Position: AI-Agent Pricing Should Become More Outcome-Dependent: An Economic Perspective

Yuheng Bu
Department of Computer Science
University of California, Santa Barbara,
Santa Barbara, CA 93106
buyuheng@ucsb.edu

Yueyuan Ma
Department of Economics
University of California, Santa Barbara,
Santa Barbara, CA 93106
yueyuanma@ucsb.edu

Abstract

AI agents have advanced rapidly and are increasingly sold as services, yet dominant pricing models still reward observable usage rather than the outcomes users actually value. This position paper argues that AI-agent pricing should become more outcome-dependent as markets mature, especially in domains where success can be measured objectively. Our position is motivated by three factors. First, pricing is a risk-sharing mechanism: when users are more risk-averse than AI-agent providers, outcome-dependent contracts can improve welfare by reallocating risk more efficiently. Second, token-based pricing under-incentivizes hidden effort, such as verification, tool use, and model selection, thereby creating moral hazard. Third, outcome-dependent pricing is often difficult to implement because performance measures may be noisy, delayed, or strategically manipulated. Using stylized principal-agent models from economics, we show how these forces shape the case for outcome-dependent pricing and why hybrid contracts that combine usage-based charges with outcome-dependent terms are often more practical than either token-only or pure pay-for-performance schemes. Overall, we argue that moving beyond token-only pricing is necessary for allocating risk more efficiently, aligning incentives, and building more trustworthy AI-agent markets.

1 Introduction

AI agents have advanced rapidly, evolving from simple chatbots into tool-using systems that can draft documents [1, 2], write and debug code [3], address customer requests [4], and support complex decision workflows [5, 6]. As these capabilities move into economically meaningful deployment, a central bottleneck for broad adoption is not only technical performance but also how agents are priced and contracted. Pricing governs adoption, allocates risk, and ultimately determines whether users trust the product.

However, there is still no consensus on the "right" pricing model. Most AI agents are sold via usage-based charges (e.g., tokens) [7, 8] or subscription-style fixed seat fees [9]. These schemes are easy to meter, but they are often poorly aligned with what users value in agentic deployments: whether tasks are completed successfully, how reliable the results are, and who bears the downside risk when the agent fails. At the same time, providers make many costly but hidden choices that affect quality, including verification, tool use, and model selection. Pricing purely on observable usage, therefore, risks rewarding visible activity rather than realized value.

Our position is that **AI-agent pricing should be analyzed through an economic lens and should become more outcome-dependent as markets mature**. Contracts should be evaluated by the welfare they generate, not by the convenience of metering usage. Because AI-agent markets require both efficient risk sharing and incentive alignment, pricing should place greater weight on verified

Preprint.

Table 1: From economic factors to contract design in AI-agent pricing.

Economic factor	Why token pricing is insufficient	Contract implication
Asymmetric exposure to outcome risk	Users may bear substantial downside risk from task failure, while providers can diversify across customers and tasks.	Adopt outcome-based contract to improve risk sharing.
Hidden quality effort	Token-based pricing does not reward unobserved quality effort.	Use hybrid contracts: usage-based charges plus outcome-dependent pay.
Imperfect outcome measurement	Outcomes may be noisy, delayed, disputed, or strategically manipulated.	Reduce outcome dependence when measurement is weak; complement with auditing and governance.

outcomes, especially where success can be measured reliably. At the same time, we do not argue that all pricing should immediately become pure pay-for-performance. As outcome measurement is often imperfect, a more practical approach is often *hybrid* contracting, which combines usage-based charges with outcome-based pricing. This view is also beginning to surface in industry discussions of AI services and SaaS commercialization, where outcome-based pricing is increasingly seen as a natural direction for agentic products [10, 11].

We motivate our position through three factors using stylized principal-agent (PA) models from economics [12], in which the user (principal) delegates a task to the AI provider (agent). We use these stylized PA models not to claim that all AI-agent markets literally satisfy their assumptions, but to isolate these factors and support our position in a transparent way. The main takeaways are summarized in Table 1.

The first is *risk sharing*, which we believe is underappreciated in current discussions of AI-agent pricing. Classical PA theory typically studies settings in which the principal is risk-neutral, and the agent is risk-averse [13, 14]. In many AI-agent markets, however, the economic roles may be *reversed*: users may be highly exposed to uncertain task outcomes, while large AI companies can diversify risk across many customers, tasks, and time. From this perspective, outcome-based pricing is valuable because it can shift risk toward the side better able to bear it.

The second factor is providing incentives for *hidden effort*. Delivering an AI agent requires more than observable token usage. Providers also make hidden choices that often drive success, including tool use, verification, model or version selection, and human-in-the-loop monitoring, many of which are difficult for users to observe or contract on directly. This is the familiar logic of moral hazard in PA models [15, 16]: if compensation depends only on what is easily measured, hidden but value-relevant effort may be underprovided. It may even encourage behavior that increases token usage or shifts requests to lower-cost AI models [17, 18, 19].

The third factor is *evaluation frictions*. Outcome-dependent pricing is only practical when outcomes are verifiable, timely, and hard to game. Some AI-agent applications admit relatively clear and auditable metrics, while others depend on subjective evaluation, delayed feedback, or proxies based on human judgment or LLM-as-judge [20, 21]. Once pricing depends on such metrics, both principals and agents may also have incentives to manipulate the evaluation signal. This measurement challenge helps explain why outcome-based pricing remains limited in practice. The relevant question, therefore, is not whether pricing should ever depend on outcomes, but when outcomes are verifiable enough to support such dependence, and when hybrid contracts plus auditing and governance are more realistic.

Finally, we emphasize the scope of this paper. Our aim is narrower than a full market-equilibrium theory (no externality) of AI-agent pricing, but broader than a purely technical contract-design exercise. Within this scope, our goal is to clarify the economic factors governing AI-agent pricing and to argue that pricing is part of the infrastructure of trustworthy AI deployment: it allocates uncertainty, shapes incentives, and determines whether providers are rewarded for the outcomes users actually value. From this perspective, moving beyond token-only pricing is not just a business-model question; it is part of how AI-agent markets become more efficient, reliable, and mature.

2 Outcome-Dependent Pricing Is Needed for Efficient Risk Sharing

Outcome-dependent pricing is often motivated by incentive alignment in contract theory: when effort is hidden, compensation should depend on outcomes rather than only on observable effort [13, 14] to mitigate moral hazard. More broadly, tying pay to performance is a standard way to align interests when individual contributions are hard to measure, as in startup compensation that emphasizes equity over cash [22, 23]. Our first claim is that **outcome-dependent pricing is needed not only for incentives, but also for efficient risk sharing**.

Consider a customer-support AI agent sold to a small business (principal)[1]. If the agent performs poorly, the principal may bear substantial losses, including unresolved tickets, dissatisfied customers, and reputational harm. By contrast, the AI provider can often diversify this uncertainty across many users, tasks, and time. The provider may therefore be better positioned to absorb or pool outcome risk than any single principal. From this perspective, a fixed-fee or purely token-based contract can be unattractive because it leaves too much uncertainty on the principal's side. Outcome-dependent pricing becomes valuable because it shifts risk toward the side better able to bear it.

We formalize this point using a stylized PA benchmark. We adopt the standard CARA-Normal framework because it yields closed-form comparisons and makes the role of risk sharing explicit.

2.1 Benchmark 0: Classical CARA-Normal PA Models

A principal contracts with an agent who chooses effort e and incurs cost $c(e)$ to complete a task. The realized outcome of the principal is assumed to follow a Normal distribution

$$\pi \mid e \sim \mathcal{N}(\mu(e), \sigma^2), \tag{1}$$

where the noise level σ^2 is independent of e and the mean $\mu(e)$ is increasing in e. As in much of the contract-theory literature, the Normal assumption is best understood as an approximation for aggregate performance measures. Our qualitative conclusions are intended to be directional rather than tied to this exact assumption. A contract specifies a wage $w(\pi, e)$ when effort is observable and contractable, and a wage $w(\pi)$ when effort is unobservable.

In the classical setting, e.g., a large firm (principal) hires workers (agents) to perform jobs, the principal is typically *risk-neutral* and maximizes expected profit $\max_{w(\cdot),\, e}\ \mathbb{E}[\pi - w(\pi, e)]$. The agent is *risk-averse* and has constant absolute risk aversion (CARA) utility

$$u_a(x) = -\exp(-r_a x), \qquad r_a > 0, \tag{2}$$

and derives utility from the net payoff $x = w(\pi, e) - c(e)$. Under CARA-Normal assumptions, expected utility admits a certainty-equivalent (CE) form: for normally distributed X, maximizing $\mathbb{E}[u_a(X)]$ is equivalent to maximizing $\mathrm{CE}_a(X) = \mathbb{E}[X] - \frac{r_a}{2}\mathrm{Var}(X)$.

Let $\bar{u}$ denote the agent's reservation utility in CE. The agent will accept the contract if

$$\mathrm{CE}_a(w(\pi, e) - c(e)) \geq \bar{u}. \tag{3}$$

This is called the individual rationality (IR) constraint.

If effort is unobservable, wages can depend only on π. The agent then chooses effort to maximize her expected utility, so the contract must satisfy an incentive compatibility (IC) constraint. Specifically, the principal solves

$$\max_{w(\cdot)}\ \mathbb{E}[\pi - w(\pi)], \quad \text{s.t.} \quad \begin{aligned} &\text{(IR) } \mathrm{CE}_a(w(\pi) - c(e)) \geq \bar{u}, \\ &\text{(IC) } e \in \arg\max_{e'}\ \mathrm{CE}_a(w(\pi) - c(e'))\,. \end{aligned}$$

In this classical benchmark, if effort is observable, the principal can implement the first-best allocation with a *fixed wage*. If effort is hidden, the principal must rely on *outcome-based pay*, but this is second-best because it exposes the risk-averse agent to outcome risk. Thus, outcome dependence arises mainly as a response to *moral hazard*. More details and an illustrative linear-quadratic example are provided in Appendix B.

[1] Examples include startups Intercom's Fin, which is priced per resolved conversation, `https://fin.ai/pricing`; and Sierra, which markets outcome-based pricing for its agents, `https://sierra.ai/`.

2.2 Benchmark 1: Risk-neutral Agent and Risk-averse Principal

The economic roles may be reversed in AI-agent markets. Principals may be highly risk-averse about uncertain task outcomes, while large AI providers can diversify across many contracts and are effectively less risk-averse. This reversal changes the role of outcome-dependent pricing.

Suppose the principal has CARA utility with risk aversion parameter $r_p > 0$ as in (2), while the agent is *risk-neutral* and maximizes expected net payoff $\mathbb{E}[w(\pi)] - c(e)$, with reservation utility $\bar{u}$.

Lemma 2.1 (Risk-averse principal and risk-neutral agent)**.** *Suppose the principal has CARA utility with risk aversion parameter $r_p > 0$, while the agent is risk-neutral with reservation utility $\bar{u}$. Then, under either observable or unobservable effort, the first-best effort*

$$e^{\mathrm{FB}} \in \arg\max_{e} \ \mu(e) - c(e) \tag{4}$$

can be implemented by the residual-claimant contract

$$w(\pi) = \alpha + \pi, \qquad \alpha = \bar{u} + c(e^{\mathrm{FB}}) - \mu(e^{\mathrm{FB}}).$$

Interpretation. Lemma 2.1 shows that if the agent (AI provider) is willing to bear all risk and payments are unrestricted, the principal (user) can offer a contract whose compensation depends only on the final outcome π. The principal's net payoff then becomes constant, $\pi - w(\pi) = -\alpha$, so first-best effort can be implemented even when effort is unobservable. This is an extreme benchmark because it ignores limited liability and other practical constraints. In effect, the principal sells the task to the agent and bears no risk. Still, it clarifies the key point: *purely outcome-based pricing can be optimal even without moral hazard, because it reallocates risk.*

2.3 Benchmark 2: Both Parties Are Risk-Averse

In practice, AI companies are owned by shareholders who are also risk-averse. Although companies can diversify risk across many tasks and customers, some systematic risks may remain and cannot be fully eliminated through diversification. Thus, a more realistic benchmark is one in which both sides are risk-averse, with the provider being less risk-averse than the user.

Suppose the principal has CARA utility with absolute risk aversion $r_p > 0$ and the agent has CARA utility with absolute risk aversion $r_a > 0$, with $r_a < r_p$. The following lemma characterizes the optimal contract under observable versus unobservable effort.

Lemma 2.2 (Both parties are risk-averse)**.** *For the CARA-Normal PA Models, it is without loss of generality to restrict attention to affine contracts $w(\pi) = \alpha + \beta\pi$.*

(i) Observable effort. *The first-best effort maximizes total surplus as in* (4)*, and the optimal outcome sensitivity is*

$$\beta^{\mathrm{FB}} = \frac{r_p}{r_a + r_p}, \tag{5}$$

with α chosen so that the agent's participation constraint binds.

(ii) Unobservable effort. *If effort is unobservable and compensation can depend only on π, then for any β, the induced effort satisfies*

$$e(\beta) \in \arg\max_{e'} \ \beta\mu(e') - c(e'). \tag{6}$$

At the optimum, the participation constraint binds, and the principal chooses $\beta > \beta^{\mathrm{FB}}$ to balance incentive provision and risk sharing.

A linear-quadratic illustration. Consider $\mu(e) = e$ and $c(e) = \frac{k}{2}e^2$. Then $e^{\mathrm{FB}} = \frac{1}{k}$. With observable effort, the first-best outcome sensitivity remains (5). With unobservable effort, the agent chooses $e(\beta) = \beta/k$, and the second-best outcome sensitivity is

$$\beta^{\mathrm{SB}} = \frac{1 + k\sigma^2 r_p}{1 + k\sigma^2(r_a + r_p)}, \qquad e^{\mathrm{SB}} = \frac{\beta^{\mathrm{SB}}}{k}. \tag{7}$$

Hence $e^{\mathrm{SB}} < e^{\mathrm{FB}}$ and $\beta^{\mathrm{SB}} > \beta^{\mathrm{FB}}$, so hidden effort pushes the contract toward stronger outcome dependence than the pure risk-sharing benchmark. More details can be found in Appendix C.

Interpretation. Lemma 2.2 highlights a distinct risk-sharing role for outcome-based contracts. Even when effort is observable and incentive alignment is irrelevant, outcome-dependent pricing can still be first-best because it allocates uncertainty more efficiently between the two parties. A contract tied to realized performance is therefore not merely a device for inducing hidden effort; it is also a mechanism for insuring the more exposed party.

2.4 Risk Sharing in Outcome-Based Pricing

This risk-sharing perspective is particularly relevant for AI-agent markets. A user purchasing an agent often cares about whether a task succeeds and may be highly exposed to failure, while the provider can pool risk across many deployments. Under this asymmetry, purely fixed-fee or token-only pricing can be inefficient because it leaves too much uncertainty on the principal's side. Outcome-dependent pricing instead shifts some of that uncertainty toward the side better able to bear it.

Many service markets already use outcome-dependent pay for exactly this reason. In **contingency-fee litigation**, the client pays little or nothing unless they win, shifting downside risk to the law firm. In **gain-sharing consulting**, compensation is partly tied to agreed KPIs, so the consultant shares performance risk. The common thread is that uncertainty shifts toward the party better able to absorb or manage it, often through expertise or diversification. Applied to AI-agent pricing, this suggests a clear adoption prerequisite: providers must credibly take on some outcome risk rather than shifting it all onto users. That requires trust, sufficient capital, and enforceable commitments. As these conditions are met, outcome-dependent contracts should become more common, and AI-agent markets should mature as adoption broadens.

3 Pricing Should Reward Hidden Quality Effort

In practice, LLM usage is most often priced through usage-based billing [7, 8], such as per-token charges. Yet token counts measure only observable usage, not the behind-the-scenes choices that often determine whether the agent actually performs well.

This is especially clear in settings such as customer support. Two AI agents may consume similar numbers of tokens per ticket, yet deliver very different outcomes because they differ in retrieval quality, verification, escalation decisions, and monitoring. These quality-relevant efforts are often costly to the AI provider but only weakly reflected in token usage. As a result, token-only pricing rewards observable activity rather than the hidden effort that often drives realized value.

This motivates our second claim: **when important quality effort is only partially observed, pricing should not rely on tokens alone. Instead, it should include an outcome-dependent component.** In many practical settings, this points toward *hybrid pricing*, which combines usage-based charges for observable effort with outcome-dependent terms that reward hidden quality effort. We formalize this point using a stylized two-effort benchmark.

3.1 A Stylized Model of Observable and Hidden Effort

An agent chooses effort $e = (e_1, e_2)^\top \in \mathbb{R}^2$, where e_1 is observable and contractable, while e_2 is hidden and non-contractable. In an AI-agent interpretation, e_1 can be viewed as token usage, while e_2 captures hidden quality effort, such as verification, retrieval quality, model selection, etc. Output is

$$\pi = \mu(e_1, e_2) + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{8}$$

where ε is independent of effort. Effort incurs a strictly convex cost $c(e_1, e_2)$, and both the principal and the agent have CARA preferences.

To obtain clear comparisons across contract forms, we specialize in a linear-quadratic benchmark:

$$\pi = \theta^\top e + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2), \tag{9}$$

with

$$c(e) = \frac{1}{2} e^\top K e, \qquad K = \begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{pmatrix} \succ 0. \tag{10}$$

We compare four contracting environments that differ in which variables are contractable and/or usable for pay:

0. (first-best) both e_1 and e_2 are contractable;
1. (token-only pricing) only e_1 is contractable, and pay cannot depend on output;
2. (outcome-only pricing) neither effort is contractable, and pay depends only on output;
3. (hybrid pricing) e_1 is contractable, and pay can also depend on output.

3.2 What the Model Shows

The full characterizations are provided in Appendix D. The next proposition summarizes the ranking implied by the benchmark, where superscripts denote the different contract forms.

Proposition 3.1 (Comparison across contract forms)**.** *Under the linear-quadratic CARA-Normal benchmark, the following comparisons hold.*

(i) Observable effort. *The effort levels e_1 satisfy $e_1^{(0)} = e_1^{(1)} = e_1^{(3)},\ e_1^{(2)} < e_1^{(0)}$.*

(ii) Hidden effort. *The hidden effort e_2 satisfies $e_2^{(2)} < e_2^{(0)},\ e_2^{(1)} < e_2^{(3)}$.*

(iii) Incentive intensity. *The optimal β satisfy $\beta^{(1)} = 0 \ < \ \beta^{(0)} \ < \ \beta^{(3)} \ \leq \ \beta^{(2)}$.*

(iv) Principal welfare. *If $r_a \leq r_p$, then $\mathrm{CE}_p^{(0)} \ \geq \ \mathrm{CE}_p^{(3)} \ \geq \ \mathrm{CE}_p^{(2)} \geq \mathrm{CE}_p^{(1)}$.*

Below, we summarize the main takeaways.

Token-only pricing (1) under-rewards hidden effort. Proposition 3.1 shows that hidden effort is strictly higher under hybrid pricing than under token-only pricing, i.e., $e_2^{(1)} < e_2^{(3)}$. The reason is that when pricing depends only on the observable component e_1, the provider chooses the hidden effort e_2 to minimize its own cost conditional on the contracted usage level. The contract, therefore, rewards what is easy to measure, but not the hidden effort that often drives quality. In AI-agent settings, token-only pricing can systematically under-incentivize efforts that improve performance but are not directly priced. It may even encourage behavior that increases token usage or shifts requests to lower-cost AI models [17, 18, 19]. Consistent with this, token-only pricing delivers the lowest principal welfare among the four contract forms.

Outcome-only pricing (2) restores incentives, but only crudely. The proposition shows that outcome-only pricing induces both observable and hidden effort, but still falls short of the first-best: $e_1^{(2)} < e_1^{(0)}$ and $e_2^{(2)} < e_2^{(0)}$. At the same time, it requires the greatest outcome sensitivity. Economically, a single outcome-dependent term must do too much at once: it must induce both types of effort while also sharing risk between the two parties. As a result, pure outcome-based pricing improves on token-only pricing on principal welfare, but remains coarse.

Hybrid pricing (3) separates the roles of usage and outcomes. The proposition also explains the advantage of hybrid pricing. Because e_1 is contractable, the contract can regulate observable usage directly, while the outcome-dependent component is reserved for the hidden effort that tokens do not capture. This is why hybrid pricing requires less outcome sensitivity than outcome-only pricing, $\beta^{(3)} \leq \beta^{(2)}$, yet still delivers higher welfare and more hidden effort than token-only pricing. In this sense, hybrid pricing separates the roles of usage and outcomes more effectively than either token-only or pure outcome-based pricing.

Overall interpretation. The proposition yields a clear ranking. Token-only pricing performs worst because it neither shares risk efficiently nor rewards hidden effort. Outcome-only pricing improves on token-only pricing by incentivizing both effort dimensions, but it does so only through a single outcome-dependent term, making it relatively coarse and requiring the greatest outcome sensitivity. Hybrid pricing combines the strengths of both approaches: it uses usage-based pricing to regulate observable effort and outcome dependence to reward hidden effort, which is why it dominates the other two contracts and is the closest to the first-best scenario where both efforts are contractable.

3.3 Implications for AI-Agent Pricing

The benchmark formalizes a basic limitation of token pricing for AI agents: *tokens measure activity, not quality*. If important aspects of agent behavior are only partially observed, then pricing only what is observable will systematically under-reward what users actually care about.

This point is especially relevant in current agent deployments. In customer support, coding assistance, and workflow automation, providers make hidden quality choices that strongly affect outcomes but are not well captured by token counts. A pricing scheme based only on usage rewards observable volume while neglecting hidden quality effort. By contrast, a purely outcome-based contract may put too much weight on noisy realized performance. Hybrid pricing offers a more natural compromise: it preserves usage-based control over observable inputs while introducing outcome dependence precisely where usage metrics are silent.

Our conclusion is not that token pricing is useless. Observable usage remains informative when it tracks real costs or when activity itself matters. The point is that **token pricing alone is generally too narrow for AI agents**. As markets mature, pricing should increasingly combine usage-based charges with outcome-dependent terms that reward the hidden quality effort that ultimately drives value.

4 Implementation Depends on Verifiable Outcomes

If risk sharing and hidden quality effort both point toward more outcome-dependent pricing, a natural question arises: why have AI-agent markets not already converged to it? Our third claim is that the main obstacle is not conceptual, but operational. **Outcome-dependent pricing is only feasible when outcomes are sufficiently verifiable, measurable, and hard to game.** Otherwise, the contract risks becoming a dispute over the metric rather than a mechanism for rewarding real performance.

Return again to the customer support example. Some outcomes are relatively concrete, such as whether a ticket is resolved or satisfaction is reported after the interaction. In those settings, an outcome-dependent contract is plausible. But in many others, the outcome is harder to contract on: quality may be subjective, satisfaction may be noisy, success may only be revealed after a significant delay, and the parties may disagree about whether the agent truly caused the outcome. Once pricing depends on such measures, both sides may also have incentives to manipulate how the signal is recorded or interpreted.

This motivates our third claim: **the case for outcome-dependent pricing depends on the quality of the evaluation signal.** When outcomes are verifiable, outcome-based terms can improve both incentives and risk sharing. When outcomes are noisy or contestable, stronger outcome dependence becomes harder to justify, and hybrid pricing supported by evaluation and auditing becomes more realistic than pure pay-for-performance.

4.1 A Stylized Model of Noisy and Gameable Outcome Measurement

To isolate this force, we consider a reduced-form benchmark in which the agent exerts hidden effort $e \in \mathbb{R}$ and produces a true output

$$y = e + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_y^2), \tag{11}$$

where σ_y^2 captures intrinsic uncertainty in task performance. Effort incurs cost $c(e) = \frac{k}{2}e^2$. The contract cannot depend directly on the true output y, but only on a measured outcome z. We model this measured outcome as

$$z = y + \nu + x_a - x_p, \tag{12}$$

where $\nu \sim \mathcal{N}(0, \sigma_m^2)$ is measurement noise, x_a is manipulation by the agent that inflates the signal, and x_p is manipulation by the principal that deflates it. Manipulation is costly:

$$\text{agent gaming cost: } \frac{d_a}{2}x_a^2, \qquad \text{principal gaming cost: } \frac{d_p}{2}x_p^2.$$

We restrict attention to linear outcome-based contracts $w(z) = \alpha + \beta z$, where both the principal and the agent have CARA utilities with $r_a < r_p$. The two parties behave as follows:

1. The principal offers a contract (α, β).
2. The agent chooses effort e and the manipulation x_a.
3. Uncertainty (ε, ν) is realized.
4. The principal chooses the principal's manipulation x_p.
5. Signal z is finalized and wage w is paid.

The full derivation is given in Appendix D. The main comparative statics are summarized below.

Proposition 4.1 (Outcome-dependent pricing under noisy and gameable evaluation)**.** *In the benchmark above, the optimal affine contract $w(z) = \alpha + \beta z$ has outcome sensitivity*

$$\beta^* = \frac{\frac{1}{k} + r_p\sigma_y^2}{\frac{1}{k} + \left(\frac{1}{d_a} + \frac{1}{d_p}\right) + (r_a + r_p)(\sigma_y^2 + \sigma_m^2)}. \tag{13}$$

The induced equilibrium effort and manipulation levels are

$$e^* = \frac{\beta^*}{k}, \qquad x_a^* = \frac{\beta^*}{d_a}, \qquad x_p^* = \frac{\beta^*}{d_p}.$$

4.2 What the Benchmark Implies

Proposition 4.1 highlights a basic issue for outcome-dependent pricing: the more pricing depends on the measured outcome, the stronger the incentives for both sides to manipulate it. As a result, the outcome signal must be reliable enough to bear contractual weight. Three implications are especially important.

Noisy evaluation σ_m weakens pure outcome-based pay. When the measured outcome is only a noisy proxy for true performance (larger σ_m), putting too much weight on it in pricing adds risk without providing much useful information. This lowers the optimal outcome sensitivity β and makes pure pay-for-performance less attractive.

Gameable metrics can trigger a costly arms race. If either side can manipulate the measurement at low cost, that is, when d_a and d_p are small, then stronger outcome dependence encourages wasteful effort aimed at the metric rather than at true performance. In the extreme case where measurements are very noisy or manipulation is nearly costless, the optimal β collapses toward zero, and we should avoid outcome-based pricing.

Not all tasks are equally contractable. The practical lesson is that outcome-dependent pricing works best in domains where outcomes are clear, timely, and auditable. In customer support, for instance, metrics such as resolved conversations, escalation rates, or post-resolution satisfaction can sometimes support outcome-dependent terms. By contrast, in more subjective, delayed, or multi-stage tasks, the evaluation signal may be too fragile to support a strong outcome-based contract.

Overall, noisy and gameable evaluation turns outcome-based contracting into a contest over the scoreboard: both sides may spend resources manipulating the measured signal rather than improving true output, which reduces surplus and forces the optimal contract to place less weight on outcomes.

4.3 Implications for AI-Agent Pricing

The main bottleneck for scalable outcome-dependent pricing is the evaluation infrastructure. A provider can only be paid on outcomes if both sides can agree on what the outcome is, when it is measured, and how it is attributed. This is why pricing design and evaluation design cannot be separated in AI-agent markets.

This point also clarifies the role of hybrid pricing. Earlier sections argued that token-only pricing is too narrow because it does not share risk efficiently and does not reward hidden quality effort. But this does not imply that the solution is always a pure outcome-only contract. When measured outcomes are imperfect, a more practical design is often hybrid: usage-based charges cover observable and meterable inputs, while an outcome-dependent term is added only to the extent that the evaluation signal is reliable enough to support it.

More broadly, the same evaluation challenge that complicates benchmarking and post-training also constrains contracting. Better benchmarks, standardized task definitions, logging, provenance, and auditing are not merely technical complements to pricing; they are what make stronger outcome-dependent contracts feasible in the first place. As AI-agent markets mature, the spread of outcome-dependent pricing will depend not only on better models, but also on better systems for measuring and verifying what those models actually achieve.

5 Alternative Views

Our position is intentionally narrow: we use stylized PA models to isolate how *risk sharing*, *hidden effort*, and *evaluation frictions* shape AI-agent pricing, and to argue that pricing should become more outcome-dependent as markets mature. At the same time, several countervailing forces help explain why current markets still rely heavily on token-based and subscription pricing. Below, we discuss three such alternative views; additional limitations of our analysis are summarized in Appendix A.

Alternative view 1: Repeated interaction and reputation may substitute for explicit outcome-dependent pricing. Our benchmark analysis focuses on *one-shot* contracting with enforceable terms. In practice, however, many AI-agent relationships are ongoing: enterprise users renew contracts, compare vendors over time, and condition future business on realized performance. This creates a *relational contracting* environment [24, 25], in which performance contingencies may be implemented implicitly rather than through an explicit outcome-dependent term. Common mechanisms include refund guarantees, service credits for missed uptime or resolution targets, and renewal price adjustments tied to realized performance. These arrangements can deliver some of the incentive and risk-sharing benefits of outcome-dependent pricing while reducing per-task measurement disputes. Still, they introduce forces outside our benchmark, including bargaining at renewal, switching costs, and dynamic trade-offs between short-run performance and long-run learning. Relational contracting can therefore partially substitute for explicit outcome-dependent pricing, but it does not eliminate its role, especially when relationships are short-lived, quality is hard to verify, or users remain highly exposed to downside risk within each contract period.

Alternative view 2: Governance and monitoring may matter more than payment form. A second view is that, when outcomes are noisy or gameable, the more effective intervention is not to redesign prices but to improve monitoring and governance. Logging, traceability, evaluation harnesses, audits, and escalation rules can make hidden effort more observable and reduce room for metric gaming [26, 27]. For example, tool-call logs and intermediate artifacts can make hidden effort more visible, standardized evaluation protocols can reduce noise, and policy constraints can reduce manipulative behavior that inflates measured outcomes. These tools can substitute for stronger pay-for-performance by constraining the action space and making deviations easier to detect. At the same time, governance does not eliminate the role of pricing; rather, it changes what pricing can credibly do. Better monitoring and auditing expand the set of outcomes that can be contracted on, thereby making stronger outcome-dependent pricing feasible. In this sense, the feasible set of pricing contracts is jointly determined by payment design and the surrounding measurement infrastructure.

Alternative view 3: Market structure may favor token pricing even when it is not contractually efficient. A third view is that pricing reflects not only bilateral contract efficiency, but also broader market forces. Providers may prefer token-based pricing because it is simple, standardized, and easy to scale across heterogeneous users. In concentrated markets, firms may also favor pricing forms that shift more risk to users, while stronger competition may encourage providers to experiment with outcome-dependent guarantees. These forces can sustain token-based pricing even when a more outcome-dependent contract would be welfare-improving in a bilateral benchmark. They also help explain why adoption may be slow, uneven, and shaped by bargaining power as much as by contractual efficiency. Such market-structure considerations lie outside our stylized models, but they may significantly affect the pace of transition toward outcome-dependent pricing in practice.

6 Related Work

Our paper is related to recent work that applies contract-theoretic and delegation perspectives to AI and ML. In delegated classification and text generation, recent papers design performance-based or statistical contracts to incentivize an AI agent under hidden action or hidden cost [28, 29]. Related ideas also appear in delegated data collection for decentralized learning, where contracts are used to handle uncertainty in model quality and performance [30]. These papers show that contract design is a useful lens for AI systems, but they focus on specific delegation problems rather than the welfare impact of AI pricing, which justifies demand for outcome-based pricing.

A second related strand studies failures of token-based pricing and dishonest LLM providers. Recent work audits pay-per-token schemes and shows that token-based billing can create incentives for providers to misreport or manipulate usage [31]. Other work studies dishonest LLM providers

through a game-theoretic lens and designs mechanisms against model substitution or degraded service [32]. These papers are closely related to our emphasis on hidden actions and strategic manipulation.

Relative to this literature, our contribution is a broader position on AI-agent pricing. We compare token-based, outcome-only, and hybrid pricing through the PA lens, and emphasize a factor largely absent from prior AI/ML contract work: *risk sharing*. Our argument is that outcome-dependent pricing matters not only for incentive alignment under hidden effort, but also for reallocating uncertainty from users to providers better able to diversify it. We then connect this to hidden quality effort and imperfect evaluation, which together motivate hybrid contracts as a practical target.

References

[1] T. Schick, J. Dwivedi-Yu, R. Dessì, R. Raileanu, M. Lomeli, E. Hambro, L. Zettlemoyer, N. Cancedda, and T. Scialom, "Toolformer: Language models can teach themselves to use tools," *Advances in Neural Information Processing Systems*, vol. 36, pp. 68539–68551, 2023.

[2] S. G. Patil, T. Zhang, X. Wang, and J. E. Gonzalez, "Gorilla: Large language model connected with massive APIs," *Advances in Neural Information Processing Systems*, vol. 37, pp. 126544–126565, 2024.

[3] J. Yang, C. E. Jimenez, A. Wettig, K. Lieret, S. Yao, K. Narasimhan, and O. Press, "Swe-agent: Agent-computer interfaces enable automated software engineering," *Advances in Neural Information Processing Systems*, vol. 37, pp. 50528–50652, 2024.

[4] L. Qin, W. Pan, Q. Chen, L. Liao, Z. Yu, Y. Zhang, W. Che, and M. Li, "End-to-end task-oriented dialogue: A survey of tasks, methods, and future directions," *arXiv preprint arXiv:2311.09008*, 2023.

[5] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. R. Narasimhan, and Y. Cao, "React: Synergizing reasoning and acting in language models," in *The eleventh international conference on learning representations*, 2022.

[6] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *Advances in neural information processing systems*, vol. 36, pp. 11809–11822, 2023.

[7] D. Bergemann, A. Bonatti, and A. Smolin, "The economics of large language models: Token allocation, fine-tuning, and optimal pricing," *arXiv preprint arXiv:2502.07736*, 2025.

[8] E. Klang, D. Apakama, E. E. Abbott, A. Vaid, J. Lampert, A. Sakhuja, R. Freeman, A. W. Charney, D. Reich, M. Kraft, *et al.*, "A strategy for cost-effective large language model use at health system-scale," *NPJ digital medicine*, vol. 7, no. 1, p. 320, 2024.

[9] B. Li and S. Kumar, "Managing software-as-a-service: Pricing and operations," *Production and operations management*, vol. 31, no. 6, pp. 2588–2608, 2022.

[10] I. Makarov, J. da Costa, and B. Pinero, "Ai is driving a shift towards outcome-based pricing (december 2024 enterprise newsletter)." Andreessen Horowitz (a16z), The Enterprise Newsletter, Dec. 2024. Accessed: 2026-01-27.

[11] A. Blaylock, "Saas: how outcome-based pricing affects revenue recognition." Ernst & Young (EY) Insights, June 2025. Accessed: 2026-01-27.

[12] A. Mas-Colell, M. D. Whinston, J. R. Green, *et al.*, *Microeconomic theory*, vol. 1. Oxford university press New York, 1995.

[13] S. Shavell, "Risk sharing and incentives in the principal and agent relationship," *The Bell Journal of Economics*, pp. 55–73, 1979.

[14] S. J. Grossman and O. D. Hart, "An analysis of the principal-agent problem," in *Foundations of insurance economics: Readings in economics and finance*, pp. 302–340, Springer, 1992.

[15] B. Holmström, "Moral hazard and observability," *The Bell journal of economics*, pp. 74–91, 1979.

[16] B. Holmstrom and P. Milgrom, "Aggregation and linearity in the provision of intertemporal incentives," *Econometrica: Journal of the Econometric Society*, pp. 303–328, 1987.

[17] W. Cai, T. Shi, X. Zhao, and D. Song, "Are you getting what you pay for? auditing model substitution in llm apis," *arXiv preprint arXiv:2504.04715*, 2025.

[18] G. Sun, Z. Wang, X. Zhao, B. Tian, Z. Shen, Y. He, J. Xing, and A. Li, "Invisible tokens, visible bills: The urgent need to audit hidden operations in opaque llm services," *arXiv preprint arXiv:2505.18471*, 2025.

[19] Y. Cao, Y. Wang, S. Liu, M. Li, Y. Tao, and T. He, "Pay for the second-best service: A game-theoretic approach against dishonest llm providers," *arXiv preprint arXiv:2511.00847*, 2025.

[20] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in neural information processing systems*, vol. 35, pp. 27730–27744, 2022.

[21] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, *et al.*, "Judging llm-as-a-judge with mt-bench and chatbot arena," *Advances in neural information processing systems*, vol. 36, pp. 46595–46623, 2023.

[22] J. Lerner and J. Wulf, "Innovation and incentives: Evidence from corporate r&d," *the Review of Economics and Statistics*, vol. 89, no. 4, pp. 634–644, 2007.

[23] Y. Ma and S. Yang, "Technology driven market concentration through idea allocation," *Available at SSRN 4978854*, 2024.

[24] G. Baker, R. Gibbons, and K. J. Murphy, "Relational contracts and the theory of the firm," *The Quarterly Journal of Economics*, vol. 117, no. 1, pp. 39–84, 2002.

[25] J. Levin, "Relational incentive contracts," *American Economic Review*, vol. 93, no. 3, pp. 835–857, 2003.

[26] E. Tabassi, "Artificial intelligence risk management framework (ai rmf 1.0)," NIST AI 100-1, National Institute of Standards and Technology, Gaithersburg, MD, Jan. 2023.

[27] European Union, "Regulation (eu) 2024/1689 of the european parliament and of the council of 13 march 2024 laying down harmonised rules on artificial intelligence and amending certain union legislative acts (artificial intelligence act)," 2024.

[28] E. Saig, I. Talgam-Cohen, and N. Rosenfeld, "Delegated classification," *Advances in Neural Information Processing Systems*, vol. 36, pp. 13200–13236, 2023.

[29] E. Saig, O. Einav, and I. Talgam-Cohen, "Incentivizing quality text generation via statistical contracts," *Advances in Neural Information Processing Systems*, vol. 37, pp. 51196–51222, 2024.

[30] N. Ananthakrishnan, S. Bates, M. Jordan, and N. Haghtalab, "Delegating data collection in decentralized machine learning," in *International Conference on Artificial Intelligence and Statistics*, pp. 478–486, PMLR, 2024.

[31] A. A. Velasco, S. Tsirtsis, and M. Gomez-Rodriguez, "Auditing pay-per-token in large language models," *arXiv preprint arXiv:2510.05181*, 2025.

[32] Y. Cao, Y. Wang, S. Liu, M. Li, Y. Tao, and T. He, "Pay for the second-best service: A game-theoretic approach against dishonest llm providers," in *Proceedings of the ACM Web Conference 2026*, pp. 249–260, 2026.

A Limitations

We use stylized PA models to isolate how *risk sharing*, *hidden effort*, and *evaluation frictions* shape AI-agent pricing, and to advocate for outcome-based pricing. However, the current models are limited from the following perspectives:

(1) No externalities and no labor-market effects. As stated in the introduction, we abstract from misuse, spillovers across users, and broader social harms, and we do not study labor-market substitution/complementarity. These forces can dominate welfare conclusions and can justify regulatory or governance constraints that alter the feasible contract set.

(2) Linear contracts and CARA-Normal structure. We restrict attention to affine contracts under CARA-Normal assumptions for tractability and closed forms. Real pricing is nonlinear (volume discounts, caps, two-part tariffs, quotas, tiered plans) and may feature state dependence (e.g., higher charges under peak load). Nonlinear contracts could strengthen (or weaken) the case for hybrid designs, especially when there is user heterogeneity or limited liability.

(3) Non-Dynamics models. We treat "effort" as contemporaneous. In agent markets, effort includes *investment* (model upgrades, tool integrations, safety tuning) with dynamic spillovers across customers. Outcome-based pay can shape what gets improved and what is neglected over time. A dynamic model could justify long-run contracts (subscriptions with performance clauses) even when short-run outcome pay is noisy.

(4) Lack of empirical validation. Our analysis is theoretical and position-oriented, and we do not provide empirical evidence showing when real AI-agent markets satisfy the assumptions of our benchmark models or how large the predicted welfare gains from outcome-dependent pricing would be in practice. As a result, the paper should be read as identifying economic mechanisms and directional implications rather than as delivering quantitatively calibrated recommendations. Empirical work on real pricing data, deployment outcomes, risk exposure, and the feasibility of evaluation pipelines would be needed to assess when the proposed contract forms are most appropriate in practice.

B CARA–Normal Principal–Agent Benchmarks

B.1 Model Primitives and Certainty Equivalents

A principal contracts with an agent who chooses effort $e \in \mathcal{E}$ and incurs cost $c(e)$. The realized outcome π is a noisy function of effort:

$$\pi \mid e \sim \mathcal{N}\big(\mu(e), \sigma^2\big), \tag{14}$$

where σ^2 is independent of e, and $\mu(e)$ is increasing in e. A wage contract is denoted $w(\pi)$ when effort is not contractable, and $w(\pi, e)$ when effort is observable and contractable.

CARA utilities and certainty equivalents. When an agent (or principal) has CARA utility

$$u(x) = -\exp(-rx), \qquad r > 0,$$

and X is normally distributed, maximizing expected utility $\mathbb{E}[u(X)]$ is equivalent to maximizing the certainty equivalent

$$\mathrm{CE}(X) = \mathbb{E}[X] - \frac{r}{2}\mathrm{Var}(X). \tag{15}$$

Throughout, $\bar{u}$ denotes the counterparty's reservation utility expressed in the appropriate metric: *certainty-equivalent utility* under CARA preferences, and *expected net payoff* under risk neutrality.

B.2 Benchmark 0: Risk-neutral Principal and Risk-averse Agent

In this benchmark, the principal is *risk-neutral* and the agent is *risk-averse* with CARA utility

$$u_a(x) = -\exp(-r_a x), \qquad r_a > 0.$$

The agent's net payoff is $w(\pi, e) - c(e)$ (or $w(\pi) - c(e)$ when effort is not contractable). The agent participates only if

$$\mathrm{CE}_a(w(\pi, e) - c(e)) \geq \bar{u}, \tag{16}$$

where CE_a is defined in (15) with $r = r_a$.

Case I: Observable effort (first-best). If effort is observable and contractable, the principal directly chooses effort and fully insures the agent. The first-best effort solves

$$e^{\mathrm{FB}} \in \arg\max_{e} \ \mu(e) - c(e). \tag{17}$$

A first-best contract is a fixed wage contingent on the prescribed action:

$$w^{\mathrm{FB}}(\pi, e) = \begin{cases} \bar{w}, & \text{if } e = e^{\mathrm{FB}}, \\ 0, & \text{otherwise,} \end{cases}$$

where $\bar{w}$ makes (16) bind. Since the wage is deterministic, the binding IR condition implies

$$\bar{w} - c\big(e^{\mathrm{FB}}\big) = \bar{u} \qquad \Rightarrow \qquad \bar{w} = \bar{u} + c\big(e^{\mathrm{FB}}\big)\,.$$

Case II: Unobservable effort (moral hazard). If effort is unobservable, the wage can depend only on π. The principal solves

$$\max_{w(\cdot)} \ \mathbb{E}[\pi - w(\pi)] \quad \text{s.t.} \quad \text{(IR) } \mathrm{CE}_a(w(\pi) - c(e)) \geq \bar{u}, \tag{18}$$

$$\text{(IC) } e \in \arg\max_{e'} \ \mathrm{CE}_a(w(\pi) - c(e'))\,. \tag{19}$$

Under CARA–Normal assumptions, it is without loss of generality to restrict attention to affine contracts

$$w(\pi) = \alpha + \beta\pi. \tag{20}$$

Conditional on effort e, the agent's certainty equivalent is

$$\mathrm{CE}_a(e; \alpha, \beta) = \alpha + \beta\mu(e) - c(e) - \frac{r_a}{2}\beta^2\sigma^2. \tag{21}$$

Hence the IC condition reduces to

$$e \in \arg\max_{e'} \ \beta\mu(e') - c(e'), \tag{22}$$

and the IR condition becomes

$$\alpha + \beta\mu(e) - c(e) - \frac{r_a}{2}\beta^2\sigma^2 \geq \bar{u}. \tag{23}$$

B.2.1 Linear-quadratic closed form

Example: Linear-quadratic Cost. Consider the common linear-quadratic specialization $\mu(e) = e$ and $c(e) = \frac{k}{2}e^2$.

Observable effort (first-best). The principal chooses

$$e^{\mathrm{FB}} = \arg\max_{e} \left\{ e - \frac{k}{2}e^2 \right\} = \frac{1}{k}.$$

The constant wage $\bar{w}$ binds IR:

$$\bar{w} = \bar{u} + c\big(e^{\mathrm{FB}}\big) = \bar{u} + \frac{1}{2k}.$$

Since the principal is risk-neutral, its certainty equivalent equals expected profit:

$$\mathrm{CE}_p^{\mathrm{FB}} = \mathbb{E}[\pi - w^{\mathrm{FB}}] = \mu\big(e^{\mathrm{FB}}\big) - \bar{w} = \frac{1}{k} - \left(\bar{u} + \frac{1}{2k}\right) = \frac{1}{2k} - \bar{u}.$$

Unobservable effort (second-best). Under the affine contract $w(\pi) = \alpha + \beta\pi$, the agent's best response from (22) is

$$e(\beta) = \frac{\beta}{k}.$$

Imposing IC and setting IR to bind yields the optimal incentive intensity

$$\beta^{\mathrm{SB}} = \frac{1}{1 + kr_a\sigma^2}, \qquad e^{\mathrm{SB}} = \frac{\beta^{\mathrm{SB}}}{k} = \frac{1}{k(1 + kr_a\sigma^2)} < e^{\mathrm{FB}}.$$

The fixed payment is

$$\alpha^{\mathrm{SB}} = \bar{u} - \frac{\left(\beta^{\mathrm{SB}}\right)^2}{2k} + \frac{r_a}{2}\left(\beta^{\mathrm{SB}}\right)^2\sigma^2,$$

so the optimal contract is

$$w^{\mathrm{SB}}(\pi) = \alpha^{\mathrm{SB}} + \beta^{\mathrm{SB}}\pi.$$

The principal's certainty equivalent (equal to expected profit) is

$$\mathrm{CE}_p^{\mathrm{SB}} = \frac{\beta^{\mathrm{SB}}}{k} - \frac{\left(\beta^{\mathrm{SB}}\right)^2}{2k} - \bar{u} - \frac{r_a}{2}\left(\beta^{\mathrm{SB}}\right)^2\sigma^2 = \frac{1}{2k(1+kr_a\sigma^2)} - \bar{u}.$$

Comparison (insurance–incentives trade-off). With observable effort, the principal fully insures the risk-averse agent and implements the first-best effort e^{FB}. With moral hazard, incentives require outcome-contingent pay; the optimal incentive intensity satisfies $\beta^{\mathrm{SB}} < 1$ and induces $e^{\mathrm{SB}} < e^{\mathrm{FB}}$. Increasing β strengthens incentives but increases the agent's risk exposure, which must be compensated via a higher expected payment.

C Outcome-Based Contracting Is Needed When the Principal Is Risk-Averse

C.1 Benchmark 1: Risk-averse Principal and Risk-neutral Agent

In this benchmark, the principal has CARA utility over profit $\pi - w(\pi)$ with risk aversion parameter $r_p > 0$:

$$u_p(x) = -\exp(-r_p x),$$

so for normally distributed profit the principal maximizes

$$\mathrm{CE}_p(\pi - w(\pi)) = \mathbb{E}[\pi - w(\pi)] - \frac{r_p}{2}\mathrm{Var}(\pi - w(\pi)).$$

The agent is risk-neutral and chooses effort to maximize expected net payoff $\mathbb{E}[w(\pi)] - c(e)$, with reservation utility $\bar{u}$.

Case I: Observable effort (first-best). If effort is contractable, the principal solves

$$\max_{e,\, w(\cdot)} \ \mathrm{CE}_p(\pi - w(\pi)) \quad \text{s.t.} \quad \mathbb{E}[w(\pi)] - c(e) \geq \bar{u}. \tag{24}$$

With a risk-neutral agent, the principal can shift all outcome risk to the agent. A convenient implementation is the residual-claimant contract

$$w(\pi) = \alpha + \pi, \tag{25}$$

under which principal profit is constant: $\pi - w(\pi) = -\alpha$, so the principal is fully insured. The participation constraint binds:

$$\alpha + \mathbb{E}[\pi] - c(e) = \bar{u} \qquad \Rightarrow \qquad \alpha = \bar{u} + c(e) - \mu(e).$$

The principal then chooses the first-best effort by maximizing total surplus:

$$e^{\mathrm{FB}} \in \arg\max_e \ \mu(e) - c(e).$$

Case II: Unobservable effort. If effort is unobservable, the contract depends only on π and the principal solves

$$\max_{w(\cdot)} \ \mathrm{CE}_p(\pi - w(\pi)) \quad \text{s.t.} \quad \text{(IC) } e \in \arg\max_{e'} \ \mathbb{E}[w(\pi)] - c(e'), \tag{26}$$

$$\text{(IR) } \mathbb{E}[w(\pi)] - c(e) \geq \bar{u}. \tag{27}$$

Restricting to affine contracts $w(\pi) = \alpha + \beta\pi$, the agent's IC depends only on the mean:

$$e \in \arg\max_{e'} \ \beta\mu(e') - c(e'). \tag{28}$$

Choosing $\beta = 1$ aligns the agent's objective with the total surplus maximization $\max_e\{\mu(e) - c(e)\}$. Moreover, when $\beta = 1$ the principal's profit becomes constant, $\pi - w(\pi) = -\alpha$, eliminating all risk for the principal. Finally, α is set so that IR binds:

$$\alpha = \bar{u} + c(e) - \mu(e).$$

Thus, absent additional constraints (e.g., limited liability), the principal can implement the first-best effort even when effort is unobservable by making the agent the residual claimant.

C.2 Benchmark 2: Principal Is More Risk-Averse

We consider a CARA–Normal principal–agent environment in which *both* parties are risk-averse. The principal has CARA utility over profit $\pi - w(\pi)$ with absolute risk aversion $r_p > 0$,

$$u_p(x) = -\exp(-r_p x), \tag{29}$$

and the agent has CARA utility over net payoff $w(\pi) - c(e)$ with absolute risk aversion $r_a > 0$,

$$u_a(x) = -\exp(-r_a x). \tag{30}$$

The outcome satisfies $\pi \mid e \sim \mathcal{N}(\mu(e), \sigma^2)$, where σ^2 is independent of e and $\mu(e)$ is weakly increasing. Let $\bar{u}$ denote the agent's reservation utility in certainty-equivalent terms.

Throughout, we use certainty equivalents under CARA–Normality. For normally distributed X,

$$\mathrm{CE}(X) = \mathbb{E}[X] - \frac{r}{2}\mathrm{Var}(X), \tag{31}$$

so the principal and agent maximize $\mathrm{CE}_p(\cdot)$ and $\mathrm{CE}_a(\cdot)$ with parameters r_p and r_a, respectively.

Case I: Observable effort (first-best). If effort is contractable, the principal chooses $(e, w(\cdot))$ to maximize the principal's certainty equivalent subject to the agent's participation constraint:

$$\begin{aligned} &\max_{e,\, w(\cdot)} \ \mathrm{CE}_p\big(\pi - w(\pi)\big) \\ &\quad \text{s.t. } \mathrm{CE}_a\big(w(\pi) - c(e)\big) \geq \bar{u}. \end{aligned} \tag{32}$$

Under CARA–Normality it is without loss to restrict attention to affine risk-sharing contracts

$$w(\pi) = \alpha + \beta\pi. \tag{33}$$

Given effort e, the agent's participation constraint binds at the optimum and can be written as

$$\alpha + \beta\mu(e) - c(e) - \frac{r_a}{2}\beta^2\sigma^2 = \bar{u}, \tag{34}$$

which pins down α as a function of (e, β). Substituting (34) into the principal's certainty equivalent yields the reduced-form first-best problem

$$\max_{e,\beta} \ \mu(e) - c(e) - \frac{r_a}{2}\beta^2\sigma^2 - \frac{r_p}{2}(1-\beta)^2\sigma^2. \tag{35}$$

The choice separates. The optimal risk-sharing slope is

$$\beta^{\mathrm{FB}} = \arg\min_{\beta} \left\{ \frac{r_a}{2}\beta^2\sigma^2 + \frac{r_p}{2}(1-\beta)^2\sigma^2 \right\} = \frac{r_p}{r_a + r_p}, \tag{36}$$

and effort maximizes total surplus,

$$e^{\mathrm{FB}} \in \arg\max_{e} \ \mu(e) - c(e). \tag{37}$$

Finally, α^{FB} is pinned down by (34) evaluated at $(e^{\mathrm{FB}}, \beta^{\mathrm{FB}})$.

Case II: Unobservable effort (second-best). If effort is not observable, the contract can depend only on π. The principal solves

$$\begin{aligned} &\max_{w(\cdot)} \ \mathrm{CE}_p\big(\pi - w(\pi)\big) \\ &\quad \text{s.t. (IR) } \mathrm{CE}_a\big(w(\pi) - c(e)\big) \geq \bar{u}, \\ &\qquad \text{(IC) } e \in \arg\max_{e'} \mathrm{CE}_a\big(w(\pi) - c(e')\big). \end{aligned} \tag{38}$$

Under CARA–Normality it is without loss to consider affine contracts (33). Under (α, β), the agent's certainty equivalent is

$$\mathrm{CE}_a(e; \alpha, \beta) = \alpha + \beta\mu(e) - c(e) - \frac{r_a}{2}\beta^2\sigma^2, \tag{39}$$

so incentive compatibility reduces to

$$e(\beta) \in \arg\max_{e'} \ \beta\mu(e') - c(e'). \tag{40}$$

Given the induced effort $e(\beta)$, the principal's certainty equivalent is

$$\mathrm{CE}_p(e;\alpha,\beta) = (1-\beta)\mu(e) - \alpha - \frac{r_p}{2}(1-\beta)^2\sigma^2. \tag{41}$$

At the optimum, the participation constraint typically binds in certainty-equivalent terms,

$$\alpha + \beta\mu(e(\beta)) - c(e(\beta)) - \frac{r_a}{2}\beta^2\sigma^2 = \bar{u}, \tag{42}$$

which eliminates α and reduces the problem to choosing β (and the induced effort $e(\beta)$) to balance incentives and risk sharing.

C.2.1 Canonical linear-quadratic closed form

We compare the two cases under the linear-quadratic specialization $\mu(e) = e$ and $c(e) = \frac{k}{2}e^2$.

Observable effort (first-best). Effort maximizes total surplus:

$$e^{\mathrm{FB}} = \arg\max_e \left(e - \frac{k}{2}e^2\right) = \frac{1}{k}. \tag{43}$$

For any fixed effort, the optimal risk-sharing slope is given by (36),

$$\beta^{\mathrm{FB}} = \frac{r_p}{r_a + r_p}, \tag{44}$$

and the fixed payment α^{FB} is determined by the binding participation constraint (34).

Unobservable effort (second-best). The agent's best response (40) implies

$$e(\beta) = \arg\max_e \left(\beta e - \frac{k}{2}e^2\right) = \frac{\beta}{k}. \tag{45}$$

Using the binding participation constraint to eliminate α, the principal chooses β to trade off incentives and risk-sharing. The resulting second-best slope and effort are

$$\beta^{\mathrm{SB}} = \frac{1 + k\sigma^2 r_p}{1 + k\sigma^2(r_a + r_p)}, \qquad e^{\mathrm{SB}} = \frac{\beta^{\mathrm{SB}}}{k}. \tag{46}$$

Finally, α^{SB} is pinned down by the binding participation constraint:

$$\alpha^{\mathrm{SB}} = \bar{u} - \beta^{\mathrm{SB}}e^{\mathrm{SB}} + \frac{k}{2}(e^{\mathrm{SB}})^2 + \frac{r_a}{2}(\beta^{\mathrm{SB}})^2\sigma^2. \tag{47}$$

Direct comparison. With observable effort, the first-best effort is independent of risk parameters:

$$e^{\mathrm{FB}} = \frac{1}{k}. \tag{48}$$

Under moral hazard, $\beta^{\mathrm{SB}} < 1$ and therefore

$$e^{\mathrm{SB}} = \frac{\beta^{\mathrm{SB}}}{k} < \frac{1}{k} = e^{\mathrm{FB}}, \tag{49}$$

so effort is distorted downward. Moreover, the second-best slope exceeds the pure risk-sharing rule:

$$\beta^{\mathrm{SB}} - \beta^{\mathrm{FB}} = \frac{1}{1 + k\sigma^2(r_a + r_p)} \ > \ 0, \tag{50}$$

showing that incentive provision under unobservable effort requires *stronger* outcome-contingent pay relative to the first-best risk-sharing benchmark.

D Model Details for Hidden Efforts

We consider a CARA–normal principal–agent model with two effort dimensions, the outcome is given by

$$\pi = \theta^\top e + \epsilon, \tag{51}$$

where $e = (e_1, e_2)^\top$, $\theta = (\theta_1, \theta_2)^\top$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. The agent incurs a quadratic effort cost

$$c(e) = \frac{1}{2} e^\top K e, \tag{52}$$

where K is symmetric positive definite, which can be written as

$$K = \begin{pmatrix} k_{11} & k_{12} \\ k_{12} & k_{22} \end{pmatrix}. \tag{53}$$

The agent and the principal have CARA preferences with risk aversion parameters $r_a > 0$ and $r_p > 0$. Let $\bar{u}$ denote the agent's reservation utility. For later use, define the scalars

$$A \triangleq \theta^\top K^{-1} \theta, \qquad B \triangleq \frac{\theta_2^2}{k_{22}}, \qquad \Delta \triangleq k_{11} k_{22} - k_{12}^2. \tag{54}$$

The following theorems characterize the optimal contract in this setup across the four cases.

Theorem D.1 (Case (0): Both efforts observable)**.** *Suppose both effort components are enforceable, and the contract is linear $w(e, \pi) = \alpha + \beta\pi$. The first-best effort chosen by the principal and the optimal incentive intensity are given by*

$$e^{(0)} = K^{-1}\theta, \qquad \beta^{(0)} = \frac{r_p}{r_a + r_p},$$

and the principal's optimal certainty equivalent is

$$\mathrm{CE}_p^{(0)} = \frac{1}{2} A - \frac{\sigma^2}{2} r_a \beta^{(0)} - \bar{u}.$$

Because both the principal and the agent are risk-averse, achieving the optimal arrangement requires an outcome-based contract to share risk efficiently.

Theorem D.2 (Case (1): only e_1 observable, no outcome-based pay)**.** *Suppose the principal can enforce e_1 but the contract cannot depend on π, i.e., $w = \alpha$ contingent only on e_1. The agent's best response in the hidden effort is*

$$e_2^{(1)}(e_1) = -\frac{k_{12}}{k_{22}} e_1,$$

the principal enforces

$$e_1^{(1)} = \frac{k_{22}\theta_1 - k_{12}\theta_2}{\Delta}, \qquad e^{(1)} = (e_1^{(1)}, e_2^{(1)})^\top,$$

and the principal's optimal certainty equivalent is

$$\mathrm{CE}_p^{(1)} = \frac{1}{2} A - \frac{\theta_2^2}{2k_{22}} - \frac{\sigma^2}{2} r_p - \bar{u}.$$

Since only e_1 is contractable, the agent will choose the unobservable effort e_2 to minimize its own cost. For example, if $k_{12} > 0$, the two efforts are substitutes, the agent has an incentive to rely heavily on token usage (high e_1) while selecting a lower-quality model (low e_2). The resulting surplus is strictly smaller compared to case (0) because e_2 is low, and the risk is borne entirely by the principal.

Theorem D.3 (Case (2): Outcome-only contract)**.** *Suppose neither effort component is contractable and the principal offers a linear outcome-based contract $w(\pi) = \alpha + \beta\pi$. The agent then chooses the following effort, and the optimal incentive intensity is given by*

$$e^{(2)}(\beta) = \beta K^{-1}\theta, \qquad \beta^{(2)} = \frac{A + r_p\sigma^2}{A + (r_a + r_p)\sigma^2},$$

and the principal's optimal certainty equivalent is

$$\mathrm{CE}_p^{(2)} = \frac{1}{2} A - \frac{\sigma^2}{2} r_a \, \beta^{(2)} - \bar{u}.$$

Table 2: Ranking comparison across four contracting cases. For each column, rank 1 denotes the highest outcome in that category.

Case	Setting	β rank	e_1 rank	e_2 rank	CE_p rank	sum of CE
(0)	e_1, e_2, π contractable;	3	1	1	1	1
(1)	e_1 contractable only;	4	1	N/A	4	4
(2)	π contractable only;	1	2	2	3	3
(3)	e_1, π contractable;	2	1	N/A	2	2

In this case, a larger $\beta^{(2)}$ is required to incentivize the agent to put more effort compared with case (0). This highlights a moral hazard problem: because all efforts are hidden, the agent has an incentive to underprovide it unless the contract places sufficient weight on performance outcomes.

Theorem D.4 (Case (3): Hybrid pricing)**.** *Suppose the principal can enforce e_1 and offers an outcome-based linear contract $w(e_1, \pi) = \alpha + \beta\pi$, while e_2 remains hidden. Then, given (e_1, β) the agent chooses*

$$e_2(e_1, \beta) = \frac{\beta\theta_2 - k_{12}e_1}{k_{22}}.$$

The principal enforces the same observable effort e_1 as in case (1), and chooses the optimal incentive intensity

$$e_1^{(3)} = \frac{k_{22}\theta_1 - k_{12}\theta_2}{\Delta}, \qquad \beta^{(3)} = \frac{B + r_p\sigma^2}{B + (r_a + r_p)\sigma^2},$$

and the principal's optimal certainty equivalent is

$$\mathrm{CE}_p^{(3)} = \frac{1}{2}A - \frac{\sigma^2}{2}r_a\,\beta^{(3)} - \bar{u}.$$

Comparing optimal contracts across the four cases yields the following results. Table 2 summarizes the overall ranking.

Proposition D.5 (Comparison across the four cases)**.** *Under the assumptions of linear-quadratic CARA-Normal, the following comparisons hold.*

(i) Incentive intensity. *The optimal β satisfy*

$$\beta^{(1)} = 0 \;<\; \beta^{(0)} \;<\; \beta^{(3)} \;\leq\; \beta^{(2)}.$$

(ii) Principal welfare. *If $r_a \leq r_p$, then the principal's optimal CE_p satisfy*

$$\mathrm{CE}_p^{(0)} \;\geq\; \mathrm{CE}_p^{(3)} \;\geq\; \mathrm{CE}_p^{(2)} \geq \mathrm{CE}_p^{(1)}.$$

(iii) Observable effort. *The effort levels e_1 satisfy*

$$e_1^{(0)} = e_1^{(1)} = e_1^{(3)} = \frac{k_{22}\theta_1 - k_{12}\theta_2}{\Delta}, \qquad e_1^{(2)} = \beta^{(2)}e_1^{(0)}.$$

(iv) Hidden effort. *The hidden effort e_2 satisfies*

$$e_2^{(2)} = \beta^{(2)}e_2^{(0)}, \qquad e_2^{(3)} = e_2^{(1)} + \frac{\beta^{(3)}\theta_2}{k_{22}}.$$

A general level ordering between $e_2^{(3)}$ and $e_2^{(2)}$ does not hold without further restrictions on (θ, K).

D.1 Case (0): Observable efforts

Suppose both effort components are observable and enforceable. Without loss of generalizability, we consider the following linear contracts

$$w = \alpha + \beta\pi. \tag{55}$$

Agent's certainty equivalent. The agent's net payoff is

$$x_a = \alpha + \beta(\theta^\top e + \epsilon) - \frac{1}{2} e^\top K e, \tag{56}$$

so the CARA–normal certainty equivalent is

$$\mathrm{CE}_a = \alpha + \beta\theta^\top e - \frac{1}{2} e^\top K e - \frac{r_a}{2}\beta^2\sigma^2. \tag{57}$$

Participation constraint and optimal α. The participation constraint $\mathrm{CE}_a \geq \bar{u}$ binds at the optimum, so

$$\alpha^{(0)}(e, \beta) = \bar{u} - \beta\theta^\top e + \frac{1}{2} e^\top K e + \frac{r_a}{2}\beta^2\sigma^2. \tag{58}$$

Principal's certainty equivalent. The principal's net payoff is

$$x_p = \pi - w = (1 - \beta)(\theta^\top e + \epsilon) - \alpha, \tag{59}$$

and the certainty equivalent is

$$\mathrm{CE}_p = (1 - \beta)\theta^\top e - \alpha - \frac{r_p}{2}(1 - \beta)^2\sigma^2. \tag{60}$$

Substituting $\alpha^{(0)}(e, \beta)$ yields

$$\mathrm{CE}_p(e, \beta) = \theta^\top e - \frac{1}{2} e^\top K e - \bar{u} - \frac{\sigma^2}{2}\Big(r_a\beta^2 + r_p(1 - \beta)^2\Big). \tag{61}$$

Optimal effort and incentive intensity. Since the risk-sharing term is independent of e, the principal chooses e to maximize $\theta^\top e - \frac{1}{2} e^\top K e$, yielding

$$e^{(0)} = K^{-1}\theta. \tag{62}$$

Given $e^{(0)}$, the principal chooses β to minimize $r_a\beta^2 + r_p(1 - \beta)^2$, so

$$\beta^{(0)} = \frac{r_p}{r_a + r_p}. \tag{63}$$

Principal's certainty equivalent at the optimum. Evaluating the deterministic surplus at $e^{(0)}$ gives

$$\theta^\top e^{(0)} - \frac{1}{2}(e^{(0)})^\top K e^{(0)} = \frac{1}{2} A, \tag{64}$$

and

$$r_a(\beta^{(0)})^2 + r_p(1 - \beta^{(0)})^2 = \frac{r_a r_p}{r_a + r_p}. \tag{65}$$

Therefore,

$$\mathrm{CE}_p^{(0)} = \frac{1}{2} A - \frac{\sigma^2}{2}\frac{r_a r_p}{r_a + r_p} - \bar{u} = \frac{1}{2} A - \frac{r_a\sigma^2}{2}\beta^{(0)} - \bar{u}. \tag{66}$$

D.2 Case (1): e_1 observable, e_2 hidden, no outcome-based pay

In this variant, the contract can depend on the observable effort e_1 but not on output π. We restrict attention to contracts of the form

$$w = \alpha \quad \text{if } e_1 = e_1^{(1)}, \tag{67}$$

where $(\alpha, e_1^{(1)})$ are chosen by the principal. The principal enforces e_1, while the agent privately chooses e_2.

Agent's certainty equivalent and best response in e_2. Because the wage is independent of π, the agent faces no risk:

$$\mathrm{CE}_a = \alpha - \frac{1}{2} e^\top K e. \tag{68}$$

For any enforced e_1, the agent chooses e_2 to maximize CE_a. The first-order condition is

$$\frac{\partial \mathrm{CE}_a}{\partial e_2} = -(k_{12}e_1 + k_{22}e_2) = 0, \tag{69}$$

so

$$e_2^{(1)}(e_1) = -\frac{k_{12}}{k_{22}}\, e_1. \tag{70}$$

Participation constraint and optimal α. The participation constraint binds $\mathrm{CE}_a \geq \bar{u}$, so

$$\alpha^{(1)}(e_1) = \bar{u} + \frac{1}{2} e^\top K e, \qquad e = (e_1, e_2^{(1)}(e_1))^\top. \tag{71}$$

Substituting $e_2^{(1)}(e_1) = -(k_{12}/k_{22})e_1$ yields

$$e^\top K e = \left(k_{11} - \frac{k_{12}^2}{k_{22}}\right) e_1^2, \tag{72}$$

and hence

$$\alpha^{(1)}(e_1) = \bar{u} + \frac{1}{2}\left(k_{11} - \frac{k_{12}^2}{k_{22}}\right) e_1^2. \tag{73}$$

Principal's certainty equivalent and optimal enforced effort $e_1^{(1)}$. The principal's net payoff is

$$x_p = \pi - w = \theta^\top e + \epsilon - \alpha, \tag{74}$$

and the principal's certainty equivalent is

$$\mathrm{CE}_p = \theta^\top e - \alpha - \frac{r_p}{2}\sigma^2. \tag{75}$$

Using the binding participation constraint,

$$\mathrm{CE}_p(e_1) = \theta^\top e - \frac{1}{2} e^\top K e - \bar{u} - \frac{r_p}{2}\sigma^2, \qquad e = (e_1, e_2^{(1)}(e_1))^\top. \tag{76}$$

Substituting $e_2^{(1)}(e_1) = -(k_{12}/k_{22})e_1$ and (72), the principal chooses e_1 to maximize

$$\max_{e_1} \left\{ \left(\theta_1 - \frac{k_{12}}{k_{22}}\theta_2\right) e_1 - \frac{1}{2}\left(k_{11} - \frac{k_{12}^2}{k_{22}}\right) e_1^2 \right\}. \tag{77}$$

Therefore, the enforced effort in this case is

$$e_1^{(1)} = \frac{\theta_1 - \frac{k_{12}}{k_{22}}\theta_2}{k_{11} - \frac{k_{12}^2}{k_{22}}} = \frac{k_{22}\theta_1 - k_{12}\theta_2}{\Delta}, \tag{78}$$

and induced hidden effort is

$$e_2^{(1)} = e_2^{(1)}(e_1^{(1)}) = -\frac{k_{12}}{k_{22}}\, e_1^{(1)} = -\frac{k_{12}\theta_1 - \frac{k_{12}^2}{k_{22}}\theta_2}{\Delta}. \tag{79}$$

Principal's certainty equivalent at the optimum.

$$\mathrm{CE}_p^{(1)} = \theta^\top e^{(1)} - \frac{1}{2}(e^{(1)})^\top K e^{(1)} - \bar{u} - \frac{r_p}{2}\sigma^2, \qquad e^{(1)} = (e_1^{(1)}, e_2^{(1)})^\top. \tag{80}$$

Equivalently, substituting $e^{(1)}$ yeilds

$$\mathrm{CE}_p^{(1)} = \frac{1}{2}\frac{(k_{22}\theta_1 - k_{12}\theta_2)^2}{k_{22}\Delta} - \frac{r_p}{2}\sigma^2 - \bar{u} = \frac{1}{2}A - \frac{\theta_2^2}{2k_{22}} - \frac{r_p}{2}\sigma^2 - \bar{u}. \tag{81}$$

D.3 Case (2): Outcome-only contract

In this variant, the contract depends on output π but cannot depend on either effort component. We restrict attention to

$$w = \alpha + \beta\pi, \tag{82}$$

and the agent privately chooses e.

Agent's certainty equivalent and incentive compatibility. The agent's certainty equivalent is

$$\mathrm{CE}_a = \alpha + \beta\theta^\top e - \frac{1}{2} e^\top K e - \frac{r_a}{2}\beta^2\sigma^2. \tag{83}$$

Given (α, β), the agent chooses e to maximize $\beta\theta^\top e - \frac{1}{2}e^\top K e$, yielding

$$e^{(2)}(\beta) = \beta K^{-1}\theta. \tag{84}$$

Participation constraint and optimal α. Imposing $\mathrm{CE}_a \geq \bar{u}$ at $e^{(2)}(\beta)$ gives

$$\alpha^{(2)}(\beta) = \bar{u} - \frac{1}{2}\beta^2 A + \frac{r_a}{2}\beta^2\sigma^2. \tag{85}$$

Principal's certainty equivalent. The principal's net payoff is

$$x_p = \pi - w = (1-\beta)(\theta^\top e + \epsilon) - \alpha, \tag{86}$$

and the principal's certainty equivalent is

$$\mathrm{CE}_p = (1-\beta)\theta^\top e - \alpha - \frac{r_p}{2}(1-\beta)^2\sigma^2. \tag{87}$$

Substituting $e = e^{(2)}(\beta)$ and $\alpha = \alpha^{(2)}(\beta)$ into the principal's certainty equivalent yields

$$\mathrm{CE}_p^{(2)}(\beta) = \frac{1}{2}\beta(2-\beta)\,A - \bar{u} - \frac{\sigma^2}{2}\Big(r_a\beta^2 + r_p(1-\beta)^2\Big). \tag{88}$$

Optimal incentive intensity and value at the optimum. The first-order condition implies

$$\beta^{(2)} = \frac{A + r_p\sigma^2}{A + (r_a + r_p)\sigma^2}. \tag{89}$$

Plugging $\beta^{(2)}$ back yields the convenient decomposition

$$\mathrm{CE}_p^{(2)} = \frac{A^2 + Ar_p\sigma^2 - r_a r_p\sigma^4}{2\big(A + (r_a + r_p)\sigma^2\big)} - \bar{u} = \frac{1}{2}\,A - \frac{\sigma^2}{2}r_a\,\beta^{(2)} - \bar{u}. \tag{90}$$

In particular, the first term is exactly the first-best deterministic surplus component $\frac{A}{2}$, and the final term is the distortion induced by moral hazard and risk sharing under outcome-only contracting.

D.4 Case (3): e_1 observable, e_2 hidden, outcome-based pay

In this case, e_1 is observable and enforceable by the principal, while e_2 is privately chosen by the agent. The wage depends only on output:

$$w = \alpha + \beta\pi. \tag{91}$$

Agent's certainty equivalent and best response in e_2. Given enforced e_1 and contract (α, β), the agent's certainty equivalent is

$$\mathrm{CE}_a = \alpha + \beta\theta^\top e - \frac{1}{2}e^\top Ke - \frac{r_a}{2}\beta^2\sigma^2. \tag{92}$$

The first-order condition for e_2 is

$$\beta\theta_2 - (k_{12}e_1 + k_{22}e_2) = 0, \tag{93}$$

so

$$e_2(e_1, \beta) = \frac{\beta\theta_2 - k_{12}e_1}{k_{22}}. \tag{94}$$

Participation constraint and optimal α. The participation constraint $\mathrm{CE}_a \geq \bar{u}$ binds:

$$\alpha^{(3)}(\beta, e_1) = \bar{u} - \beta\theta^\top e + \frac{1}{2}e^\top Ke + \frac{r_a}{2}\beta^2\sigma^2, \qquad e = (e_1, e_2(e_1, \beta))^\top. \tag{95}$$

Principal's certainty equivalent and optimal enforced effort. Substituting $\alpha^{(3)}(\beta, e_1)$ into the principal's certainty equivalent yields

$$\mathrm{CE}_p^{(3)}(\beta, e_1) = \theta^\top e - \frac{1}{2}e^\top Ke - \bar{u} - \frac{\sigma^2}{2}\Big(r_a\beta^2 + r_p(1-\beta)^2\Big), \qquad e = (e_1, e_2(e_1, \beta))^\top. \tag{96}$$

Fixing β, the principal chooses e_1 to maximize the deterministic surplus $\theta^\top e - \frac{1}{2}e^\top Ke$. Differentiating with respect to e_1 yields

$$\left(k_{11} - \frac{k_{12}^2}{k_{22}}\right)e_1 = \theta_1 - \frac{k_{12}}{k_{22}}\theta_2, \tag{97}$$

so

$$e_1^{(3)} = \frac{k_{22}\theta_1 - k_{12}\theta_2}{\Delta}, \tag{98}$$

which is independent of β. The induced hidden effort is

$$e_2^{(3)}(\beta) = \frac{\beta\theta_2 - k_{12}e_1^{(3)}}{k_{22}}. \tag{99}$$

Reduced-form objective and optimal $\beta^{(3)}$. Evaluating the deterministic surplus at $e_1^{(3)}$ gives

$$\theta^\top e - \frac{1}{2} e^\top K e = \frac{(k_{22}\theta_1 - k_{12}\theta_2)^2}{2k_{22}\Delta} + \frac{(2\beta - \beta^2)\theta_2^2}{2k_{22}}. \tag{100}$$

Therefore,

$$\mathrm{CE}_p^{(3)}(\beta) = \frac{(k_{22}\theta_1 - k_{12}\theta_2)^2}{2k_{22}\Delta} + \frac{(2\beta - \beta^2)}{2} B - \bar{u} - \frac{\sigma^2}{2}\Big(r_a\beta^2 + r_p(1-\beta)^2\Big), \tag{101}$$

where $B \triangleq \frac{\theta_2^2}{k_{22}}$. The first-order condition is

$$B(1-\beta) = \sigma^2\Big((r_a + r_p)\beta - r_p\Big), \tag{102}$$

so

$$\beta^{(3)} = \frac{B + r_p\sigma^2}{B + (r_a + r_p)\sigma^2}. \tag{103}$$

Principal's certainty equivalent at the optimum. Substituting $\beta^{(3)}$ into Principal's certainty equivalent and simplifying yields

$$\mathrm{CE}_p^{(3)} = \frac{(k_{22}\theta_1 - k_{12}\theta_2)^2}{2k_{22}\Delta} + \frac{1}{2}\frac{\left(\frac{\theta_2^2}{k_{22}} + r_p\sigma^2\right)^2}{\frac{\theta_2^2}{k_{22}} + (r_a + r_p)\sigma^2} - \frac{\sigma^2}{2} r_p - \bar{u}. \tag{104}$$

Recall that

$$A = \theta^\top K^{-1}\theta = \frac{k_{22}\theta_1^2 - 2k_{12}\theta_1\theta_2 + k_{11}\theta_2^2}{\Delta}, \tag{105}$$

so we have the identity

$$\frac{(k_{22}\theta_1 - k_{12}\theta_2)^2}{k_{22}\Delta} = A - \frac{\theta_2^2}{k_{22}} = A - B. \tag{106}$$

Therefore, we obtain the compact decomposition

$$\mathrm{CE}_p^{(3)} = \frac{1}{2}A - \frac{\theta_2^2}{2k_{22}} + \frac{1}{2}\frac{\left(\frac{\theta_2^2}{k_{22}} + r_p\sigma^2\right)^2}{\frac{\theta_2^2}{k_{22}} + (r_a + r_p)\sigma^2} - \frac{\sigma^2}{2} r_p - \bar{u} = \frac{1}{2}A - \frac{\sigma^2}{2} r_a\beta^{(3)} - \bar{u}. \tag{107}$$

In case (2), we have

$$\beta^{(2)} = \frac{A + r_p\sigma^2}{A + (r_a + r_p)\sigma^2}, \qquad \mathrm{CE}_p^{(2)} = \frac{1}{2}A - \bar{u} - \frac{\sigma^2}{2} r_a\,\beta^{(2)}. \tag{108}$$

Since $B \le A$ (see (106)), it follows that $\beta^{(3)} \le \beta^{(2)}$, and therefore

$$\mathrm{CE}_p^{(3)} \ge \mathrm{CE}_p^{(2)}. \tag{109}$$

E Evaluation Frictions: Noisy Measurement and Strategic Gaming

Environment (CARA–Normal). The agent chooses unobservable effort $e \in \mathbb{R}$, generating true output

$$y = e + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_y^2). \tag{110}$$

Effort incurs a quadratic cost

$$c(e) = \frac{k}{2}e^2, \qquad k > 0. \tag{111}$$

The principal cannot contract on y and instead observes a noisy measurement z that can be strategically distorted by both parties:

$$z = y + \nu + x_a - x_p, \qquad \nu \sim \mathcal{N}(0, \sigma_m^2), \tag{112}$$

where ν is independent of ε. The agent chooses x_a to inflate the signal, while the principal chooses x_p to deflate it. Manipulation is costly:

$$\text{Agent gaming cost: } \frac{d_a}{2}x_a^2, \qquad \text{Principal gaming cost: } \frac{d_p}{2}x_p^2, \qquad d_a, d_p > 0. \tag{113}$$

Define the total exogenous noise variance in the metric as

$$\sigma^2 := \sigma_y^2 + \sigma_m^2. \tag{114}$$

Both principal and agent have CARA utility. Let w denote the wage and let principal profit be

$$\pi = y - w. \tag{115}$$

The agent's utility is

$$U_a = -\exp\left\{-r_a\left[w - c(e) - \frac{d_a}{2}x_a^2\right]\right\}, \tag{116}$$

and the principal's utility is

$$U_p = -\exp\left\{-r_p\left[\pi - \frac{d_p}{2}x_p^2\right]\right\}. \tag{117}$$

Under CARA–Normal assumptions, both parties maximize certainty equivalents.

Contract and timing. Because effort is unobservable, the contract can depend only on z. We restrict attention to linear contracts

$$w = \alpha + \beta z, \tag{118}$$

where β is the incentive slope. The timing is: (i) the principal offers (α, β); (ii) the agent chooses e and x_a; (iii) (ε, ν) are realized; (iv) the principal chooses x_p; (v) z is finalized and w is paid.

Equilibrium manipulation for a given slope β. Fix β. Given (118)–(112), manipulation enters wages linearly through z.

Lemma E.1 (Two-sided signal gaming)**.** *Fix a contract slope β. In equilibrium,*

$$x_a^*(\beta) = \frac{\beta}{d_a}, \qquad x_p^*(\beta) = \frac{\beta}{d_p}, \tag{119}$$

and therefore

$$z = y + \nu + \frac{\beta}{d_a} - \frac{\beta}{d_p}. \tag{120}$$

Proof. The agent chooses x_a to maximize $\beta x_a - \frac{d_a}{2}x_a^2$, yielding $x_a^*(\beta) = \beta/d_a$. After (ε, ν) are realized, the principal chooses x_p to minimize wage plus manipulation cost, equivalently maximize $\beta x_p - \frac{d_p}{2}x_p^2$, yielding $x_p^*(\beta) = \beta/d_p$. □

A useful implication is the expected deadweight loss from two-sided manipulation:

$$\frac{d_a}{2}\big(x_a^*(\beta)\big)^2 + \frac{d_p}{2}\big(x_p^*(\beta)\big)^2 = \frac{\beta^2}{2}\left(\frac{1}{d_a} + \frac{1}{d_p}\right). \tag{121}$$

Effort response for a given slope β.

Lemma E.2 (Effort response)**.** *Fix a contract slope β. The agent's optimal effort is*

$$e^*(\beta) = \frac{\beta}{k}. \tag{122}$$

Proof. Under CARA–Normality, the agent maximizes a certainty equivalent. The only e-dependent term is $\beta e - \frac{k}{2}e^2$, which is maximized at $e^*(\beta) = \beta/k$. □

Optimal incentive slope under two-sided gaming. In the unique interior equilibrium, the optimal contract $w = \alpha + \beta z$ has slope

$$\beta^* = \frac{\frac{1}{k} + r_p\sigma_y^2}{\frac{1}{k} + \left(\frac{1}{d_a} + \frac{1}{d_p}\right) + (r_a + r_p)(\sigma_y^2 + \sigma_m^2)}, \tag{123}$$

with induced equilibrium actions

$$x_a^* = \frac{\beta^*}{d_a}, \qquad x_p^* = \frac{\beta^*}{d_p}, \qquad e^* = \frac{\beta^*}{k}. \tag{124}$$

Proof. Fix (α, β) and consider the induced equilibrium. By Lemmas E.1 and E.2,

$$e = \frac{\beta}{k}, \qquad x_a = \frac{\beta}{d_a}, \qquad x_p = \frac{\beta}{d_p}.$$

Since $z = e + \varepsilon + \nu + x_a - x_p$, the wage is

$$w = \alpha + \beta z = \alpha + \beta\big(e + x_a - x_p\big) + \beta(\varepsilon + \nu),$$

so

$$\mathrm{Var}(w) = \beta^2\,\mathrm{Var}(\varepsilon + \nu) = \beta^2(\sigma_y^2 + \sigma_m^2).$$

The agent's certainty equivalent is

$$CE_a = \mathbb{E}[w] - \frac{k}{2}e^2 - \frac{d_a}{2}x_a^2 - \frac{r_a}{2}\mathrm{Var}(w)\,.$$

Substituting $\mathbb{E}[w] = \alpha + \beta(e + x_a - x_p)$, $\mathrm{Var}(w) = \beta^2(\sigma_y^2 + \sigma_m^2)$, and the equilibrium actions gives

$$\mathrm{CE}_a = \alpha + \beta\Big(\frac{\beta}{k} + \frac{\beta}{d_a} - \frac{\beta}{d_p}\Big) - \frac{k}{2}\Big(\frac{\beta}{k}\Big)^2 - \frac{d_a}{2}\Big(\frac{\beta}{d_a}\Big)^2 - \frac{r_a}{2}\beta^2(\sigma_y^2 + \sigma_m^2),$$

hence

$$\mathrm{CE}_a = \alpha + \frac{\beta^2}{2k} + \frac{\beta^2}{2d_a} - \frac{\beta^2}{d_p} - \frac{r_a}{2}\beta^2(\sigma_y^2 + \sigma_m^2).$$

At the optimum, the IR constraint binds ($\mathrm{CE}_a = \bar{u}$), so

$$\alpha(\beta) = \bar{u} - \frac{\beta^2}{2k} - \frac{\beta^2}{2d_a} + \frac{\beta^2}{d_p} + \frac{r_a}{2}\beta^2(\sigma_y^2 + \sigma_m^2). \tag{125}$$

Principal profit is $\pi = y - w = (e + \varepsilon) - w$. Using $w = \alpha + \beta z$ and $z = e + \varepsilon + \nu + x_a - x_p$,

$$\pi = (1 - \beta)e - \alpha - \beta(x_a - x_p) + (1 - \beta)\varepsilon - \beta\nu.$$

Thus

$$\mathbb{E}[\pi] = (1 - \beta)e - \alpha - \beta(x_a - x_p), \qquad \mathrm{Var}(\pi) = (1 - \beta)^2\sigma_y^2 + \beta^2\sigma_m^2,$$

and principal manipulation cost is $\frac{d_p}{2}x_p^2 = \frac{\beta^2}{2d_p}$. Therefore the principal's certainty equivalent is

$$\mathrm{CE}_p(\beta) = \mathbb{E}[\pi] - \frac{d_p}{2}x_p^2 - \frac{r_p}{2}\mathrm{Var}(\pi)\,.$$

Substituting $e = \beta/k$, $x_a = \beta/d_a$, $x_p = \beta/d_p$, and $\alpha(\beta)$ from (125), and dropping constants independent of β, yields

$$\mathrm{CE}_p(\beta) = \Big(\frac{1}{k} + r_p\sigma_y^2\Big)\beta - \frac{1}{2}\left[\frac{1}{k} + \Big(\frac{1}{d_a} + \frac{1}{d_p}\Big) + (r_a + r_p)(\sigma_y^2 + \sigma_m^2)\right]\beta^2 + \text{const.}$$

The objective is strictly concave in β, hence the maximizer is unique and satisfies the first-order condition

$$0 = \frac{d\mathrm{CE}_p}{d\beta} = \Big(\frac{1}{k} + r_p\sigma_y^2\Big) - \left[\frac{1}{k} + \Big(\frac{1}{d_a} + \frac{1}{d_p}\Big) + (r_a + r_p)(\sigma_y^2 + \sigma_m^2)\right]\beta.$$

Solving gives (123). Plugging β^* into $e = \beta/k$, $x_a = \beta/d_a$, $x_p = \beta/d_p$ gives (124). □