At $C{=}1$: the profitability threshold is $G_k > 50$, so bribery still pays at first, and corruption bleeds only the bribe bill $C\beta = 10$ per level — $400 \to 390 \to 380 \to \cdots$ — for 35 levels, until G_k reaches $50 = CB$ and the geometric phase takes over for $\lceil \log 50 / \log 1.25 \rceil = 18$ more: 53 levels, $53 \times 50 = 2650$ of bond capital [verified, arithmetic; regenerates from

`b2_tower.py`]. At C=8: the threshold is $G_k > 400$, not met even at level 0, so there is no linear phase at all and the collapse is geometric from the start at $(1 - \rho d) = 0.8$ per level: $400 \to 320 \to 256 \to 204.8 \to \cdots$, reaching below one unit in $\lceil \log 400 / \log 1.25 \rceil = 27$ levels, so total bond capital $27 \times 50 = 1350$ [verified, arithmetic].

What the clique multiplier actually buys. Set the two runs side by side and the multiplier is precisely a factor of two in certified depth and capital — 53 levels and 2650 against 27 and 1350 — not the difference between a tower that converges and one that does not. Both converge. Raising C *removes the linear phase*, because the profitability threshold CB rises past G_0; the geometric phase, and the rate $(1 - \rho d)$ that drives it, are common to both. That is a real and budgetable saving, and it is the whole of the claim. *Now you try:* at C=2, what is the per-level bleed while bribery remains profitable, and where does the linear phase stop? ($C\beta = 20$ per level: $400 \to 380 \to 360 \to \cdots$, until $G_k \le CB = 100$ after 15 levels, where the decay turns geometric for a further 21 — 36 levels, 1800 of bond.)

Why the per-level condition suffices. $V(c)$ scores a *single* level's deviation, so $G_k > CB$ is an iff for one-level bribery, not for a bribery strategy. The distinction matters because buying level k+1 protects G_k only from that level's audit: level k+2 still threatens it, so a briber committed to the programme pays $C\beta$ at every level for as long as the tower runs. Scoring the strategy rather than the level makes bribery *less* attractive, not more, so the contraction is unaffected — at C=1 the all-bribe path spends $35 \times 10 = 350$ to shepherd $G_0 = 400$ down to 50 and then loses the 50 geometrically anyway, burning 87.5% of the protected value to arrive where it would have arrived regardless [verified, arithmetic]. The per-level condition is therefore the weaker and the safer of the two to state: it is what the affine argument establishes, and any multi-level bribery programme is dominated by it a fortiori.

6 Reputation is amortized verification

The stage game prices one grade. A platform judges T of them per judge lifetime, and flat auditing at $\rho^\star$ costs $a\rho^\star T = \Theta(T)$. But an old judge is not a newcomer: it has a verified history of length t, and with per-period reputation rent v its accumulated stake vt — the future income its record commands — is forfeit on capture. That stake is punishment capital the platform did not fund, and an incentive-compatible audit schedule may spend it.

Theorem 3 (Amortization)**.** *Let a judge at history length t have reputation-at-stake vt, and let the audit schedule ρ_t be chosen so the one-shot deviation is weakly unprofitable at every t. Then:*

1. ***(Flat.)*** *The history-blind schedule $\rho_t \equiv \rho^\star = G/(dB)$ is incentive compatible and spends $a\rho^\star T = \Theta(T)$.*
2. ***(Model A: loss only if audited.)*** *If a cheat is penalized only when an audit catches it, and capture forfeits both bond and stake, the binding schedule is $\rho_t = G/\big(d(B+vt)\big)$, and lifetime spend is $a\sum_{t<T}\rho_t = \frac{aG}{dv}\ln\!\big(1+\frac{vT}{B}\big) + O(1) = \Theta(\log T)$.*
3. ***(Model B: independent revelation.)*** *If cheats additionally surface without audits at rate r per period (whistle-blowers, downstream failures, tombstoned provenance), destroying the stake vt when they do, the binding schedule is $\rho_t = \max\!\big(0, \frac{G-rvt}{dB}\big)$, which reaches zero at $t^\star = G/(rv)$; lifetime spend is bounded by the closed form*

$$a\sum_{t\ge 0}\rho_t \;\longrightarrow\; \frac{aGt^\star}{2dB} \;=\; \frac{aG^2}{2dBrv} \;=\; O(1).$$

Under each schedule the deviation value is ≤ 0 pointwise at every t: reputation does not

weaken deterrence, it funds *it.*

Proof. (2) The deviation value at t is $G-\rho_t d(B+vt)$; setting it to zero gives the schedule, and the spend is a harmonic-type sum: $a\sum_t G/(d(B+vt)) \approx \frac{aG}{dv}\ln(1+vT/B)$. (3) The deviation value is $G-\rho_t dB - rvt$ (audit slash plus the expected independent loss); the binding schedule is linear and hits zero at $t^\star = G/(rv)$, so the spend is the area of a triangle with base $t^\star$ and height $a\rho^\star = aG/(dB)$: $\frac{1}{2}\cdot\frac{G}{rv}\cdot\frac{aG}{dB} = \frac{aG^2}{2dBrv}$, independent of T. □

Numbers by hand — with the caveat a referee would otherwise find. Parameters G=10, d=0.8, B=50, a=1, v=0.6, r=0.05, horizon T=200. Flat spend: $1\cdot 0.25\cdot 200 = 50.00$ [verified]. Model A: 25.58 measured against closed form $\frac{aG}{dv}\ln(1+vT/B) = 25.50$ [internal, `b2_tower.py` / verified]. Model B: measured spend 35.08 against the $O(1)$ limit $aG^2/(2dBrv) = 41.67$ [internal/verified]. **Note that** $35.08 \neq 41.67$**, and the discrepancy is not error:** at these parameters $t^\star = G/(rv) = 333 > T = 200$, so the audit schedule has not yet reached zero within the measured horizon — *the* $O(1)$ *curve had not saturated at* T=200. The measured 35.08 is the partial triangle; the closed form is its full area, attained only for $T \geq t^\star = 333$. Run the horizon past $t^\star$ and the cumulative spend flattens and stays there. Figure 3 shows both halves of this: panel A is the measured range, where all three curves are near-linear and the $O(1)$ curve is in fact the second-steepest; panel B runs the same three schedules to T=1500, where flat reaches 375.00, Model A 61.46 and is still climbing, and Model B stops dead at 41.79 from $t^\star$ onward and never spends again [verified, `paper3_amortization_figures.py`]. (The plateau is 41.79 rather than 41.67 because the closed form is the area of the continuous triangle while the schedule is summed period by period; the excess is exactly half the first term, $\rho^\star/2 = 0.125$.) That is why the pre-asymptotic gap at T=200 is a check on the closed form rather than a strike against it: read against $t^\star$, not against the limit, 35.08 is precisely the partial area the theorem predicts. The incentive checks pass pointwise: maximum deviation value $+1.78\times 10^{-15}$ under both schedules — machine-epsilon zeros, i.e. ≤ 0 up to float epsilon, exactly the binding-constraint signature [internal]. *Now you try:* at what T do Model A and flat differ by a factor of 4? (Solve $\frac{aG}{dv}\ln(1+vT/B) = \frac{1}{4}a\rho^\star T$; at these parameters $T \approx 779$ — the log curve's lead widens without bound.)

The fork is empirical. Whether a deployment lives in Model A or Model B is a question about the world, not the model: *do cheats surface without audits?* If bad grades eventually reveal themselves — failed downstream builds, provenance tombstones, user reports — then $r > 0$ and lifetime verification spend per honest judge is a constant. If a cheat undetected at audit time is silent forever, only the $\Theta(\log T)$ schedule is safe. The deployment telemetry that estimates r (the rate at which sanctioned outcomes were first flagged by non-audit channels) is therefore part of this paper's measurement obligation, not an implementation detail.

7 The discount-factor connection: claim signaling, mechanized

The program's repeated-game track prices the same deterrence with a different currency. In the corrected claim-signaling stage game — payoffs $(3,3)$ mutual-truth, $(0,4)/(4,0)$ sucker/defector, $(1,1)$ mutual-claim — the one-shot deviation gain is $g = 1$ and the per-round punishment loss is 2. By the one-shot deviation principle, a grim trigger is subgame-perfect — cooperation survives every history, not just the equilibrium path — iff $1 \leq 2\,\delta/(1-\delta)$, i.e. $\delta \geq 1/3$; the gentler three-round graduated trigger is subgame-perfect iff $1 \leq 2(\delta+\delta^2+\delta^3)$, whose threshold is the unique real root in $(0,1)$ of $2\delta^3+2\delta^2+2\delta-1=0$, numerically $\delta^\star = 0.3425080314$. These are not folklore numbers: they are CI-mechanized in the repository's `proofs/economics/` directory — `delta-threshold.z3` proves existence, uniqueness in $(0,1)$, and location in $[0.34, 0.35]$ of the root by real-algebraic decision procedure; `claim_signaling.tla` model-checks the discounted one-shot-deviation invariant under both TLC and Apalache; and `sweep-delta.sh` confirms the empirical crossover ($\delta = 0.35$ on the integer grid) matches the closed form. Every claim in this paragraph is [verified] by that CI suite.

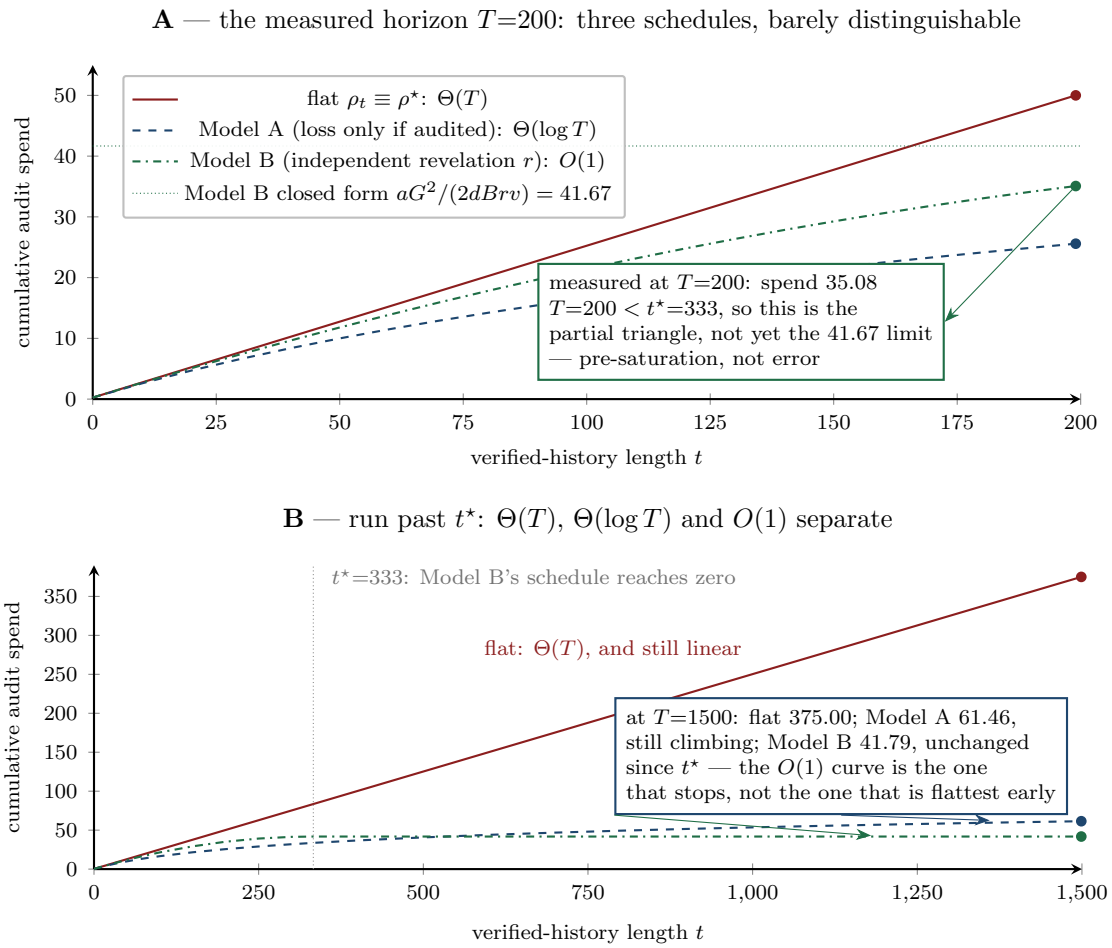


Figure 3: Cumulative audit spend against verified-history length t under the three incentive-compatible schedules of Theorem 3, at the running parameters. **A:** the horizon the paper measures, T=200. All three curves are near-linear here and Model B — the $O(1)$ schedule — is the *second-steepest*: on this range the asymptotics are not visible, and the measured 35.08 sits below the closed-form 41.67 because saturation requires $t^\star = G/(rv) = 333$. That gap is pre-asymptotic, not error. **B:** the same three schedules run to T=1500, past $t^\star$. Now the growth classes separate on the page: flat reaches 375.00 and keeps climbing linearly, Model A reaches 61.46 and is still climbing (logarithmically), and Model B stops dead at 41.79 from $t^\star$ onward and never spends again. Model A's cumulative spend crosses *above* Model B's at t=532: past that point the $O(1)$ schedule is permanently the cheaper of the two, which is the reversal panel A cannot show. The plateau 41.79 exceeds the continuous closed form 41.67 by 0.125 = $\rho^\star/2$, the expected discrete-sum overshoot on a linear ramp. Line styles differ (solid / dashed / dash-dot) so the three series survive greyscale printing [verified: `paper3_amortization_figures.py`, seed 20260816; cross-checks against `b2_tower.py`].

The connection to Theorem 3 is exact, not analogical. In Kreps–Wilson [3] and Fudenberg–Levine [4], deterrence is funded by *continuation value*: the discounted future surplus a deviator forfeits. In the inspection game it is funded by *bond plus stake*: $d(B + vt)$ at audit, rvt by revelation. Model A's schedule $\rho_t = G/(d(B + vt))$ is precisely the statement that as endogenous continuation value vt accumulates, exogenous verification ρ_t can retreat while the total deterrent stays pinned at G — the bonded game continuously deforms into the repeated game. The discount thresholds $\delta \geq 1/3$ and $\delta^\star = 0.3425$ are the limiting case $B = 0$, $\rho = 0$: all deterrence carried by the future. Reputation is amortized verification in both formalisms; the inspection game just itemizes the invoice.

8 Conjecture III.11.1, parameterized

The source volume's central open question — Conjecture III.11.1, the grading-oracle problem — asked whether a recursive rate-the-raters scheme can terminate. Two external reviews proposed answers that seemed to compete: one mandated a heterogeneous judge mix (distinct model families, to break correlated mode collapse) with VRF-selected honeypot tasks so a judge never knows when it is audited; the other derived the local condition $\rho dB > G$ and insisted a recursive tower needs a per-level contraction plus a grounded termination. Neither had the theorem. Fused, they *are* it:

- *Model heterogeneity supplies the cliques.* The disjointness that Theorem 2's C counts is statistical and economic independence of judge pools — exactly what distinct principals running distinct model families provide, and what a homogeneous pool (one vendor, one model, C=1) does not. The cost of C=1 is a linear bribery phase before the geometric one, roughly doubling certified depth and bond capital at the worked parameters (53 levels and 2650 against 27 and 1350); heterogeneity buys the removal of that phase, not the difference between a tower that converges and one that does not (§5).
- *VRF honeypots implement the sealed audit.* A verifiable random function [6] selecting honeypot tasks at rate $\rho^\star$ is precisely the sealed sampling the contraction hypothesis requires: the draw is unpredictable to the briber (sealing) yet publicly auditable after the fact (so the sampler itself cannot be quietly corrupted).
- *The termination is named, not wished.* The tower roots in an exogenously honest anchor — the operator at n=1, a bonded arbitration market at scale — and Theorem 2 prices the depth: 27 levels and bond capital 1350 at the running parameters [verified, arithmetic]. Concretely, this is what the judge ceremony of the reputation chapter is buying: the re-audit sampling rate is $\rho^\star$, the "distinct architectural weights" quorum is the C in CB, the VRF injection is the seal, and the depth-and-capital line item ($27 \times B$) is a number a platform can put in a budget rather than an assumption it has to hope holds.

The conjecture thereby becomes a *parameterized theorem* — deterrence at $\rho^\star = G/(dB)$ per level, contraction $(1 - \rho d)$ under sealed C-clique sampling with the bribery cutoff $G_k \leq CB$, amortized lifetime cost by Theorem 3 — plus a *measurement obligation*: does the deployed judge pool actually satisfy sealed multi-clique sampling, and what are d, r, and the cross-clique collusion correlation? Those are numbers to estimate from audit telemetry, not assumptions to wave at.

9 Reproducibility

Every number in this paper regenerates deterministically from one self-contained script, `skills/harbor-result` (numpy + matplotlib; program seed convention 20260816, though the computation is deterministic). Its output at the running parameters, verbatim:

```
rho* (keep-gain)     = G/(dB)       = 0.2500
rho* (confiscation)  = G/(d(G+B))   = 0.2083
q*   (mixed NE)      = a/(dL)       = 0.0312
judge indifferent at rho*:  cheat payoff = +0.000000
C=1:  400.0  390.0  380.0  370.0  360.0  350.0  340.0  330.0  320.0
C=8:  400.0  320.0  256.0  204.8  163.8  131.1  104.9   83.9   67.1
levels to drive corrupt value below 1 unit (geometric regime):  27;
total bond capital = B per level * levels = 1350
flat:       50.00   (Theta(T))
model A:    25.58   (Theta(log T); closed form 25.50)
model B:    35.08   (O(1); closed form aG^2/(2dBrv) = 41.67; hits 0 at t*=333)
IC check A: max deviation value = +1.78e-15 (<=0)
IC check B: max deviation value = +1.78e-15 (<=0)
```

The script also verifies the mixed equilibrium by direct best-response check (auditor indifference $-1.25 = -1.25$) and renders the three-panel summary figure. The claim-signaling thresholds of §7 regenerate independently via the CI suite in `proofs/economics/` (Z3, TLC, Apalache, and the δ-sweep).

10 Honest boundaries

What this paper does not say.

- *The headline is convention-dependent.* $\rho^\star = 0.25$ under keep-gain, 0.2083 under confiscation — a ∼17% swing from the settlement rule alone. Quote neither without stating which convention your slashing contract implements; a platform whose escrow claws back the gain but which budgets audits at $G/(dB)$ is over-auditing, and one that budgets at $G/(d(G+B))$ without clawback is under-deterring.
- *Sealed sampling is a hypothesis, not a garnish.* A briber who learns the audit draw before placing bribes faces threshold $G_k > B$, not $G_k > CB$: the clique math collapses to C=1 (§5). The sampler — and its VRF, and whoever operates it — is inside the trusted computing base of Theorem 2.
- *The tower needs an honest root.* Contraction shrinks corruption level over level; something exogenously honest must sit at the top — the operator in a small deployment, a bonded arbitration market at scale. The theorem prices the depth to that root; it does not conjure the root.
- *d, r, v, and cross-clique correlation are measured quantities, not assumptions.* Detection probability comes from audit outcomes; the revelation rate r from non-audit provenance of sanctions; cross-clique collusion correlation from judge-pair agreement statistics. "Disjoint cliques" is falsifiable and must be monitored: two nominally rival benches fine-tuned from the same base model may fail disjointness in exactly the correlated-error cases that matter.
- *The $\Theta(\log T)$-vs-$O(1)$ fork is empirical.* Model B's constant lifetime spend holds only if cheats surface without audits ($r > 0$). And at the session parameters the $O(1)$ claim is demonstrated pre-asymptotically: $t^\star = 333 > T = 200$, so the measured spend 35.08 sits below, not at, the 41.67 limit (§6).
- *One-shot gains, risk-neutral bribers, stationary parameters.* G is per-grade; an adversary who can correlate many small cheats below audit granularity, or who values capture asymmetrically, needs the repeated-game extension of §7 with the stake re-derived. The

machine-epsilon IC residuals ($+1.78 \times 10^{-15}$) are zeros of a float computation, stated as "≤ 0 up to float epsilon," not as exact algebra.

References

[1] R. Avenhaus, B. von Stengel, and S. Zamir. Inspection games. In *Handbook of Game Theory with Economic Applications*, Vol. 3, Ch. 51, pp. 1947–1987. Elsevier, 2002.

[2] G. S. Becker. Crime and punishment: an economic approach. *Journal of Political Economy* 76(2):169–217, 1968.

[3] D. M. Kreps and R. Wilson. Reputation and imperfect information. *Journal of Economic Theory* 27(2):253–279, 1982.

[4] D. Fudenberg and D. K. Levine. Reputation and equilibrium selection in games with a patient player. *Econometrica* 57(4):759–778, 1989.

[5] R. Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics* 14(4):436–440, 1943.

[6] S. Micali, M. Rabin, and S. Vadhan. Verifiable random functions. In *Proceedings of the 40th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 120–130. IEEE, 1999.