Verified Incentives in Hierarchical Bandits: Action-Contingent Transfers with Random Attestations

Liz Lemma Future Detective

January 16, 2026

Abstract

We study decentralized online contracting in hierarchical principal–agent systems when the key observability assumption—principals directly observing agents' actions—fails. Motivated by 2026-era agentic toolchains and API ecosystems, we model verifiable-but-private actions via receipts (e.g., signed logs, TEEs, zero-knowledge proofs) that attest an agent's action with probability p, and are otherwise absent. We extend the tree-structured principal–agent bandit framework with transfers and regret guarantees: contracts pay only upon receipt-confirmed compliance. We show a sharp verification–payment tradeoff: the minimal payment needed to induce a target action inflates by $1/p$, and payment-learning requires $1/p$ more samples to achieve the same confidence. Building on the regret decomposition (action/payment/deviation) and the transfer-based efficiency restoration logic in MAIL, we propose Verified-MAIL, a decentralized algorithm that (i) learns verification-adjusted inducing transfers and (ii) runs a bandit subroutine on shifted rewards. Verified-MAIL achieves sublinear welfare regret and no-regret utilities for all players under mild conditions, with rates that degrade polynomially in $1/p$. Finally, we provide information-theoretic lower bounds: without action verification ($p = 0$), outcome-only contracting can force linear welfare loss. Our results connect mechanism design and online learning to concrete infrastructure choices: how much attestation capability is needed to make local bonuses restore efficiency in modern AI supply chains.

Table of Contents

design; connections to secure attestations (TEE/ZK) at a modeling level.

3. 3. Model: tree delegation, stochastic rewards, receipt process, contract space, and information structure; define welfare and regret notions (utility regret vs welfare regret).

4. 4. Benchmark: full observability (p=1) recovers MAIL; restate optimal incentives as shadow prices and recall why transfers cancel in welfare.

5. 5. Verification-adjusted optimal incentives: characterize minimal inducing payments under receipts; closed-form τ^*/p scaling; discuss extensions (false positives/negatives; endogenous verification).

6. 6. Verified-MAIL algorithm: (i) receipt-subsampled payment exploration; (ii) shifted-reward bandit subroutine; (iii) design choices when receipts are missing; parameter schedules by depth.

7. 7. Theory: regret decomposition with receipts; bounds on deviation probability and payment estimation error; per-node regret and welfare regret bounds; explicit dependence on p and depth.

8. 8. Lower bounds / impossibility: outcome-only contracting (p=0) yields $\Omega(T)$ welfare regret on a constructed instance; optional lower bound showing τ must scale $\geq 1/p$ when $p>0$.

9. 9. Comparative statics and design implications: optimal p vs verification cost; where to deploy attestation in a hierarchy; policy/engineering takeaways for marketplaces.

10. 10. Extensions and discussion: small DAGs, budgets/limited liability, noisy verification, and empirical evaluation plan; conclusion.

1 Introduction and motivation

Modern economic activity is increasingly mediated by *agentic toolchains*: a user delegates a task to a high-level system, which decomposes it into subtasks and delegates further—to specialized APIs, third-party services, or downstream agents that themselves subcontract. The resulting organizational form is naturally hierarchical and recursive. At the same time, these toolchains operate in *API markets* where interactions are cheap, repeated, and data-rich, but also fragmented across administrative boundaries. A principal typically does not *observe* a subcontractor's internal action; it observes only outcomes (latency, a final answer, a transaction receipt) and, occasionally, some verifiable trace of what was done (a signed log entry, an attested execution report, a zero-knowledge proof of a computation). This combination of deep delegation and imperfect action observability motivates a contracting-and-learning problem that differs from the canonical principal–agent model in a simple but consequential way: *verification is stochastic and partial rather than deterministic and complete.*

We take seriously the practical premise that "auditability" is not binary. In many digital settings, verification arrives as a probabilistic event. A request may be accompanied by an authenticated trace only some of the time (because logging is sampled), an attestation may fail due to system constraints, or a proof may be too costly to generate for every interaction. Even when the principal can request evidence, evidence production can be subject to congestion and privacy constraints, and it is often most realistic to model verification as a random subsample of transactions. When verification is absent, the principal cannot condition transfers on the agent's action; when verification is present, the principal can enforce a contingent transfer in a way that is difficult to manipulate. This partial-verification regime is precisely where we expect the most acute tradeoffs between incentives and learning: the principal must pay enough to discipline behavior *in expectation*, yet cannot rely on frequent, high-quality feedback to calibrate those payments quickly.

Our goal is to illuminate this tradeoff in a setting with *nested delegation.* Each node is simultaneously an agent (to its parent) and a principal (to its children), and the organization must coordinate across layers while each participant learns from local feedback. This is the economic analogue of a multi-hop toolchain: downstream decisions create externalities upstream, but no single actor has global observability or centralized control. In such environments, one expects two frictions to interact: (i) *incentive misalignment* due to local objectives, and (ii) *statistical uncertainty* because payoffs must be learned online. A key modeling choice is therefore to represent not only stochastic rewards, but also stochastic *contract enforceability*: the principal can promise a transfer that is actually paid only when an unforgeable receipt arrives and confirms compliance.

This receipt-based view connects naturally to real implementation primitives. Trusted execution environments, secure enclaves, signed API responses, and hardware-backed logging all provide a notion of "receipt" that is difficult to forge when present, yet may be unavailable due to cost, latency, or system-level failures. Zero-knowledge proofs likewise offer strong correctness guarantees conditional on successful proof generation, but do not eliminate the resource tradeoff that makes universal verification impractical. Our abstraction deliberately suppresses engineering detail in favor of an economic object: a *verification probability* that indexes the strength of contract conditioning. This allows us to ask an economic question of first-order importance: when verification is sporadic, how much additional transfer must be promised to induce desired actions, and how does that requirement scale through a delegation hierarchy?

The first contribution is a transparent characterization of the *verification–payment frontier*. When compliance can be rewarded only on verified rounds, the expected incentive wedge created by a promised payment is attenuated by the verification probability. As a result, to replicate a given incentive constraint one must inflate payments by a factor that is inversely proportional to verification frequency. This yields a clean comparative static: weaker verifiability translates mechanically into higher promised payments (and, in equilibrium, higher expected burn), even holding fixed the underlying payoff externalities. The frontier is economically interpretable: verification and transfers are substitutes for achieving effort and coordination, and improving verification quality relaxes the need for monetary (or resource) incentives.

The second contribution is algorithmic: we show that decentralized learning can remain effective under partial verification, but only after modifying how incentives are explored and estimated. In nested delegation, a principal does not merely choose an action; it also chooses contracts for children, and must learn the minimal transfers that implement those contracts. Partial verification slows this learning because observed compliance is a Bernoulli subsample of true compliance events. The implication is that any procedure that searches for an inducing payment (for example, via batched tests or binary search) must run longer batches to obtain the same statistical confidence. Our Verified-MAIL construction formalizes this intuition by rescaling both payments and exploration schedules in a way that preserves the recursive structure of delegation.

The third contribution is a sharp boundary: without any verifiable action information, outcome-only contracting may be fundamentally insufficient for efficiency restoration. In many applied discussions, it is tempting to assume that enough outcome data will eventually reveal the right incentives. Our results caution against this optimism in hierarchical settings: if the principal cannot ever condition transfers on verified actions—even rarely—then there exist environments in which no decentralized learning-and-contracting scheme can avoid linear welfare loss. Economically, the obstruction is identification and enforceability: outcomes can be statistically compatible with multiple hidden action profiles that have different welfare implications, and without an action-contingent lever, incentives cannot be targeted to internalize externalities.

We summarize our main messages as follows.

- **Partial verification has a first-order price.** When only a fraction of interactions are verifiable, promised payments must increase to deliver the same expected incentive, generating a quantitative verification–payment frontier.

- **Learning slows predictably.** Because receipts subsample compliance observations, payment estimation and deviation control degrade with verification probability; algorithmic fixes require longer exploration and more conservative confidence.

- **Some verification is essential.** In the absence of any verifiable actions, there are instances where outcome-only schemes cannot coordinate the organization, implying a strict role for audit/attestation infrastructure.

Beyond theory, the model speaks to practical design choices. In platform governance and API ecosystems, one can often invest in better audit trails (higher receipt rates, lower false negatives), but this investment is costly. Our framework clarifies what such investments buy: they reduce the level of transfers required to align incentives and accelerate the learning of those transfers, especially in deep toolchains where errors propagate upstream. Conversely, if verification is expensive and remains sparse, one should expect either higher compensation budgets, more conservative delegation, or persistent inefficiencies.

We also acknowledge limitations that help delineate the scope of the conclusions. We treat receipts as unforgeable when present and independent across interactions; real systems may have correlated outages, strategic manipulation of evidence generation, or noisy attestations with false positives/negatives. We restrict attention to nonnegative, receipt-triggered transfers, abstracting from richer contract forms (e.g., penalties, dynamic reputation) that may partially substitute for verification. Finally, our focus is on local contracting in a fixed hierarchy; in practice, principals may rewire toolchains, switch providers, or centralize some decisions to economize on incentive costs. These extensions are natural next steps, but the baseline model already isolates the core tradeoff: *when verifiability is probabilistic, incentives become more expensive and learning becomes slower, yet even a small amount of verification can qualitatively change what is achievable.*

2 Related work

Our paper sits at the intersection of three literatures that are often studied separately: (i) principal–agent models with imperfect observability, (ii) online learning and bandits with strategic responses, and (iii) hierarchical (multi-tier) delegation with transfers. We also draw motivation from a growing body of systems work on secure attestations and verifiable computation, which provides concrete primitives that resemble the "receipt" abstraction we formalize.

Principal–agent theory and imperfect observability. The canonical principal–agent problem emphasizes that actions are typically hidden, so incentives must be provided through outcome-contingent compensation and dynamic considerations **??**. A complementary line studies *auditing* and *random inspection* as a way to partially restore action observability: even rare audits can discipline behavior if penalties/rewards conditional on audit outcomes are sufficiently strong **??**. Our model adopts a closely related logic—verification arrives only stochastically—but differs in two respects that matter for learning and for toolchain applications. First, we focus on *receipt-triggered transfers* that are paid only when verification arrives and confirms compliance, rather than on penalties imposed upon detected deviation. This matches many API-market settings in which clawbacks or fines are infeasible, while bonuses contingent on proofs or logs are feasible. Second, we embed this auditing friction in a *delegation hierarchy* in which each agent is simultaneously a principal downstream, so incentive costs and information distortions compound across layers.

Online contracting and principal–agent bandits. A growing literature studies how a principal can learn to contract when payoff-relevant parameters are unknown and must be inferred from repeated interaction **??**. In the bandit tradition, work on principal–agent learning emphasizes that the principal optimizes over contracts while facing stochastic rewards and strategic best responses by agents **??**. Related models include "bandits with strategic arms" where each arm is controlled by a self-interested agent who can manipulate outcomes or participation **??**. These papers clarify that learning incentives is qualitatively harder than learning payoffs: the principal must account for the agent's response mapping, which itself may be unknown or history-dependent.

Our contribution is aligned with this message but targets a different bottleneck: *verification-limited enforceability.* In many online contracting models, once the principal specifies a contract, the relevant performance signal is always observed (even if noisy). Here, by contrast, the principal may fail to observe the key compliance signal on most rounds, so the *same contract* produces different realized incentives across time. This has two implications:

(i) inducing payments scale as $1/p$ because the effective expected bonus is attenuated by the verification rate, and (ii) learning those payments is slowed because compliance is only observed on a Bernoulli subsample of rounds. In this sense, partial verification creates a distinct statistical friction on top of the usual exploration–exploitation and strategic-response issues.

Nested delegation with transfers and learning (MAIL). The closest conceptual ancestor is work on recursive contracting and learning in hierarchies, where local incentives are used to align behavior in a multi-layer organization. Classical economics emphasizes that delegation chains create wedges because each layer optimizes locally and passes distorted incentives downstream **??**. Recent algorithmic work brings this logic into repeated stochastic settings: under suitable contract spaces and no-regret learning assumptions, local incentive schemes can restore near-efficient behavior even without centralized observability. Our framework follows this recursive perspective and directly builds on the idea that a principal can reduce its problem to a bandit over *shifted* rewards once it internalizes the payments needed to induce desired downstream actions (as in MAIL-style recursion **?**).

Relative to this line, our main departure is to model action observability as *stochastic* even conditional on the contract: the parent can enforce payment only when a receipt arrives. This changes both the economics and the algorithmics. Economically, the relevant object is no longer the full-information minimal inducing transfer $\tau_b^\star(w)$, but its verification-adjusted counterpart $\tau_b^{\star,p}(w) = \tau_b^\star(w)/p$, which can be substantially larger when verification is rare. Algorithmically, procedures that estimate inducing transfers (e.g., batched tests that detect whether a child is deviating under a candidate payment) must be lengthened by approximately $1/p$ to maintain comparable confidence, because only receipt rounds are informative about compliance. The Verified-MAIL construction preserves the recursive tractability of the original approach while making this verification dependence explicit.

Auditing, verification, and mechanism design. Our receipt model also connects to broader work on verification in mechanism design and information economics. The "costly state verification" tradition studies when principals optimally verify reported information and how contracts trade off verification costs against incentive provision **??**. Other work analyzes audit probabilities and enforcement in regulatory settings, showing that stochastic audits can implement compliance when penalties are credible and sufficiently steep **?**. We can be viewed as importing a similar verification lever into an online, decentralized environment, with two twists: verification is not chosen endogenously in the baseline model (though it can be in extensions), and enforcement uses nonnegative, receipt-triggered transfers rather than punishments. These restrictions are not innocuous: they rule out some classical implementations that rely on large fines, and they sharpen the role of even small verification rates in enabling identification and incentive targeting.

Secure attestations as a modeling primitive. Finally, the receipt abstraction is motivated by concrete cryptographic and systems primitives that increasingly mediate real-world contracting in digital ecosystems. Trusted execution environments (TEEs) can provide signed attestations about code execution; secure logging and provenance systems can produce authenticated traces; and zero-knowledge proofs (ZK) can certify that a computation was performed correctly without revealing sensitive inputs **???**. In practice, however, these mechanisms are often *intermittent*: attestations may be sampled, proofs may be generated only for high-value requests, and system outages or latency constraints may prevent universal verification. Our parameter p is intended to capture this operational reality at a reduced-form level. The resulting comparative statics—verification as a substitute for monetary burn, and partial verification as a drag on learning speed—translate directly into design guidance: increasing attestation coverage or reliability reduces the transfers required to align incentives and improves the rate at which those transfers can be calibrated.

Taken together, these literatures suggest a common lesson: incentives and information are complements, and weakening one forces the other to work harder. The novelty in our setting is that the information friction is not merely noisy outcomes but *missing enforceability events*, and the organizational form is explicitly hierarchical. This combination produces a simple but powerful frontier—payments inflate as $1/p$, learning slows by $1/p$, and when $p = 0$ there are environments where efficiency restoration is impossible—that we formalize in the model section that follows.

3 Model

We study repeated delegation on a rooted tree $\mathcal{T} = (V, E)$. Each node $v \in V$ is simultaneously an *agent* (relative to its parent $P(v)$, when $v \neq \text{root}$) and a *principal* (relative to its children $C(v)$). Time is discrete $t = 1, \ldots, T$. In each round, every player chooses an action $A_t^v \in \mathcal{A}$, where $|\mathcal{A}| = K$. We index depth so that leaves are at depth 1 and the root is at depth D. For simplicity, one may think of a regular branching factor B, although nothing in the model definition requires regularity.

Local payoff externalities and stochastic rewards. The payoff of node v depends on its own action and the actions of its direct children:

$$X_t^v\big(A_t^v, A_t^{C(v)}\big) = \theta_v\big(A_t^v, A_t^{C(v)}\big) + z_t^v, \qquad \theta_v(\cdot) \in [0,1]. \tag{1}$$

The noise z_t^v is conditionally zero-mean sub-Gaussian, capturing the standard bandit-type uncertainty that persists even if the action profile were observed. The restriction to dependence on $C(v)$ encodes *local* externalities along edges; it is precisely this locality that makes recursive contracting feasible, but it does not eliminate strategic tension because each child action can shift the parent's reward.

Local contracts: recommendations and receipt-triggered transfers. In each round t, each principal v issues a separate contract to each child $w \in C(v)$. The contract consists of a recommended action $B_t(w) \in \mathcal{A}$ and a nonnegative transfer $\tau_t(w) \geq 0$. Transfers are *receipt-triggered*: v pays $\tau_t(w)$ to w only when a verifiable receipt arrives and attests that w complied with the recommendation.

Formally, on each edge (v, w) and round t, an attestation signal (receipt) $S_t^{v \to w} \in \mathcal{A} \cup \{\perp\}$ is realized according to

$$\Pr[S_t^{v \to w} = A_t^w] = p, \qquad \Pr[S_t^{v \to w} = \perp] = 1 - p, \tag{2}$$

independently across edges and time. The symbol $\perp$ denotes "no receipt" (unverifiable). Unforgeability is built in: whenever $S_t^{v \to w} \neq \perp$, it equals the child's true action. Payments are then settled by the rule

$$\text{payment from } v \text{ to } w \text{ at time } t = \mathbb{I}\{S_t^{v \to w} = B_t(w)\}\ \tau_t(w). \tag{3}$$

This contract space is intentionally one-sided (bonuses only) and local (edge-by-edge). The nonnegativity restriction reflects limited liability or the practical difficulty of enforcing fines/clawbacks in many digital contracting environments.

Timing and information. Within each round t: (i) each v observes the contract offered by its parent $(B_t(v), \tau_t(v))$ (for the root this term is absent); (ii) each v chooses A_t^v and simultaneously chooses $\{(B_t(w), \tau_t(w))\}_{w \in C(v)}$ for its children; (iii) rewards $\{X_t^v\}_{v \in V}$ realize; (iv) receipts $\{S_t^{v \to w}\}_{(v,w) \in E}$ realize; (v) transfers are paid according to receipts; (vi) players update their learning states.

The information structure is decentralized. Player v observes its own realized reward X_t^v, its incoming receipt $S_t^{P(v) \to v}$ (if applicable), and its outgoing receipts $\{S_t^{v \to w}\}_{w \in C(v)}$. When $S_t^{v \to w} = \perp$, the parent receives no direct evidence of w's action in that round. Thus, even holding fixed a contract $(B_t(w), \tau_t(w))$, the realized incentive provided to the child is random over time because the transfer is only paid on a random subset of rounds.

Utility with transfers. Each node's per-round utility equals its stochastic reward plus any incoming transfer received upon verified compliance, minus any outgoing transfers paid to children upon verified compliance:

$$U_t^v = X_t^v\big(A_t^v, A_t^{C(v)}\big) + \mathbb{I}\Big\{S_t^{P(v)\to v} = B_t(v)\Big\}\ \tau_t(v) - \sum_{w\in C(v)} \mathbb{I}\{S_t^{v\to w} = B_t(w)\}\ \tau_t(w), \tag{4}$$

with the convention that the incoming term is 0 for the root. The transfer terms encode the core enforcement friction: a child can deviate without immediate financial consequence whenever no receipt arrives, and the parent can economize on payments only to the extent that verification is frequent.

Continuation values and induced utilities. Because each node is both agent and principal, a useful summary object is the *induced expected utility* of a child under downstream optimal play. Fix a node w and imagine that its parent recommends an action $b \in \mathcal{A}$ with transfer τ. Let $\mu_w(a)$ denote w's expected continuation value if it chooses action a and then optimally contracts with its own descendants given the same receipt structure. This recursion is what allows "local" contracts to be analyzed in terms of effective payoffs. Under a stationary contract recommending b, the incentive-compatibility constraints for w compare the expected utility of choosing b to any deviation $a \neq b$:

$$\mu_w(b) + p\,\tau \ \geq\ \mu_w(a), \qquad \forall a \in \mathcal{A}, \tag{5}$$

since the receipt-triggered transfer is obtained only with probability p. We emphasize that (5) is an *ex post* behavioral condition (given a continuation value function), and it is the key place where verification enters multiplicatively.

Welfare. We evaluate system performance by welfare defined over *real rewards*, not transfers. The realized welfare in round t is

$$W_t \ = \ \sum_{v\in V} X_t^v\big(A_t^v, A_t^{C(v)}\big), \tag{6}$$

with corresponding expected welfare $\sum_v \theta_v(\cdot)$. This choice reflects the standard perspective that transfers are internal to the organization/marketplace: they redistribute utility but do not create surplus. Indeed, summing (4) across all $v \in V$ cancels every transfer term edge-by-edge, leaving $\sum_v U_t^v = \sum_v X_t^v$. This accounting identity underlies why we can simultaneously study incentive costs (which matter for individual utilities and feasibility) and efficiency (which depends only on realized rewards).

Regret notions: individual utility regret and welfare regret. Two regret metrics will be useful. First, for each node v, we define an *individual (utility) regret* relative to the best fixed action in hindsight against the realized environment it faced:

$$R_v(T) = \max_{a \in \mathcal{A}} \mathbb{E}\left[\sum_{t=1}^{T} U_t^v(a)\right] - \mathbb{E}\left[\sum_{t=1}^{T} U_t^v\right], \tag{7}$$

where $U_t^v(a)$ denotes the (counterfactual) utility at time t if v had played a while everything else (including contracts offered by others, receipt realizations, and downstream learning dynamics) evolves as specified.[1] This definition captures the learning objective we impose on each agent: maximize its own realized utility stream given partial verification and stochastic rewards.

Second, we define *welfare regret* against the best fixed joint action profile in hindsight under the mean rewards:

$$R_{\mathrm{wel}}(T) = \mathbb{E}\left[\sum_{t=1}^{T}\left(\max_{a \in \mathcal{A}^{|V|}} \sum_{v \in V} \theta_v\big(a_v, a_{C(v)}\big) - \sum_{v \in V} \theta_v\big(A_t^v, A_t^{C(v)}\big)\right)\right]. \tag{8}$$

Welfare regret is the metric that speaks most directly to efficiency restoration: it asks whether the decentralized contracts and learning dynamics drive play toward the welfare-optimal profile, despite missing receipts and strategic incentives. The main results in the sequel establish that $p > 0$ is enough to make $R_{\mathrm{wel}}(T) = o(T)$ achievable (with explicit dependence on $1/p$ and depth), while $p = 0$ admits environments where linear welfare regret is unavoidable.

4 Benchmark: full observability ($p = 1$)

To build intuition for the role of receipts, it is helpful to begin from the limiting case of *full observability*, in which every recommended action can be verified ex post. In our notation, setting $p = 1$ implies that on each edge (v, w) and round t we have $S_t^{v \to w} = A_t^w$ almost surely, so a parent can condition payments on the child's realized action without statistical thinning. This benchmark is not merely a technical convenience: it isolates the *strategic* reason transfers are needed (local payoff externalities along edges) from the *informational* reason they become expensive (missing receipts). In this regime, our model reduces to the setting studied by MAIL, and the recursive contract structure admits a particularly transparent "shadow price" interpretation.

Fix an edge (v, w) and suppose v wants to induce w to take some target action $b \in \mathcal{A}$. Under full observability, the receipt-triggered bonus is paid

[1] This is the standard "policy regret" subtlety in strategic environments; our results focus on settings where the recursive contract structure makes the induced environment sufficiently stable to admit no-regret guarantees, as in MAIL-style analyses.

whenever w complies, which now happens deterministically conditional on w's choice. Thus the incentive constraints (5) reduce to

$$\mu_w(b) + \tau \;\geq\; \mu_w(a), \qquad \forall a \in \mathcal{A}, \tag{9}$$

and the minimal bonus that makes b optimal (breaking ties in favor of compliance) is the familiar "gap" between the best deviation and the recommended action:

$$\tau_b^\star(w) \;=\; \max_{a\in\mathcal{A}} \mu_w(a) - \mu_w(b). \tag{10}$$

Equation (10) is the first place where the recursion across the tree becomes operational: $\mu_w(\cdot)$ already incorporates the fact that w will optimally contract with its descendants, so $\tau_b^\star(w)$ is the *net* inducement cost after downstream incentives are optimally set. In other words, even though v writes a contract only with w, the object μ_w internalizes all lower-level strategic interactions into a single continuation value function.

From the parent's perspective, inducing w to play b yields a benefit through v's own payoff function $\theta_v(\cdot)$, but it also requires paying the bonus $\tau_b^\star(w)$. Under compliance, we can therefore summarize v's induced expected utility (net of incentive costs to children) by the "shifted" payoff

$$\mu_v\big(a, b_{C(v)}\big) \;=\; \theta_v\big(a, b_{C(v)}\big) - \sum_{w\in C(v)} \tau_{b_w}^\star(w), \tag{11}$$

which is exactly the MAIL recursion specialized to $p = 1$. This representation is conceptually important: it converts a hierarchical contracting problem into a local choice problem in which the principal selects recommended child actions $b_{C(v)}$ as if facing a standard bandit objective, but with payoffs reduced by the *endogenous prices* $\tau_{b_w}^\star(w)$ required to sustain those recommendations.

These transfers admit a natural interpretation as *shadow prices* of incentive constraints. To see this, imagine a local planner (or equivalently a principal writing a mechanism) who would like to impose the constraint "child w plays b" but must respect w's best response. The constraint (9) says that the mechanism must compensate the child for the opportunity cost of foregoing its best alternative action. The minimal transfer $\tau_b^\star(w)$ is precisely the smallest Lagrange multiplier that relaxes the binding deviation constraint, and thus can be read as the (endogenous) unit price of implementing action b at node w. This "price" is not exogenous: it depends on downstream contractibility and learning through $\mu_w(\cdot)$. When a descendant is easy to incentivize (because its downstream continuation values align), $\mu_w(b)$ is large relative to $\max_a \mu_w(a)$, and the corresponding shadow price is small; when incentives are misaligned, the price rises.

This shadow-price view also clarifies why, in the welfare analysis, transfers are simultaneously indispensable and irrelevant. They are indispensable because without them the recommended actions need not be best responses, so local externalities would not be internalized. Yet transfers are irrelevant for welfare because they are pure redistribution within the tree. Formally, summing per-round utilities (4) over $v \in V$ cancels every transfer on every edge: each outgoing payment from a parent appears as an incoming payment to the child with the opposite sign. Consequently,

$$\sum_{v \in V} U_t^v \; = \; \sum_{v \in V} X_t^v\big(A_t^v, A_t^{C(v)}\big) \; = \; W_t, \tag{12}$$

independently of the contract terms and independently of whether $p = 1$ or $p < 1$. This identity is the accounting backbone of the MAIL-style welfare guarantee: once we can ensure that each node experiences vanishing *utility* regret relative to its induced environment, the sum of utilities tracks the sum of realized rewards, and efficiency statements can be made without separately tracking the flow of transfers.

At the same time, we emphasize a limitation that becomes central once we depart from the benchmark: although transfers cancel in *welfare*, they do not cancel in any agent's *individual* utility, nor are they innocuous from the perspective of feasibility. Large τ values can create significant "payment burn" for principals and large variance in realized utilities for agents, even when the system-wide surplus is unchanged. Under full observability, (10) shows that the required payments are exactly the minimal wedges needed for incentive compatibility, so payment burn is as low as it can be given the continuation values. This is one reason full monitoring is often treated as the idealized target in practice: strong auditing, reliable logging, or cryptographic attestations effectively push systems toward the $p = 1$ regime, reducing the incentive cost of enforcing compliance.

Finally, this benchmark clarifies what changes when we move to partial verification. When $p < 1$, the same shadow-price logic still applies, but the "unit price" of implementing an action is inflated because the bonus is received only on the subset of rounds where a receipt arrives. Put differently, full observability is the case in which the mechanism designer can pay exactly when an action is taken; partial observability forces the designer to pay only when an action is *verified*. The next section formalizes how this statistical thinning transforms (10) into a $1/p$ scaling, and how the learning dynamics must be modified accordingly.

5 5. Verification-adjusted optimal incentives: characterize minimal inducing payments under receipts; closed-form τ*/p scaling; discuss extensions (false positives/negatives; endogenous verification).

When receipts arrive only intermittently, the parent can still use the same local "pay-for-compliance" logic as in the benchmark, but the incentive constraint is diluted by statistical thinning. Fix an edge (v, w) and a target recommendation $b \in \mathcal{A}$. Consider a stationary contract of the form: the parent recommends b and promises a nonnegative bonus $\tau \geq 0$ that is paid if and only if a receipt arrives and attests that w played b, i.e., if $S_t^{v \to w} = b$. Under our receipt technology, if w actually plays action a, then $S_t^{v \to w} = a$ with probability p and $S_t^{v \to w} = \perp$ otherwise; moreover, when $S_t^{v \to w} \neq \perp$ it is unforgeable. Hence the expected transfer received by w equals $p\tau$ if $a = b$, and 0 if $a \neq b$. In terms of continuation values, w's expected utility from choosing a is therefore

$$\mu_w(a) \; + \; p\,\tau\,\mathbb{I}\{a = b\},$$

which yields the receipt-adjusted incentive constraints

$$\mu_w(b) + p\,\tau \; \geq \; \mu_w(a), \qquad \forall a \in \mathcal{A}. \tag{13}$$

Relative to the full-observability case, the only change is that the "effective" bonus is $p\tau$ rather than τ: the payment is perfectly targeted when it occurs, but it occurs only on a p-fraction of rounds.

Solving (13) immediately gives a closed form for the smallest payment that induces b (again breaking ties in favor of compliance):

$$\tau_b^{\star,p}(w) \; = \; \frac{\max_{a \in \mathcal{A}} \mu_w(a) - \mu_w(b)}{p} \; \equiv \; \frac{\tau_b^{\star}(w)}{p}, \qquad p > 0. \tag{14}$$

Two observations are worth isolating. First, if b is already optimal for w given its downstream environment (i.e., $\mu_w(b) = \max_a \mu_w(a)$), then $\tau_b^{\star,p}(w) = 0$ for every p: receipts are irrelevant when incentives are already aligned. Second, whenever there is a strictly positive temptation to deviate, partial verification inflates the required *promised* transfer by a factor $1/p$. This scaling is not an artifact of our proof technique; it is the direct implication of paying only upon verification. The expected reward wedge needed for incentive compatibility is a property of preferences (the gap $\tau_b^{\star}(w)$), while the instrument delivers only $p\tau$ units of expected wedge per unit of promised payment.

The same logic also clarifies why we insist on $p > 0$ for general implementability. If $p = 0$, then no contract of this form can ever condition on compliance, so (13) collapses to $\mu_w(b) \geq \mu_w(a)$ for all a, i.e., the principal can induce b only if b is already a best response. In other words, when receipts never arrive, transfers cannot relax any deviation constraint at all; incentives are purely "outcome-driven" through μ_w, which in general does not align with the parent's objective. This is precisely the channel behind our later impossibility: without any verification, the contractible margin disappears.

Equation (14) also sharpens a practical distinction between welfare and feasibility. Even though transfers cancel in aggregate welfare, the *gross* promised payments can become large when p is small: implementing the same action profile requires higher posted bonuses, and realized transfers become more sporadic but larger when they occur. Thus partial verification can increase principals' payment burn, tighten liquidity constraints, and inflate variance in realized utilities even if the induced action profile (and therefore welfare) remains essentially the same. In applications, this trade-off is often salient: weak auditability forces organizations either to tolerate misaligned behavior or to post very high contingent bonuses that are rarely triggered but expensive when triggered.

Our baseline receipt model assumes perfect correctness conditional on being present (unforgeable, no misclassification). It is useful, however, to understand how (14) changes with noisy attestations. One natural extension introduces false negatives and false positives. Suppose that when w plays b, the receipt correctly attests b with probability $p(1 - \varepsilon_{\text{fn}})$ (otherwise it is missing or uninformative), and when w plays $a \neq b$, the receipt incorrectly attests b with probability $p\,\varepsilon_{\text{fp}}$. Under the same "pay iff the receipt says b" rule, the expected transfer from playing b is $p(1 - \varepsilon_{\text{fn}})\tau$, while the expected transfer from deviating is $p\varepsilon_{\text{fp}}\tau$. The net incentive wedge delivered by the instrument is therefore $p(1 - \varepsilon_{\text{fn}} - \varepsilon_{\text{fp}})\tau$, and the IC condition becomes

$$\mu_w(b) + p(1 - \varepsilon_{\text{fn}})\tau \;\geq\; \mu_w(a) + p\varepsilon_{\text{fp}}\tau, \qquad \forall a \neq b,$$

so that a sufficient (and essentially tight) minimal transfer is

$$\tau_b^{\star,p,\varepsilon}(w) \;=\; \frac{\max_{a\in\mathcal{A}} \mu_w(a) - \mu_w(b)}{p\,(1 - \varepsilon_{\text{fn}} - \varepsilon_{\text{fp}})}, \tag{15}$$

provided $1 - \varepsilon_{\text{fn}} - \varepsilon_{\text{fp}} > 0$. This formula highlights a sharp qualitative boundary: false negatives act like a reduction in effective verification frequency, while false positives undermine the targeting of payments and can make incentives non-implementable if misclassification is too severe (when $\varepsilon_{\text{fn}} + \varepsilon_{\text{fp}} \geq 1$, raising τ no longer increases the expected advantage of compliance over deviation). Because our main focus is on learning under statistical thinning rather than misclassification, we keep the unforgeable-receipt assumption throughout, but (15) indicates how the guarantees would deform under more realistic auditing errors.

A second extension makes verification endogenous. In many systems, a principal can invest in higher p via audits, monitoring infrastructure, or cryptographic attestation, at some cost $c(p)$ (e.g., increasing and convex). Under (14), the expected *promised* inducement cost for action b is proportional to $\tau_b^\star(w)/p$, suggesting a simple reduced-form tradeoff: verification substitutes for transfers. If a parent chooses p and τ jointly to implement b, a natural objective is to minimize total implementation cost $c(p)+\tau$ subject to $p\tau \geq \tau_b^\star(w)$, which yields $\tau = \tau_b^\star(w)/p$ and hence the scalar problem

$$\min_{p\in(0,1]} c(p) + \frac{\tau_b^\star(w)}{p}.$$

When c is differentiable and strictly convex, an interior optimum satisfies $c'(p) = \tau_b^\star(w)/p^2$: the marginal cost of improving verification is set equal to the marginal savings in incentive payments. This perspective connects our model to practice: better logging, stronger attestations, and more frequent audits reduce the financial (or contractual) burden needed to align behavior, but may themselves consume resources.

Finally, we emphasize the methodological implication for learning. The object $\tau_b^{\star,p}(w)$ depends on $\mu_w(\cdot)$, which is endogenous and unknown ex ante because it incorporates downstream learning and contracting. Under partial verification, the parent additionally observes compliance only on receipt rounds. Thus even estimating which payments induce which actions must proceed from a subsampled signal. The next section operationalizes this point: Verified-MAIL combines (i) an exploration procedure that lengthens payment-search batches by a factor $1/p$ to compensate for missing receipts with (ii) the same shifted-reward bandit reduction as in MAIL, thereby restoring no-regret guarantees despite verification frictions.

Verified-MAIL: overview. Verified-MAIL is the receipt-robust analogue of MAIL: each node v simultaneously (a) *learns* which local action/recommendation vector is optimal for its upstream objective and (b) *implements* that recommendation vector downstream using receipt-triggered transfers. The receipt friction affects both components. On the implementation side, the principal cannot directly observe compliance in most rounds, so it must (i) post higher promised bonuses (as quantified above) and (ii) *learn* those bonuses from a thinned stream of verifications. On the learning side, even when the induced action profile is stable, realized utilities contain additional receipt-driven randomness; we therefore separate (A) a payment-learning routine that is explicitly designed for subsampled compliance feedback and (B) a bandit routine that treats the induced payment burden as a known offset (up to estimation error). We run these two routines on a carefully separated time scale, with depth-dependent batch lengths to control the propagation of estimation error up the tree.

(i) Receipt-subsampled payment exploration (ExpSub). Fix a node v, a child $w \in C(v)$, and a target recommendation $b \in \mathcal{A}$. The object we need for downstream implementation is the receipt-adjusted inducing transfer $\tau_b^{\star,p}(w)$. The difficulty is that v only observes w's action when $S_t^{v\to w} \neq \perp$, which occurs on a p-fraction of rounds in expectation; moreover, w is itself learning, so short-run play can deviate from its eventual best response even under a correctly calibrated contract. ExpSub addresses both issues via long batches and one-sided (upper) estimation.

Operationally, ExpSub maintains an interval $[\underline{\tau}, \overline{\tau}] \subseteq [0, \tau_{\max}]$ that is guaranteed (on a high-probability good event) to contain $\tau_b^{\star,p}(w)$, and repeatedly queries the midpoint $\tau = (\underline{\tau} + \overline{\tau})/2$ for a batch of rounds. During a query batch, v posts the stationary contract $(B_t(w), \tau_t(w)) = (b, \tau)$ and records only the receipt rounds. Let

$$N \;=\; \sum_{t\in\text{batch}} \mathbb{I}\{S_t^{v\to w} \neq \perp\}, \qquad Y \;=\; \sum_{t\in\text{batch}} \mathbb{I}\{S_t^{v\to w} = b\},$$

so that Y/N is an empirical compliance rate *conditional on verification.* Because missing receipts are neither evidence of compliance nor deviation, we simply treat them as censored observations and exclude them from N. The decision rule is intentionally conservative: if Y/N is sufficiently high (above a tolerance level $1-\eta$), we declare τ *feasible* and set $\overline{\tau} \leftarrow \tau$; otherwise we set $\underline{\tau} \leftarrow \tau$. The tolerance η absorbs both statistical error from finite N and the child's transient deviations due to its own no-regret learning. In particular, because N itself concentrates around $p \cdot$ (batch length), we choose batch lengths inflated by a factor $1/p$ so that, with high probability, N is large enough for Hoeffding-style control of Y/N.

We run this binary search on a geometric accuracy schedule. Writing $\varepsilon_m = 2^{-m}$ for the target precision in phase m, we use batch length of order

$$L_m \;=\; \Theta\!\left(\frac{1}{p} \cdot \frac{1}{\varepsilon_m^2} \log \frac{1}{\delta_m}\right),$$

so that the number of receipt observations N is $\tilde{\Theta}(\varepsilon_m^{-2} \log(1/\delta_m))$ and the acceptance/rejection decision is reliable. The output after m phases is an *upper* estimate $\hat{\tau}_b(w)$ satisfying $\hat{\tau}_b(w) \geq \tau_b^{\star,p}(w)$ with probability at least $1 - \sum_{j \leq m} \delta_j$. This one-sidedness is deliberate: underestimating payments risks sustained deviations that contaminate the principal's reward samples, whereas overestimating payments only incurs additional transfer burn and slows regret rates polynomially.

(ii) Shifted-reward bandit subroutine at each node. Payment exploration produces, for each child w and each recommendation b, a current inducing-payment proxy $\hat{\tau}_b(w)$. Conditional on these proxies, node v faces a standard bandit problem over *meta-actions*

$$\alpha \;=\; (a, b_{C(v)}) \in \mathcal{A}^{1+|C(v)|},$$

where a is v's own action and $b_{C(v)}$ is the vector of recommendations to its children. The conceptual objective is to pick α that maximizes the downstream-induced value $\theta_v(a, b_{C(v)})$ net of the payment burden needed to make the recommendations self-enforcing. Accordingly, in exploitation rounds we feed the bandit algorithm a *shifted* sample

$$\widetilde{X}_t^v(\alpha) \;=\; X_t^v\big(a, A_t^{C(v)}\big) \;-\; \sum_{w \in C(v)} \hat{\tau}_{b_w}(w),$$

where $A_t^{C(v)}$ are the realized child actions (which, on the high-probability event of successful inducement, coincide with $b_{C(v)}$ up to a vanishing fraction of rounds). This shift removes from the learning objective the need to track the stochastic realization of transfers (which are instrument noise from the perspective of identifying θ_v) and instead internalizes the *implementation cost* through the deterministic offsets $\hat{\tau}_{b_w}(w)$.

Because $\hat{\tau}$ can be as large as $O(1/p)$, $\widetilde{X}_t^v$ need not lie in $[0, 1]$. We therefore normalize the shifted reward by a known range bound R_v (e.g., $R_v = 1 + |C(v)|\tau_{\max}$) and run any adversarial-robust no-regret bandit method (such as EXP3) on the rescaled samples; equivalently, we may run UCB-style algorithms with appropriately scaled confidence radii. The key design choice is that the bandit learner treats $\hat{\tau}$ as fixed within an epoch, so that within-epoch rewards are conditionally i.i.d. up to sub-Gaussian noise and rare deviation events.

(iii) What to do when receipts are missing. Two implementation details matter in practice and in the subsequent regret analysis. First, transfers are *never* paid on $S_t^{v \to w} = \perp$. This is not merely a modeling constraint but a robustness feature: when verification is absent, the principal cannot condition payments on unverifiable claims, so any attempt to "fill in" missing receipts would reopen moral-hazard channels. Second, missing receipts produce missing compliance observations, not negative evidence. Consequently, ExpSub updates only on receipt rounds, and the principal's bandit learner does not interpret $S = \perp$ as a deviation signal. Instead, deviations are controlled indirectly by (a) maintaining upper estimates $\hat{\tau}$ and (b) choosing batch lengths large enough that, even after thinning, the verified rounds suffice to detect systematic noncompliance with high probability. In effect, we treat receipts as a Bernoulli subsampling device: they slow learning, but they do not bias it.

Depth-dependent parameter schedules. Finally, we set parameters by depth to manage error propagation. Let $d(v)$ denote the distance from v

to the leaves. As in MAIL, nodes closer to the leaves stabilize faster because they face fewer layers of downstream strategic learning. Verified-MAIL mirrors this by allocating shorter epochs and tighter confidence budgets to shallow nodes, and longer, more conservative epochs to deeper nodes. Concretely, we choose a decreasing accuracy schedule $\varepsilon_m(d)$ and a summable failure budget $\delta_m(d)$ (e.g., $\delta_m(d) \propto (TK)^{-cd}2^{-m}$) and set both (a) Exp-Sub batch lengths and (b) bandit-epoch lengths to scale like p^{-1} times the corresponding MAIL lengths. This ensures that every layer receives, in expectation, the same number of *verified* observations per phase as under full verification, up to logarithmic factors, while keeping the overall horizon feasible.

These choices yield an algorithm that is modular (payment estimation plus bandit selection) and explicitly receipt-aware. The next section uses this modularity to decompose regret into (i) bandit learning regret under shifted rewards, (ii) payment-estimation error, and (iii) losses from the residual probability of undetected deviations, with each term exhibiting a transparent dependence on p and depth.

Regret decomposition with receipt noise. We analyze nodewise regret with respect to the *implementation-adjusted* objective, i.e., the value of an action/recommendation vector net of the *minimal* receipt-adjusted inducement costs. For a node v, let a meta-action be $\alpha = (a, b_{C(v)}) \in \mathcal{A}^{1+|C(v)|}$, and define the corresponding benchmark mean

$$\mu_v(\alpha) \triangleq \theta_v\big(a, b_{C(v)}\big) - \sum_{w \in C(v)} \tau^{\star,p}_{b_w}(w),$$

where $\tau^{\star,p}_b(w)$ is the minimal promised transfer that makes b optimal for w under receipt probability p. We measure regret against $\max_\alpha \mu_v(\alpha)$, which is the natural analogue of MAIL's recursion: it compares v's realized performance to the best locally implementable recommendation profile, accounting for the enforcement burden created by receipts.

The key technical change relative to full verification is that v's realized utility contains two additional sources of randomness: (i) *receipt-triggered transfers* paid/received with Bernoulli probability p, and (ii) *strategic deviations* that are only partially observed. Our analysis therefore conditions on a high-probability event $\mathcal{E}$ under which (a) all payment estimates remain one-sided ($\hat{\tau}_b(w) \geq \tau^{\star,p}_b(w)$ for all queried (w, b)), (b) receipt counts concentrate in every exploration batch, and (c) every child's no-regret guarantee holds uniformly over epochs. Under $\mathcal{E}$, we can write a clean regret decomposition:

$$R_v(T) \leq \underbrace{R_v^{\text{bandit}}(T)}_{\text{learning } \alpha} + \underbrace{R_v^{\text{pay}}(T)}_{\text{overpay / estimation error}} + \underbrace{R_v^{\text{dev}}(T)}_{\text{residual deviations}},$$

where each term admits a receipt-explicit bound.

Payment estimation error and its $1/p$ cost. The ExpSub routine produces upper estimates $\hat{\tau}_b(w)$ from censored compliance data: only receipt rounds contribute to the empirical compliance statistic. For a batch of length L, the number of usable samples is $N = \sum_{t \in \text{batch}} \mathbb{I}\{S_t^{v \to w} \neq \perp\}$, with $\mathbb{E}[N] = pL$. Concentration of N implies that with probability at least $1 - \delta$,

$$N \;\geq\; \tfrac{p}{2} L \qquad \text{whenever} \qquad L \;\gtrsim\; \frac{1}{p} \log \frac{1}{\delta}.$$

Thus, to achieve the same statistical precision as under $p = 1$, every binary-search query must be lengthened by a factor $\Theta(1/p)$. This is the first, unavoidable appearance of $1/p$ in the regret rate.

On $\mathcal{E}$, the one-sided property implies we never *under*-incentivize in exploitation epochs; the cost of estimation is therefore purely *overpayment*. For a fixed epoch in which v recommends b_w to child w, we incur an additive mean offset

$$\hat{\tau}_{b_w}(w) - \tau^{\star,p}_{b_w}(w) \;\geq\; 0$$

in the shifted objective. Summing across children and time,

$$R_v^{\text{pay}}(T) \;\leq\; \sum_{t=1}^{T} \sum_{w \in C(v)} \Big(\hat{\tau}_{b_{w,t}}(w) - \tau^{\star,p}_{b_{w,t}}(w) \Big),$$

so any schedule that drives $\hat{\tau} - \tau^{\star,p}$ to zero sufficiently fast yields sublinear payment regret. With a geometric accuracy schedule ε_m and batch lengths $L_m = \tilde{\Theta}((1/p)\varepsilon_m^{-2})$, the cumulative overpayment is dominated by the final precision achieved before time T, giving a typical bound of the form

$$R_v^{\text{pay}}(T) \;=\; \tilde{O}\Big(|C(v)| \, T \, \varepsilon_{\text{final}} \Big) \;+\; \tilde{O}\Big(\frac{|C(v)|}{p} \sum_m \varepsilon_m^{-2} \Big),$$

where the second term reflects the exploration time spent to learn payments from a p-thinned stream.

Deviation control with partial observability. Even when $\hat{\tau}$ is an upper bound, children may deviate transiently because they are learning. The crucial point is that once $\hat{\tau}_b(w)$ exceeds $\tau_b^{\star,p}(w)$ by a margin, the induced utility gap in favor of b is *uniformly positive*. Indeed, for any deviation $a \neq b$,

$$\mathbb{E}[U^w \mid A^w = b] - \mathbb{E}[U^w \mid A^w = a] \;\geq\; p\big(\hat{\tau}_b(w) - \tau_b^{\star,p}(w)\big) \;\triangleq\; \Delta_{w,b}.$$

If w runs a no-regret algorithm with regret $r_w(T)$ with respect to its realized utility, then a standard gap-to-count argument implies that the number of rounds in which w plays an action with expected utility at least $\Delta_{w,b}$ below the best response is at most $O(r_w(T)/\Delta_{w,b})$. Consequently, on $\mathcal{E}$,

$$R_v^{\text{dev}}(T) \ \le \ \sum_{w \in C(v)} \tilde{O}\Big(\frac{r_w(T)}{\Delta_{w,b}}\Big),$$

up to the per-round loss scale of $\theta_v \in [0,1]$. This highlights the second role of p: smaller p weakens incentive gaps linearly, so even holding overpayment fixed, deviation control slows by a factor $1/p$. Verified-MAIL compensates by maintaining $\Delta_{w,b}$ bounded away from zero through conservative (upper) payment updates, which is precisely why underestimation is far more damaging than overestimation.

Bandit regret under shifted and enlarged ranges. Conditioning on $\mathcal{E}$, exploitation rounds present v with an adversarially-robust bandit problem over meta-actions α, with rewards equal to θ_v minus a known offset $\sum_w \hat{\tau}_{b_w}(w)$. Because $\hat{\tau}$ can be as large as $\Theta(1/p)$, the effective reward range is

$$R_v \ = \ 1 + \sum_{w \in C(v)} \tau_{\max}(w) \ = \ 1 + \tilde{O}\Big(\frac{|C(v)|}{p}\Big),$$

and standard bandit guarantees scale linearly with R_v. For example, an EXP3-style bound yields

$$R_v^{\text{bandit}}(T) \ = \ \tilde{O}\Big(R_v \sqrt{T\,|\mathcal{A}|^{1+|C(v)|}}\Big)$$

within each epoch, and the MAIL-style epoching/induction over depth converts this into a sublinear rate even though the meta-action space is large: the recursion ensures that only a controlled set of recommendation profiles is explored at each depth, and deeper nodes are allocated longer epochs to stabilize downstream learning.

Per-node and welfare guarantees (dependence on p and depth). Putting the three terms together and inducting on $d(v)$ (distance to leaves), we obtain receipt-robust analogues of MAIL's nodewise regret bounds.

Theorem 5.1 (Sketch of per-node regret with receipts)**.** *Fix $p > 0$. Under the stated sub-Gaussian reward model, unforgeable receipts, and assuming each node's bandit learner is no-regret with respect to its realized utility, Verified-MAIL achieves for every node v,*

$$\mathbb{E}[R_v(T)] \ = \ \tilde{O}\Big(T^{\,1-\frac{1}{2d(v)^2}}\ p^{-\gamma_{d(v)}}\Big),$$

for an explicit exponent γ_d that grows at most polynomially in d (and depends on the chosen confidence schedule). In particular, for fixed depth d, the regret remains $o(T)$ and degrades monotonically as $p \downarrow 0$.

Finally, welfare regret follows by summing nodewise deviations from the welfare-optimal profile and using the standard cancellation of transfers in aggregate welfare. The only receipt-specific complication is that transfers are paid intermittently; however, intermittency affects *utilities* but not *welfare*, since welfare depends on $\sum_v \theta_v$ and transfers sum to zero along edges ex post. Thus, the welfare analysis reduces to bounding (i) the time spent exploring payments and meta-actions and (ii) the externality losses in the (rare) deviation rounds; both are controlled by the same $1/p$-inflated batch lengths and the same incentive-gap arguments above. This yields sublinear welfare regret $o(T)$ for any fixed $p > 0$, with an explicit polynomial deterioration in $1/p$ that becomes sharper as depth increases.

Why some verification is indispensable. Our positive results hinge on a very weak, but qualitatively important, signal: with probability $p > 0$ the parent learns (and can condition transfers on) whether the child complied with the recommendation. When $p = 0$, this channel disappears entirely. At that point, a principal attempting to internalize downstream externalities faces a familiar *identification* problem: if no observable variable is statistically linked to the child's hidden action, then no outcome-contingent scheme can create differential incentives across actions. In a hierarchy, this difficulty is amplified because the agents who *do* observe the relevant consequences (typically descendants) cannot directly contract with the agent who generates them, as transfers only move locally along edges.

Impossibility at $p = 0$: linear welfare regret under outcome-only contracting. We formalize this intuition with a simple constructed instance in which (i) welfare critically depends on an intermediate agent's action, (ii) that action has no effect on any payoff observed by its parent, and (iii) the beneficiary is a descendant who cannot compensate the intermediate agent (since payments cannot flow upward). Even if principals are allowed to use arbitrarily rich history-dependent contracts based on their own observed outcomes, the absence of any action verifiability renders the welfare-optimal profile unenforceable and, crucially, unlearnable from the principal's perspective.

Theorem 5.2 (Outcome-only contracting can incur $\Omega(T)$ welfare regret when $p = 0$)**.** *Suppose $p = 0$ on every edge (so receipts never arrive). There exists a depth-3 path $v \to w \to u$ with binary actions $\mathcal{A} = \{0, 1\}$ and bounded rewards $\theta \in [0, 1]$ such that for any decentralized contracting/learning scheme in which transfers from v to w may depend on v's observed history but not on w's unverifiable action, the induced welfare regret is at least cT for some universal constant $c > 0$.*

Construction. Let v (the parent) be indifferent to w's action and observe no informative outcome about it:

$$\theta_v(\cdot) \equiv 0.$$

Let w (the intermediate agent) privately prefer action 0 regardless of any downstream effects:

$$\theta_w(0, \cdot) = 1, \qquad \theta_w(1, \cdot) = 0.$$

Finally, let u (the descendant) be the welfare beneficiary of w's costly action:

$$\theta_u(\cdot) \;=\; \mathbb{I}\{A^w = 1\}.$$

All remaining dependencies can be taken trivial, and noise can be set to zero (or added sub-Gaussianly without changing the argument). In this instance, per-period welfare is

$$W(A^w) \;=\; \theta_v + \theta_w + \theta_u \;=\; \begin{cases} 1 & \text{if } A^w = 0, \\ 1 & \text{if } A^w = 1, \end{cases}$$

so this raw specification by itself does not create a welfare gap. To produce a strict welfare preference for $A^w = 1$, we can instead set $\theta_w(0, \cdot) = 0$, $\theta_w(1, \cdot) = 0$ (so w is privately indifferent absent transfers) and keep $\theta_u(\cdot) = \mathbb{I}\{A^w = 1\}$, while introducing a tiny private bias for w toward 0 via a constant additive term $\varepsilon > 0$ to action 0 in its realized utility (equivalently, a small action-dependent cost for 1). Then welfare is strictly higher under $A^w = 1$ by $1 - \varepsilon$, while w strictly prefers 0 by ε.

Proof sketch (enforcement failure implies welfare loss). Fix any scheme. Because $p = 0$, any contract from v to w cannot condition on verified compliance (no receipts arrive), and by construction v's own observed reward process is independent of A^w. Hence, for any history h_t^v observed by v, any transfer rule $\tau_t(w) = \tau_t(h_t^v)$ has the same conditional distribution regardless of whether w plays 0 or 1. Therefore, the expected *incremental* transfer received by w from switching actions is identically zero:

$$\mathbb{E}[\text{transfer} \mid A^w = 1, h_t^v] - \mathbb{E}[\text{transfer} \mid A^w = 0, h_t^v] = 0.$$

Consequently, w's best response is determined solely by its private payoff (and any downstream contracts it offers, which cannot create positive net incentives here because u cannot pay w and u's action does not affect w's observed outcomes). Thus w plays its privately preferred action in essentially every round, independent of the principal's algorithm. Since the welfare gap between the welfare-optimal action and w's privately preferred action is a constant per round, the welfare regret is linear: $\Omega(T)$.

Interpretation and limitations. The force of Theorem 5.2 is not the particular numbers but the informational structure: when upstream nodes observe no statistic correlated with the hidden action, *no amount of learning* can recover the missing incentive lever. In practice, this is exactly the situation in large marketplaces and supply chains when a platform (or regulator) sees only coarse outcomes and cannot audit or attest effort/behavior at the relevant margin. Of course, the lower bound is not universal: if outcomes observed by the principal were sufficiently informative about the agent's action (or if agents could make credible reports, or if side-payments could travel upward), then outcome-only mechanisms could succeed. Our point is that hierarchical externalities naturally generate environments where such informativeness fails, making even a tiny verification probability $p > 0$ qualitatively transformative.

A matching lower bound: inducement must inflate by at least $1/p$ when $p > 0$. The second negative result is a tight necessity statement behind the $1/p$ scaling we exploit algorithmically. Even allowing arbitrary history dependence and randomization, if payments are nonnegative and can be conditioned on compliance only when a receipt arrives, then the maximal *expected* incremental bonus from complying in any single round is at most $p\tau$. Thus, whenever the child has a one-shot temptation gap $g > 0$ in favor of some deviation, the promised transfer must satisfy $\tau \geq g/p$.

Lemma 5.3 ($1/p$ inflation is information-theoretically necessary)**.** *Fix an edge (v, w) and a target action b. Suppose that in the absence of transfers, w's continuation-value gap between its best alternative and b is*

$$g \triangleq \max_{a \in \mathcal{A}} \mu_w(a) - \mu_w(b) > 0.$$

Consider any contract that (i) pays w only upon receipt-confirmed compliance with b, (ii) uses a nonnegative payment $\tau \geq 0$, and (iii) faces receipt probability $p \in (0, 1]$. Then any incentive-compatible contract must satisfy $\tau \geq g/p$.

Proof sketch. If w plays b, it receives the transfer only when the receipt arrives and confirms compliance, which happens with probability p; if it deviates, it receives no compliant payment. Hence the *largest* possible increase in expected utility from choosing b rather than a deviation a coming from transfers is $p\tau$. Incentive compatibility requires this to offset the temptation gap g, i.e., $p\tau \geq g$, yielding $\tau \geq g/p$. This matches the closed-form $\tau_b^{\star,p}(w) = \tau_b^{\star}(w)/p$ and shows there is no mechanism-design "trick" that can avoid the $1/p$ blow-up under our payment and verifiability restrictions.

Takeaway. Together, these lower bounds delineate the frontier for decentralized learning with local incentives: some verifiability is essential for efficiency restoration, and when verification is rare, the required transfers (and the time needed to learn them reliably) must deteriorate proportionally. This sets the stage for design questions: when verification is costly, where should we spend it, and how does the optimal choice of p vary with depth and externalities?

Comparative statics and design implications. Our results can be read as a design map: verification probability p is not merely a technical parameter, but the lever that trades off *contracting power* against *measurement effort*. When p rises, two forces move in tandem. First, the *nominal* transfer needed to create a given incentive wedge shrinks as $\tau_b^{\star,p}(w) = \tau_b^\star(w)/p$. Second, the *statistical* speed at which a principal can learn the right transfers (and detect misalignment) improves because the principal sees a less aggressively subsampled stream of compliance evidence. These two channels are complementary: higher p reduces both the size of the payments we must promise and the time we spend in overpaying/underpaying regimes while learning.

Choosing p when verification is costly. A natural extension—often the practical starting point—is that principals can invest in verification. Formally, imagine that on each edge (v, w) the principal can choose a verification probability $p_{v\to w} \in [0, 1]$ at a per-round resource cost $c_{v\to w}(p_{v\to w})$, with $c(\cdot)$ increasing and typically convex (reflecting that pushing audit rates toward 1 is disproportionately expensive). Even without solving the full dynamic program, our comparative statics identify the marginal benefit of increasing p: it reduces the "friction" terms that scale polynomially in $1/p$ (payment estimation error and undetected deviations), while leaving the underlying payoff externalities unchanged.

A useful back-of-the-envelope objective is to pick p to balance verification cost against the p-dependent learning loss. If, for a node at effective depth d, the algorithmic loss behaves like

$$\text{Loss}(p) \;\approx\; \tilde{O}\big(T^{1-\alpha_d}\, p^{-\gamma_d}\big)\,,$$

and verification cost accumulates linearly as $T\,c(p)$, then the optimal p is pinned down by equating marginal cost and marginal benefit:

$$T\,c'(p) \;\approx\; \gamma_d\,\tilde{O}\Big(T^{1-\alpha_d}\, p^{-(\gamma_d+1)}\Big)\,.$$

This highlights two qualitative predictions. (i) Longer horizons favor higher p only weakly: as T grows, the learning term shrinks relative to $T\,c(p)$ if c is not too steep, so even modest verification can be enough asymptotically.

(ii) Depth matters sharply: larger d (more layers of incentive propagation) increases γ_d, raising the marginal value of verification, so hierarchies with long chains should either verify more or accept substantial efficiency losses.

Transfers, burn, and liquidity: why small p can still be painful. A subtle point is that the $1/p$ inflation concerns the *promised* payment conditional on a receipt, not necessarily the *expected* payment under perfect compliance. Indeed, if a child faces a one-shot temptation gap g and we set $\tau = g/p$, then the expected transfer paid upon compliance is $p\tau = g$, apparently independent of p. This observation is correct but incomplete for design. Small p increases the *variance* and *peakiness* of payouts: the principal pays rarely, but in large lumps. This creates engineering and institutional constraints that do not appear in the expectation. For example, limited liability, escrow requirements, capital costs, or risk-sensitive agents can make high conditional payments infeasible even if expected payments are modest. Moreover, in learning mode, principals typically maintain *upper confidence* payments $\hat{\tau} \geq \tau^{\star,p}$ to guarantee compliance with high probability; since the time to tighten these upper bounds expands as $1/p$, low p can induce substantial cumulative *overpayment* before convergence. In short, even when expected burn under full information is stable, low verifiability raises real-world frictions through variance, constraints, and slower calibration.

Where to deploy verification in a hierarchy. If verification resources are limited, the central allocation question is: *which edges should be attested more heavily?* Our model suggests three prioritization principles.

First, verify *bottleneck* edges: those whose child subtrees generate large welfare externalities upstream. A convenient statistic is the marginal welfare value of aligning w's action with v's recommendation, aggregated over the descendants whose payoffs depend on w. Edges that sit above large or high-stakes subtrees amplify any deviation and therefore have higher returns to verification.

Second, verify where the *temptation gaps* are large. If $g_{v\to w} = \max_a \mu_w(a) - \mu_w(b)$ is large, then both the required conditional transfer and the risk of miscalibration are more severe. Verification is especially valuable in such edges because it directly reduces the necessary τ and accelerates accurate estimation of τ.

Third, verify closer to the *top* when errors propagate. In a deep tree, a small compliance failure rate at low levels can percolate into large effective noise in the rewards perceived at higher levels. Increasing p on upper edges can stabilize the contracting interface that "summarizes" an entire subtree for its parent, thereby reducing the compounding of uncertainty. Conversely, when externalities are largely local within subtrees, investing in verification deeper down may be more efficient. Which effect dominates is an empirical question, but the framework clarifies what must be measured: subtree externality strength versus propagation sensitivity.

Dynamic and heterogeneous verification. Verification need not be uniform over time. Because the learning burden is front-loaded, it can be optimal to use a high verification rate early to estimate inducing payments quickly and then taper p downward once contracts have stabilized. This mirrors classical "explore-then-exploit" logic, but applied to monitoring rather than actions. Heterogeneity across edges is equally natural: some relationships admit cheap, reliable receipts (e.g., digitally logged tasks), while others require expensive audits (e.g., quality inspections). In such environments, the model predicts that the organization should *reshape* incentive schemes to route effort toward verifiable margins, reserving scarce audits for the most distortionary hidden actions.

Marketplace and platform takeaways. Many online marketplaces already approximate our receipt process: a platform can sometimes attest to behavior (GPS pings, time stamps, completion logs, cryptographic proofs of delivery), but not always, and not for every dimension of quality. The p-logic suggests two operational implications. First, "random audits" are not merely deterrence devices; they are also *learning accelerators*. Even a small audit probability can make incentive schemes identifiable and thus optimizable, but the calibration time and the required conditional bonuses scale roughly with $1/p$. Second, engineering choices that slightly increase effective verifiability—better instrumentation, tamper-resistant logs, third-party attestations, standardized reporting APIs—can have outsized welfare effects in deep contracting chains because they reduce both incentive inflation and compounding statistical error.

Policy implications. From a regulatory perspective, the results frame minimum audit requirements as an efficiency instrument rather than a purely compliance-oriented constraint. If a market is organized as a hierarchy of subcontracting with limited upward transfers, then even small mandated verification (or standardized attestations) can unlock mechanisms that otherwise fail. At the same time, mandating very high p may be wasteful when verification costs are convex; our comparative statics point toward targeted verification: require strong attestations on bottleneck relationships (e.g., where hidden actions create large externalities), while allowing lighter-touch monitoring elsewhere.

Design summary. Verification probability p affects decentralized learning through (i) the size and feasibility of receipt-contingent incentives and (ii) the statistical rate at which principals can safely reduce overpayments and control deviations. Optimal system design therefore treats verification as a scarce, allocable resource: invest in it where it stabilizes the most consequential incentive interfaces, use it aggressively early when learning, and complement it with contract structures and platform instrumentation that raise effective p on the margins that matter.

Extensions: from trees to small DAGs. Our baseline analysis exploits the rooted-tree structure to define a clean upstream/downstream recursion: each node summarizes its entire subtree to its parent through an induced continuation value $\mu_v(\cdot)$. Many real contracting networks are not trees but *small DAGs*, where a unit may report to (or affect) multiple principals, or where downstream actions generate spillovers to multiple upstream parties. A direct extension is feasible when the in-degree is bounded and cycles are absent: we can interpret each node as signing multiple local contracts $\{(B_t^{(i)}(v), \tau_t^{(i)}(v))\}_i$ with different parents $P_i(v)$, and its bandit objective becomes the sum of its own reward plus the receipt-triggered transfers from each parent. The key technical change is that the child's incentive constraint now depends on a *vector* of promised payments, but under additive transfers the wedge remains linear, and the minimal inducing *aggregate* payment still scales as $1/p$. What becomes nontrivial is *cost sharing*: which parent should finance which part of the inducing payment, especially when each parent's benefit from compliance differs. One pragmatic resolution is to designate a "lead principal" on each shared child and allow side transfers among principals (or an internal accounting rule) so that the effective contract still resembles the tree case. Alternatively, one can work with a spanning arborescence for contracting and treat remaining DAG edges as externalities in rewards; this preserves tractability but may increase the temptation gaps $\tau_b^\star(\cdot)$ that must be covered. By contrast, true cycles raise a distinct difficulty: contracts can become self-referential because a node's continuation value depends on incentives that depend on that continuation value. Handling cycles likely requires either equilibrium selection assumptions (e.g., a fixed point in stationary contracts) or centralized clearing/escrow that breaks circular dependence. We view bounded-width DAGs as the most relevant near-term extension, and the tree model as the appropriate first-order approximation when organizations enforce a primary reporting line even if "dotted-line" influence exists.

Budgets, limited liability, and liquidity constraints. Receipt-contingent payments that inflate as $1/p$ immediately bring feasibility constraints to the foreground. Two common constraints are (i) limited liability on the agent side (already respected in our baseline by restricting to $\tau \geq 0$), and (ii) limited liquidity or budgets on the principal side, e.g.,

$$\sum_{w\in C(v)} \tau_t(w) \;\le\; \bar{B}_v \qquad \text{or} \qquad \tau_t(w) \;\le\; \bar{\tau}_{v\to w}.$$

Under such constraints, low p can create *hard* incentive failures: even when the expected burn $p\tau$ is moderate, the conditional transfer τ required for compliance may exceed $\bar{\tau}$ or violate the per-round budget. This is not merely a technicality; it predicts sharp regime changes in practice. When the constraint binds, principals must either (a) increase verification p, (b) recommend a second-best action b with a smaller temptation gap, or (c) redesign the contracting interface (e.g., reduce the action space to verifiable proxies, simplify tasks, or shift to repeated-relationship discipline where future access substitutes for payments). These constraints also interact with learning: algorithms like Verified-MAIL naturally maintain conservative (upper-confidence) payments to protect compliance, but with caps $\bar{\tau}$ the algorithm may be forced into under-incentivizing during exploration. A simple adaptation is *feasibility-aware exploration*: when the current upper estimate $\hat{\tau}_b$ exceeds the cap, the principal treats action b as infeasible under current p and reallocates exploration either toward increasing p (if choice is available) or toward alternative recommended actions. In short, budgets and liquidity turn the smooth $1/p$ comparative statics into a constrained design problem in which verification can be the only lever that preserves implementability.

Noisy verification: false negatives and false positives. The receipt model can also be relaxed to allow noisy attestations, a natural concern when monitoring is automated or partially manipulable. Suppose that with probability p a signal arrives, but conditional on arrival it may be wrong: if the agent plays the recommended action b, the receipt matches b only with probability $1-\varepsilon_{\mathrm{fn}}$ (false negatives), and if it plays $a \neq b$, the receipt falsely reports b with probability $\varepsilon_{\mathrm{fp}}$ (false positives). Under the same "pay only when the receipt says b" rule, the incentive difference between choosing b and deviating to $a \neq b$ becomes

$$\big(\mu_w(b) - \mu_w(a)\big) \;+\; p\big((1-\varepsilon_{\mathrm{fn}}) - \varepsilon_{\mathrm{fp}}\big)\tau.$$

Thus the minimal inducing payment inflates to

$$\tau_b^{\star,p,\varepsilon}(w) \;=\; \frac{\max_{a\in\mathcal{A}} \mu_w(a) - \mu_w(b)}{p\big(1-\varepsilon_{\mathrm{fn}}-\varepsilon_{\mathrm{fp}}\big)},$$

provided $1-\varepsilon_{\mathrm{fn}}-\varepsilon_{\mathrm{fp}} > 0$. This formula clarifies an operational threshold: if false positives plus false negatives approach one, then receipts cease to create a reliable incentive wedge regardless of nominal audit rate p. Learning is also affected asymmetrically. False negatives slow detection and inflate payments much like a reduction in effective p. False positives are more pernicious because they subsidize deviation, increasing the payment needed *and* inducing payment burn on off-path behavior. Algorithmically, the same batched estimation logic can be applied by replacing observed compliance counts with de-biased estimators or by explicitly modeling the confusion matrix $(\varepsilon_{\mathrm{fn}}, \varepsilon_{\mathrm{fp}})$ when it is known or can be calibrated. The broader message is that "verification quality" is a first-class design primitive: investments that reduce false positives can be more valuable than equivalent increases in audit frequency.

Empirical evaluation: what we would measure and how. Our framework yields sharp, testable predictions about how monitoring interacts with incentives and learning dynamics. An empirical evaluation can be conducted either in a platform environment (marketplaces with random audits and digital traces) or within a firm (multi-layer task assignment with spot checks). The core design is an experiment that varies verification probability p on a subset of edges (teams, vendors, or task categories) while holding task pools and recommended policies fixed. We would track: (i) *promised conditional bonuses* τ and their dispersion across time; (ii) *realized payments* (which should scale much less strongly with p than promised payments under high compliance); (iii) *calibration time*, measured as the time until promised payments stabilize within an ε-band; and (iv) *performance externalities*, i.e., how downstream compliance affects upstream rewards in the hierarchy. The model predicts that (a) conditional bonuses scale approximately like $1/p$, (b) convergence times scale approximately like $1/p$ holding confidence targets fixed, and (c) deeper or more externally coupled subtrees exhibit a higher marginal value of verification. A practical implementation of Verified-MAIL would require only local logging of receipts and transfers, making it amenable to field deployment; welfare can be proxied by aggregate objective metrics (delivery times, defect rates, customer satisfaction) net of transfers if transfers are internal, or gross of transfers if payments are external and represent real resource costs.

Conclusion. The central tradeoff illuminated by the model is that partial verifiability is not a binary obstacle but a quantitative friction that simultaneously constrains *incentive power* and *learning speed*. A strictly positive receipt probability $p > 0$ is enough to restore asymptotic efficiency in deep hierarchies, but the path to that limit can be costly when p is small because the mechanism must promise large, rare payments and must learn them from a thinned stream of evidence. The extensions above emphasize where the theory is likely to bend in practice: networked (non-tree) influence patterns require cost sharing and careful interface design; budgets and liquidity impose hard feasibility constraints that make verification an essential substitute for transfers; and noisy verification introduces a quality threshold beyond which auditing ceases to discipline behavior. Taken together, these considerations suggest a concrete organizational lesson: investing in reliable attestations and designing verifiable work interfaces can be as important as optimizing the incentives themselves, particularly in long contracting chains where small distortions compound.