Pricing AI as a Service

Ilan Lobel[1], Yanwei Sun[2], and Jiahua Wu[3]

[1]NYU Stern School of Business, ilobel@stern.nyu.edu
[2]Imperial Business School, yanwei@imperial.ac.uk
[3]Imperial Business School, j.wu@imperial.ac.uk

Abstract

We study *AI as a Service*, in which a monopolistic provider sells AI capabilities that perform tasks for downstream buyers. Buyers privately differ in both their value from the service and the probability that the AI resolves their task. The seller designs a menu of capabilities and outcome-contingent prices and controls how true task outcomes are mapped into reported outcomes; buyers may dispute and verify the reported outcome at a cost. We first compare two pricing formats commonly used in practice: *usage-based pricing* (UBP), under which the buyer pays per use regardless of the outcome, and *pay-for-success pricing* (PFS), under which the buyer pays only when the deployed AI reports the task as successful. When the AI-reported outcome fully aligns with the buyer's task outcome, PFS becomes more attractive as per-success values become less heterogeneous and higher per-success values coincide with higher success probabilities. In a customer-service application, where an AI front end resolves requests before they join the human service queue, PFS is favored when AI capability and human service are complements, and UBP is favored when they are substitutes. Allowing the seller to choose the billable success metric can further increase PFS profit. When buyer verification is sufficiently costly, a single success payment combined with an appropriately chosen metric can reproduce the fully aligned limited-liability two-price benchmark, so optimal PFS weakly dominates UBP. Our main result shows that this misalignment motive disappears in the unrestricted contract space: separate success and failure prices make full outcome alignment without loss of optimality and eliminate dispute waste. We characterize the optimal mechanism, which reduces to a convex program under mild conditions: the lowest- and highest-success-probability types receive first-best allocations, while interior types may be distorted upward or downward according to curvature virtual types. These results extend to settings with more than two outcomes.

Keywords: Pricing AI; outcome alignment; human–AI interplay; contract with post-allocation outcomes; information design.

1 Introduction

Artificial intelligence is becoming a foundational layer of digital infrastructure. Across sectors, firms deploy AI to draft communications, triage requests, retrieve internal knowledge, and execute routine workflows (McKinsey & Company, 2025). A distinctive feature of this shift is the rise of agentic systems: rather than producing a one-off prediction or text response, these systems act within workflows, attempt to complete tasks end to end, and escalate unresolved cases (Sun, 2025). These systems are increasingly delivered through a service model in which providers offer ongoing access to AI capabilities that perform tasks within buyers' workflows rather than selling the underlying models and infrastructure outright (Deloitte Insights, 2025; Boston Consulting Group, 2025). This emerging delivery model is commonly referred to as *AI as a Service* (AIaaS).

AIaaS inherits the software-as-a-service (SaaS) delivery model, in which software is hosted by the provider and accessed over the cloud (Mell and Grance, 2011; Armbrust et al., 2010), but differs from traditional SaaS in two economically important respects. Traditional SaaS typically monetizes seats and subscriptions: fees are largely decoupled from what the software accomplishes, and the marginal cost of serving an additional user is close to zero. Under AIaaS, by contrast, each request incurs a variable inference cost, and the buyer's willingness to pay depends on whether the AI completes the task. The service being sold is therefore not merely access to software, but completion of work.

This shift has brought two pricing formats to the forefront in practice: *usage-based pricing* (UBP) and *pay-for-success pricing* (PFS). Under UBP, the buyer pays per use or conversation regardless of whether the task is resolved. This format is simple to implement but leaves the performance risk with the buyer, who pays even when the AI fails. Such exposure can discourage adoption when the AI's effectiveness is uncertain. PFS addresses this barrier by charging the buyer only when the deployed success metric records the task as successful—for example, a resolved support ticket, a qualified sales lead, or a completed handoff.[1] PFS thereby shifts performance risk toward the seller and links charges more closely to realized business value. Its appeal, however, depends on a credible and shared definition of success.

Customer service makes this challenge concrete: a request reaches the AI agent, which either resolves it or escalates it to a human server, and PFS bills only when the request is classified as resolved. Whether a request is truly resolved, however, is often not directly observable. The provider must translate this context-dependent outcome into a billable metric using signals such as customer feedback, inactivity, escalation, or an AI-based evaluation. Under Zendesk's automated-resolution policy, for example, a ticket is classified as a billable success when the end customer leaves positive feedback and as a failure when the feedback is negative; when the customer leaves no feedback, the

[1] PFS is already common in customer service; see, e.g., Fin AI and Can I Help, which bill per resolved conversation. Beyond customer service, outcome-based billing is also emerging in consulting; see Louise Lucas, "How AI is forcing McKinsey and its peers to rethink pricing," *Financial Times*, May 24, 2026.

AI agent itself evaluates whether the request was resolved. Even without strategic manipulation, the resulting metric may differ from the buyer's value-relevant notion of task completion. This alignment concern is amplified because the seller designs and operates the metric while its payment depends on the AI-reported outcome, giving it an incentive to resolve ambiguity in its own favor and creating scope for disputes.

We formalize this tension as a joint contract- and alignment-design problem. A profit-maximizing monopolistic AI provider offers a menu to a privately informed buyer whose type determines both her value from the service and the probability that the AI resolves her task. Each menu item combines a contract—an AI capability, a success price, and a failure price—with an outcome-alignment rule[2] that maps the true task outcome according to the buyer's criterion into the billable outcome reported by the AI. For each item, this rule is summarized by two probabilities: the probability that the AI reports success after a true success and the probability that it reports success after a true failure. The rule captures post-training, prompting, evaluation-rule design, threshold selection, or product-policy choices that shape how the deployed metric records task completion. After observing the AI-reported outcome, the buyer either pays the corresponding bill or disputes it. A dispute verifies the true outcome and bases payment on that outcome, but imposes costs on both parties, reflecting the effort required to audit the interaction, review records, adjudicate the claim, or negotiate a billing adjustment. The seller chooses the menu to maximize expected profit subject to the buyer's optimal pay-or-dispute decisions, incentive compatibility, and individual rationality.

We first use this framework to study the practical pricing tradeoff introduced above: the choice between UBP and PFS. Under UBP, the success and failure bills are identical, so the AI-reported outcome is payoff-irrelevant and the problem reduces to standard one-dimensional screening on the buyer's value. Under fully aligned PFS, the seller charges only after a true success, and the relevant screening index becomes the buyer's *per-success value*—her value divided by her success probability. This creates a wedge between the buyer's willingness to accept a success price and the seller's expected profit: a buyer with high per-success value may accept a high success price but pay it rarely, while the seller incurs the deployment cost whenever the service is provided. This wedge yields an asymmetric worst-case result. With n types, UBP guarantees at least a $1/n$ share of the unrestricted second-best profit, and therefore at least a $1/n$ share of the profit from fully aligned PFS. By contrast, fully aligned PFS can perform arbitrarily poorly relative to UBP, even with only two types. More generally, the comparison depends on the relationship between per-success value and success probability. Fully aligned PFS becomes more attractive when per-success values are less heterogeneous and buyers with higher per-success values are also more likely to succeed, so high

[2]Throughout, *outcome alignment* refers narrowly to how closely the AI-reported outcome tracks the buyer's contractual success criterion. It is distinct from the broader use of "AI alignment" in the safety literature—aligning model behavior with human intent and preventing misuse or misbehavior—which is outside our scope.

willingness to pay coincides with a high probability that the seller collects the success payment. Allowing the seller to design the billable success metric can, however, increase PFS profit. If the metric records some true failures as billable successes, the seller can collect payments in states that fully aligned PFS leaves unbilled. This gain is weakly increasing in the buyer's verification cost. When verification becomes sufficiently costly, a single success payment combined with an appropriately chosen metric can reproduce the fully aligned limited-liability two-price benchmark. Optimal PFS then attains this benchmark and weakly dominates UBP.

We apply the fully aligned UBP–PFS comparison to a customer-service setting. A downstream firm deploys an AI agent as the front end for incoming requests: the agent resolves some requests and escalates the rest to capacity-constrained human servers. By diverting resolved requests from the human queue, AI adoption reduces congestion, and the buyer's value is the resulting reduction in congestion cost. The application features two buyer types: downstream firms that differ in how often the AI resolves their requests and in how congested their human service queues are. In this two-type application, PFS weakly outperforms UBP when the buyer type whose requests the AI resolves more often also has sufficiently high baseline human utilization; below this utilization threshold, UBP weakly outperforms PFS. Thus, PFS is more attractive when AI capability and human service are *complements*: the AI performs well precisely where the human service system is more congested. By contrast, UBP is more attractive when they are *substitutes*: the strengths of AI and human service overlap.

The PFS analysis sharpens the credibility concern introduced above: under success-only pricing, the seller may profit from a billable metric that differs from the buyer's value-relevant success criterion. Buyers may therefore be reluctant to trust provider-defined success. Our main design result shows that this misalignment motive arises from the restriction to a single success payment rather than from outcome-contingent pricing itself. Once the seller can charge separate success and failure fees—equivalently, a baseline fee plus a success premium—any feasible menu can be replaced, item by item, by a fully aligned two-price menu that preserves allocations and each assigned type's expected cost, makes deviations weakly less attractive, and eliminates dispute waste. Thus, *full outcome alignment is without loss of optimality*: the incentive role previously played by metric design can instead be embedded transparently in the payment schedule. This result is *not* an application of the revelation principle, which concerns the buyer's reports; the object here is the seller's own rule for reporting outcomes. The argument is instead geometric: each item's optimized buyer cost is a concave lower envelope of pay-or-dispute costs, and a supporting line at the intended type yields two outcome-contingent prices that preserve her cost, weakly raise every mimic's cost, and eliminate dispute waste.

We then characterize the optimal unrestricted mechanism. The analysis rests on a geometric implementability condition: an allocation-rent profile is implementable if and only if, for every menu item, the point associated with its intended type admits an affine expected-payment line

that passes through it and lies weakly above the corresponding points for all mimicking types. We call this the *upper-support condition*. The types with the lowest and highest success probabilities are upper-supported at any positive-service item because they face mimics on only one side of the success-probability axis. Consequently, both endpoint types—the buyers with the hardest and the easiest tasks—receive first-best allocations, in contrast to standard one-dimensional screening under UBP, where only the highest-value type is undistorted. Under separable primitives, the upper-support condition becomes a set of linear chord constraints in the space of allocation-rent profiles, and the problem becomes a convex program and is thus computationally tractable. Each interior type's optimal allocation maximizes a virtual surplus with an adjusted valuation parameter—its *curvature virtual type*—which encodes how buyer valuations curve in success probabilities across types; accordingly, interior allocations can be distorted either upward or downward. If the value shifters are discretely concave in the success-probability shifters, the seller can implement the first-best allocation with zero rents and attain the first-best profit. These results extend to settings with more than two outcomes.

The remainder of the paper is organized as follows. Section 2 reviews the related literature, and Section 3 presents the model. Section 4 compares UBP and PFS and applies the comparison to customer service. Section 5 analyzes the unrestricted two-price mechanism, establishes that full outcome alignment is without loss, and characterizes the optimal allocations. Section 6 concludes. All proofs appear in the appendix.

2 Related Literature

Our paper sits at the intersection of several literatures: the economics and pricing of AI, human–AI interaction and AI service operations, performance-based contracting, outcome alignment and information design, and mechanism design and pricing with post-allocation signals.

Economics and pricing of AI. A fast-growing literature studies the economics and pricing of AI. A first strand treats AI as a strategic *pricing agent* and asks what happens when firms delegate pricing to it. Calvano et al. (2020) show that Q-learning pricing algorithms can sustain supracompetitive prices without communication, and Banchio and Mantegazza (2022) identify spontaneous coupling as a mechanism behind such outcomes. More recently, Fish et al. (2024) show experimentally, and Cao and Hu (2026) show theoretically, that LLM-based pricing can sustain tacit collusion and supracompetitive prices. A second strand prices the AI *itself* on the supply side. Gürkan and de Véricourt (2022) study pricing and contracting for outsourced ML products when adoption generates usage data that improves future algorithmic accuracy but also changes the provider's effort incentives. Bergemann et al. (2025b) show that, under a homogeneous token-production technology, heterogeneous task valuations collapse to a scalar aggregate type, so LLM providers can screen users with committed-spend token-budget menus; their extensions rationalize model and usage tiers such as those used by Anthropic and OpenAI. Mahmood (2024) studies pricing competition among

generative-AI developers across heterogeneous downstream tasks, and Demirer et al. (2025) document price dispersion, model differentiation, and adoption patterns in the LLM market. Related work designs token-level auctions for LLM systems (Dütting et al., 2024b).

These papers price AI access, model use, or upstream inputs such as tokens, model choice, and training data. We instead study the downstream sale of AI task *completion*: the buyer faces outcome-contingent prices. Closer to this question, Iyer et al. (2025) study how to sell a service of uncertain quality and show that the profit-maximizing scheme necessarily pairs an upfront price with outcome-dependent prices. The two papers differ along three main dimensions. First, Iyer et al. (2025) study voluntary usage: after observing realized outcome, the buyer may reject the outcome, forgo access, and avoid the corresponding usage payment. In our model, participation is voluntary ex ante, but billing is enforceable after the service is provided. Upon observing the AI-reported outcome, the buyer may either pay the bill or dispute it and pay according to the verified outcome; she cannot simply reject the service to avoid payment. Second, their buyer directly observes realized quality, whereas in our model neither party initially observes the true task outcome at billing. Instead, the seller's AI produces the billable success metric, and the seller chooses how closely this metric aligns with the buyer's value-relevant success criterion. This introduces an information-design and verification layer absent from their model. Third, buyer type in their model affects valuations over outcomes but not the action-induced outcome distribution. In our model, type affects both the buyer's value and the probability that the AI completes the task. Private information therefore concerns both willingness to pay and the AI's type-dependent performance.

Human–AI interaction and AI service operations. Our paper connects to two related strands. The first studies *human–AI interaction*—how AI agents are overseen, trusted, and communicated with: Bastani and Cachon (2025) show that sustaining human oversight becomes prohibitively costly as AI errors grow rare; Overman and Bayati (2025) model AI autonomy and human oversight as a two-player Markov game; and Castro et al. (2024) study how users' interaction with generative AI shapes its outputs. The second studies *AI in service systems*—the operational setting of our application—where an AI front end absorbs part of the workload and reshapes the human servers behind it: from classic call centers with automated front ends (Khudyakov et al., 2010) to recent work on scheduling AI-based queues under prediction error (Lee et al., 2024), automation and augmentation in customer-intensive services (Hu et al., 2025), and when to defer to and how to schedule human review (Lykouris and Weng, 2025; Gocmen et al., 2025). Our customer-service application in Section 4.3 focuses on the pricing part, investigating how the substitutability and complementarity between AI capability and human capacity shape the optimal pricing mechanism. Closest to our application, Feldman et al. (2026) study a firm that sells an AI–human hybrid service directly to consumers, who privately observe whether the AI resolved their problem while the firm may not. Our setting has a different vertical structure: a third-party AI provider sells AI-agent

services to a downstream firm, rather than the same firm owning the AI and serving end customers. The provider's AI reports the billable outcome, which the downstream firm may dispute at a cost, and the downstream firm privately knows both its value from the service and the AI's probability of resolving its tasks.

Performance-based contracts. A central benchmark for performance-based contracts is moral hazard: hidden effort shifts a contractible performance signal, and outcome-contingent pay trades off incentives and risk sharing (Holmström, 1979; Holmström and Milgrom, 1987; Grossman and Hart, 1983; Dütting et al., 2024a). Related service and supply-chain work applies this moral-hazard lens to availability-based after-sales contracts with hidden inventory and cost-reduction choices (Kim et al., 2007) and to collaborative services with hidden buyer and vendor inputs (Roels et al., 2010); for related empirical work on performance-based contracts and reliability, see Guajardo et al. (2012). Our setting departs from this moral-hazard paradigm: the seller operates the AI and offers the contract, so there is no hidden effort; instead, the buyer is privately informed about her type. The core friction is thus adverse selection over hidden types rather than moral hazard over hidden actions.

The closest analogue to our UBP–PFS comparison is online advertising (Edelman et al., 2007): pay-per-impression charges for exposure, like usage-based pricing, while pay-per-click or pay-per-action charges only after a user action, like pay-for-success (Asdemir et al., 2012; Dellarocas, 2012). The analogy holds at the level of the pricing *format*, but the frictions differ. In Asdemir et al. (2012), pay-per-impression leaves advertising targeting decisions, e.g., the intensity, with the imperfectly informed advertiser, whereas pay-per-click delegates it to the better-informed publisher—who then aims to maximize clicks rather than advertiser surplus. In Dellarocas (2012), pay-per-action advertising is modeled as a vertical relationship between a publisher and an advertiser who sets a downstream product price; tying the advertising fee to clicks can distort that product price relative to pay-per-exposure. Our setting differs on both sides: the privately informed party is the *buyer*, whose type sets both value and task difficulty—a hidden-type problem—and billable success is *recorded by the seller's own AI success metric*, so the seller also designs the outcome-alignment rule that turns value-relevant outcomes into billable outcomes.

Outcome alignment and information design. Our model contains an outcome-alignment layer: the seller commits to an outcome-alignment rule that maps the buyer's value-relevant task outcome into the billable outcome used for pricing. Viewed as the choice of a signal structure, this layer falls under the umbrella of the Bayesian-persuasion and information-design literature, which studies how a designer's choice of signal shapes receivers' beliefs and actions (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2016, 2019; Candogan, 2020). Putting the contract-design problem aside, our method for solving the information-design layer in the unrestricted contract space differs from the usual concavification approach (Kamenica and Gentzkow, 2011) and the Bayes-correlated-equilibrium approach (Bergemann and Morris, 2016, 2019). We instead use a

representation result, see Theorem 2: for any outcome-alignment rule and contract, we construct another contract with adjusted prices and *a fully aligned success metric* (equivalent to full information revealing in the information-design language) that preserves the incentive constraints and achieves weakly higher profit.

Screening and mechanism design with post-allocation signals. Abstracting from the information-design layer, our contract-design problem is one of nonlinear pricing (screening) with post-allocation information. When payments cannot depend on realized outcomes, the problem reduces to standard screening (Mussa and Rosen, 1978; Maskin and Riley, 1984; Haghpanah and Siegel, 2025). The distinguishing feature of our setting is that payments can also condition on an outcome signal. The related literature can be organized by *the source of this signal*.

When the signal is *exogenous*—its distribution varies with the buyer's type but cannot be influenced by the designer—outcome-contingent payments operate as side bets that exploit differences in types' beliefs about the signal. This logic first appeared in Myerson's auction example (Myerson, 1981, Section 7) and was subsequently developed more generally by Crémer and McLean (1985, 1988) and McAfee and Reny (1992). Under suitable conditions, such payments can discipline type reports and enable full-surplus extraction. See Lopomo et al. (2022) for a later synthesis and Sinander (2026, Section 1.8) for an overview. The same force appears in procurement, where Riordan and Sappington (1988) show that a public ex post signal correlated with the contractor's privately known cost implements the buyer's first best whenever the signal is sufficiently informative.

More recent work moves beyond pure first-best extraction. Alaei et al. (2024) and Belloni et al. (2025) study settings in which the post-allocation signal is the buyer's private valuation itself, motivated by the seller's post-allocation inspection. In repeated interactions, such inspections gain force through exclusion: Li et al. (2025) show that although current payments cannot condition on a post-allocation inspection, the threat of excluding a detected misreporter from future rounds disciplines reports and raises revenue. Alaei et al. (2026) study the sale of a single item when the seller observes a signal correlated with the buyer's type. They characterize the optimal mechanism by the shape of the belief function: convexity permits full-surplus extraction, concavity yields an outcome-contingent posted price, and the general case partitions the type space into regions governed by these two regimes. Post-allocation measurement also matters in procurement. Gupta et al. (2021) study contractors privately informed about both their costs and their estimates of a noncost attribute, such as completion time, that is realized only after project completion (allocation). They show that treating the realized attribute as tamper-proof can induce the winning contractor to manipulate it; indeed, the private estimate generates information rent in their model only because manipulation is possible. In our setting, control over the payment-relevant outcome instead rests with the AI seller, and Theorem 2 shows that this control generates no additional profit when separate success and failure prices are available. When delivered quality is noncontractible, realized outcomes instead facilitate learning rather than enter the contract directly (Gur et al., 2022).

Our setting differs because the signal is *endogenous*: the seller offers a menu of AI capability levels, and the assigned capability determines the distribution of payment-relevant outcomes. Capability therefore affects not only value and cost but also expected payments across types, making allocation itself a screening instrument. This endogeneity generates distinctive allocation patterns: both the lowest- and highest-success-probability types receive first-best allocations, while interior types may be distorted either upward or downward according to curvature virtual types. The resulting design problem is generally nonconvex in its original variables, although under separable primitives it admits a convex reformulation. Relatedly, Mezzetti (2004) and Bergemann et al. (2025a) also study endogenous post-allocation signals, but focus on implementing welfare-maximizing allocations; we instead study profit maximization. We further endogenize how the AI's value-relevant outcome is mapped into a billable success metric, adding an information-design layer to the contract-design problem. This layer introduces an additional buyer decision: after choosing a menu item and observing the AI-reported outcome, the buyer decides whether to accept the report or incur a cost to dispute it and verify the true outcome.

3 Model

We study a single AI provider (the seller) that sells access to AI agents that complete downstream tasks for a buyer. Although the same AI model can perform differently across task types—for example, coding versus writing—for a specific task, AI models' performance is usually well ordered. We thus adopt the simplifying assumption that AI capability is one-dimensional, representing it by a scalar $a \in \mathcal{A} \subseteq \mathbb{R}_+$. Providing capability a costs $c(a) \geq 0$. The set $\mathcal{A}$ may be discrete, representing model or service tiers, or continuous, representing adjustable inference effort or other scalable deployment resources.

The buyer privately knows her type $t \in \mathcal{T} = \{1, \ldots, n\}$, drawn with probability $\pi_t > 0$. The buyer has a task. Deploying AI capability $a \in \mathcal{A}$ generates a binary task outcome $\omega \in \{0, 1\}$, where 1 denotes success and 0 denotes failure according to the buyer's success criterion; for a type-t buyer, the task is resolved with probability $\gamma_t(a) \in [0, 1]$. Upon outcome ω, the buyer's realized value is $\widetilde{v}_t(a, \omega)$, where $\widetilde{v}_t(a, 1)$ and $\widetilde{v}_t(a, 0)$ denote her values after a success and after a failure, respectively; both may depend on the deployed capability a, since how the task is handled—not only whether it is resolved—can affect the buyer's payoff. Because the buyer decides whether and what to purchase *before* the outcome realizes, the relevant payoff object is her expected value from deploying capability a,

$$v_t(a) := \gamma_t(a)\widetilde{v}_t(a, 1) + (1 - \gamma_t(a))\widetilde{v}_t(a, 0).$$

The analysis therefore works with the pair $(v_t(a), \gamma_t(a))$. We relax the binary-outcome assumption in Section A.3.

This formulation allows value and success probability to vary independently across types and

AI capabilities. A buyer may derive high value from AI because successful resolution relieves a severe operational bottleneck, even when success is unlikely. Conversely, AI may succeed often on tasks that are already easy or inexpensive to handle without it, generating relatively low buyer value. We maintain throughout that a no-service allocation is available, denoted by $0 \in \mathcal{A}$, with $v_t(0) = 0$, $\gamma_t(0) = 0$, and $c(0) = 0$ for every type t.

Our model also admits a *dynamic* interpretation, in which deployment generates an infinite sequence of binary task outcomes: each task is resolved with probability $\gamma_t(a)$, and is billed on its own reported outcome. In that case, $v_t(a)$ is the buyer's steady-state expected value—in the customer-service application of Section 4.3, the steady-state reduction in congestion cost. The analysis applies unchanged with $(v_t(a), \gamma_t(a))$ as per-task primitives.

We assume that both the buyer's expected value from AI and the probability that the AI resolves the task are private knowledge of the buyer. This is because both objects may depend on the buyer's operational environment, not only on the AI agent's technical capability. In customer service, for example, the seller may have a reasonable estimate of how well its AI agent handles a particular class of requests. However, downstream firms may face different mixes of requests, and this task mix is often known only to the firm itself. Similarly, the value of deploying the AI agent depends on the firm's own operating conditions. The agent is more valuable when human staff are capacity constrained, or when human resolution is slow or costly. These conditions are also typically private to the downstream firm.

Contracts and outcome alignment. By the revelation principle (Myerson, 1979), we can restrict attention to direct menus in which each buyer type reports her type by choosing the corresponding menu item. The seller offers a menu

$$\{(a_t, p_t, q_t, \sigma_t^1, \sigma_t^0)\}_{t \in \mathcal{T}}.$$

If a buyer with true type t reports type r, she receives item r, given by $(a_r, p_r, q_r, \sigma_r^1, \sigma_r^0)$. The first three components form the contract: a_r is the deployed AI capability, p_r is the fee charged after a success, and q_r is the fee charged after a failure. The outcome used for payment is the AI-reported outcome if the buyer does not dispute, and the verified outcome if the buyer disputes. We describe the AI-reporting and dispute process below.

Recall that $\omega \in \{0, 1\}$ denotes the true task outcome according to the buyer's value-relevant success criterion, with $\omega = 1$ denoting success and $\omega = 0$ denoting failure. Before deployment, the buyer specifies an operational criterion for success, which may include requirements for task completion, escalation, follow-up, brand voice, compliance, and customer retention. The billable outcome may nevertheless be based on imperfect signals. In customer service, for example, resolution may be inferred from customer feedback, inactivity, the absence of escalation, or an AI-based evaluation of whether the response was satisfactory. These signals need not coincide with the buyer's criterion. A buyer may regard a request as unresolved because the AI violated its service standards, while the

seller may classify the same interaction as resolved because the customer provided positive feedback or stopped responding and an AI evaluator judged the response relevant.[3]

Let $\widehat{\omega} \in \{0, 1\}$ denote the AI-reported outcome used for billing, with $\widehat{\omega} = 1$ denoting a reported success and $\widehat{\omega} = 0$ denoting a reported failure. Because the seller controls the system's post-training, prompts, thresholds, evaluation rules, and other product-policy choices, $\widehat{\omega}$ may differ from ω. For each menu item r, the mapping from the buyer's value-relevant outcome to the billable outcome is summarized by the outcome-alignment rule $\sigma_r = (\sigma_r^1, \sigma_r^0)$, where

$$\sigma_r^1 = \mathbb{P}(\widehat{\omega} = 1 \mid \omega = 1, r), \qquad \sigma_r^0 = \mathbb{P}(\widehat{\omega} = 1 \mid \omega = 0, r).$$

Thus, σ_r^1 is the probability that the metric reports success after a true success, whereas σ_r^0 is the probability that it reports success after a true failure. We assume that the seller can commit to this rule and that buyers observe it before choosing a menu item. This assumption is natural in our setting because the billable success metric is implemented by an AI system whose behavior is pre-coded by the seller. Buyers therefore know each item's deployed rule when selecting a contract. We define full alignment in our setting as follows.

Definition 1 (Full outcome alignment)**.** *A menu exhibits* full outcome alignment *if, for every item $r \in \mathcal{T}$, $\sigma_r^1 = 1$ and $\sigma_r^0 = 0$. Under full outcome alignment, the AI-reported outcome coincides with the buyer's value-relevant task outcome.*

We model the reporting technology flexibly by allowing any $(\sigma_r^1, \sigma_r^0) \in [0, 1]^2$. This specification follows the standard information-design approach of allowing the designer to choose any stochastic mapping from the underlying outcome to the reported signal, and provides a useful benchmark for assessing the value of success-metric design. Despite allowing arbitrary outcome-alignment rules, Theorem 2 shows that the unrestricted problem admits an optimal mechanism with full outcome alignment.

After observing the AI-reported outcome $\widehat{\omega}$, the buyer either pays the corresponding fee or disputes. At billing time, neither party observes the true outcome ω; the buyer's pay-or-dispute decision is therefore based on her posterior belief about ω given the report $\widehat{\omega}$, her type, and the committed rule. A dispute incurs cost $\kappa_B \geq 0$ on the buyer side and cost $\kappa_S \geq 0$ on the seller side, but allows the buyer to verify the true outcome ω for billing. These costs reflect verification costs and capture the resources required to audit the interaction, review transcripts or logs, contact the customer, adjudicate the claim, or negotiate a billing adjustment. The timeline of the game is summarized in Figure 1. Ties in the buyer's dispute decision are broken in favor of the seller.

The seller's problem. We now write the seller's profit-maximization problem by first defining the buyer's expected cost and the seller's expected payment from each menu item. Fix a menu

[3]For example, Zendesk's automated-resolution documentation states that a request may count as resolved after a period of inactivity when the customer provides no feedback and an AI evaluation confirms relevance; see Zendesk, "About automated resolutions for AI agents."

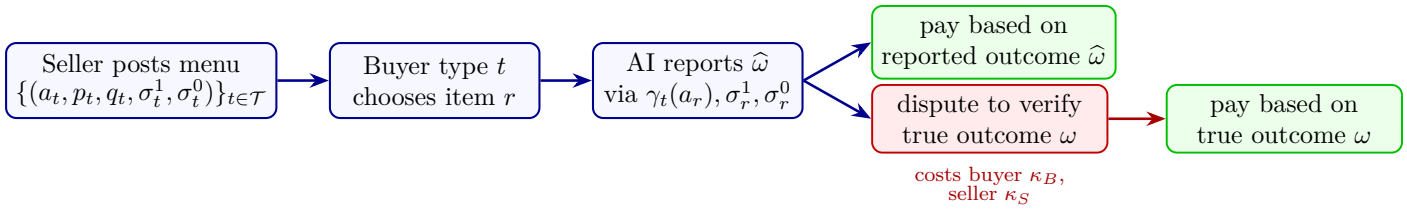


Figure 1: Timeline of the game

Notes. The seller first commits to a contract and an outcome-alignment rule. A buyer with true type t chooses item r. The AI report $\widehat{\omega}$ is generated from the following process: the task succeeds with probability $\gamma_t(a_r)$, and the reported outcome follows the success-metric rule (σ_r^1, σ_r^0). After observing $\widehat{\omega}$, the buyer either pays the bill or disputes to verify the true outcome ω. If the buyer disputes, billing is based on the verified true outcome ω; the buyer incurs verfication cost κ_B and the seller incurs verification cost κ_S. Neither party observes the true outcome ω at billing time: the buyer's pay-or-dispute decision is based on the reported outcome $\widehat{\omega}$ and her posterior belief about ω, and the true outcome is revealed only if the buyer disputes.

item r. Let $P_r^R(\widehat{\omega}) := \widehat{\omega}p_r + (1-\widehat{\omega})q_r$ be the payment based on the reported outcome, and let $P_r^V(\omega) := \omega p_r + (1-\omega)q_r$ be the payment based on the verified true outcome. When a buyer of type t chooses this item, her success probability is $\gamma_t(a_r)$; for the moment, write this probability generically as γ. After observing the AI-reported outcome, the buyer chooses whether to pay directly or dispute. Write $\delta = (\delta_0, \delta_1) \in \{0,1\}^2$ for a dispute rule,[4] where, for each $j \in \{0,1\}$, $\delta_j = 1$ means that the buyer disputes after observing $\widehat{\omega} = j$. The buyer's expected cost from item r is

$$C_r(\gamma) := \min_{\delta\in\{0,1\}^2} \mathbb{E}_{\omega,\widehat{\omega}}\left[(1-\delta_{\widehat{\omega}})P_r^R(\widehat{\omega}) + \delta_{\widehat{\omega}}(P_r^V(\omega)+\kappa_B)\right]. \tag{1}$$

The expectation is over $(\omega, \widehat{\omega})$ induced by $\omega \sim \text{Bernoulli}(\gamma)$ and the outcome-alignment rule (σ_r^1, σ_r^0). The buyer's optimal dispute rule is given by[5]

$$\delta_r^*(\gamma) \in \arg\min_{\delta\in\{0,1\}^2} \mathbb{E}_{\omega,\widehat{\omega}}\left[(1-\delta_{\widehat{\omega}})P_r^R(\widehat{\omega}) + \delta_{\widehat{\omega}}(P_r^V(\omega)+\kappa_B)\right]. \tag{SR}$$

Write $\delta_r^*(\gamma) = (\delta_{r,0}^*(\gamma), \delta_{r,1}^*(\gamma))$, where $\delta_{r,j}^*(\gamma)$ is the optimal dispute decision after observing $\widehat{\omega} = j$. The seller's expected payment from the same item, net of seller dispute costs but before the cost of providing AI capability, is

$$G_r(\gamma) := \mathbb{E}_{\omega,\widehat{\omega}}\left[(1-\delta_{r,\widehat{\omega}}^*(\gamma))P_r^R(\widehat{\omega}) + \delta_{r,\widehat{\omega}}^*(\gamma)P_r^V(\omega)\right] - \kappa_S\mathbb{E}_{\omega,\widehat{\omega}}\left[\delta_{r,\widehat{\omega}}^*(\gamma)\right]. \tag{2}$$

Thus, if type t chooses item r, her expected utility is $v_t(a_r) - C_r(\gamma_t(a_r))$, and the seller's expected profit from that choice is $G_r(\gamma_t(a_r)) - c(a_r)$. With this notation, the seller's profit-maximization

[4] Restricting the dispute rule to be pure-strategy δ_j to $\{0,1\}$ is without loss: the buyer either pays or disputes with certainty after each reported outcome.

[5] Equivalently, after observing reported outcome $j \in \{0,1\}$, the buyer compares the reported-outcome payment $P_r^R(j)$ with the expected verified-outcome payment plus the dispute cost, $\mathbb{E}[P_r^V(\omega) \mid \widehat{\omega} = j] + \kappa_B$, where the posterior is induced by $\omega \sim \text{Bernoulli}(\gamma)$ and the committed outcome-alignment rule (σ_r^1, σ_r^0). The ex ante minimization in (1) is this contingent decision rule written before $\widehat{\omega}$ is realized.

problem is

$$R_{\mathsf{SB}} := \max_{\{a_t,p_t,q_t,\sigma_t^1,\sigma_t^0\}_{t\in\mathcal{T}}} \quad \sum_{t\in\mathcal{T}} \pi_t\big[G_t(\gamma_t(a_t)) - c(a_t)\big] \tag{3}$$

$$\text{s.t.} \quad v_t(a_t) - C_t(\gamma_t(a_t)) \geq v_t(a_r) - C_r(\gamma_t(a_r)), \qquad \forall t, r \in \mathcal{T}, \qquad (\mathsf{IC})$$

$$v_t(a_t) - C_t(\gamma_t(a_t)) \geq 0, \qquad \forall t \in \mathcal{T}. \qquad (\mathsf{IR})$$

We refer to R_{SB}, the value of problem (3), as the seller's unrestricted second-best profit.[6] Equivalently, at a high level, the seller's profit-maximization problem can be read as

$$\max_{\{(a_t,p_t,q_t,\sigma_t^1,\sigma_t^0)\}_{t\in\mathcal{T}}} \quad \sum_{t\in\mathcal{T}} \pi_t \left[\mathbb{E}\{\text{payment}_t\} - \kappa_S \mathbb{P}\{\text{dispute}_t\} - c(a_t)\right]$$

$$\text{s.t.} \quad \text{buyer optimally pays or disputes after observing } \widehat{\omega}, \quad \text{(Sequential Rationality)}$$

$$\text{buyer-}t \text{ selects the item } t\ (a_t, p_t, q_t, \sigma_t^1, \sigma_t^0), \qquad \text{(IC)}$$

$$\text{expected utility is nonnegative.} \qquad \text{(IR)}$$

The objective is expected profit: expected payments net of the seller-side dispute costs induced by potential buyer disputes and the costs of providing AI capability. The constraints are sequential rationality (SR), incentive compatibility (IC), and individual rationality (IR). Sequential rationality requires that, after observing the AI-reported outcome, the buyer optimally chooses whether to pay or dispute; this requirement is encoded in $C_r(\cdot)$ and in $G_r(\cdot)$ through the optimal dispute rule in (SR).

Contract classes. Two contract classes receive substantial attention and are widely used in practice: *usage-based pricing* (UBP) and *pay-for-success pricing* (PFS). Under UBP, the buyer makes the same payment regardless of the outcome:

$$\mathsf{UBP}: \quad p_t = q_t \geq 0 \qquad \forall t.$$

This format is arguably the easiest to implement because payment depends only on usage, not on whether the AI completes the task. However, buyers are increasingly wary of usage-based contracts because spending can scale with activity even when the AI fails to deliver business value. This concern creates demand for contracts that link payment more directly to success. In response, many AI-agent providers have adopted PFS, under which the buyer pays only after the task is reported as successfully completed:

$$\mathsf{PFS}: \quad q_t = 0, \quad p_t \geq 0 \qquad \forall t.$$

PFS also admits a simple microfoundation: for tasks that generate value only upon completion, it is easy to show that PFS is the only pricing format compatible with nonnegative payments and ex-post buyer participation. PFS, however, raises its own buyer-side concerns. In practice, success

[6] Throughout the paper, max denotes sup when the maximum is not attained, and statements about optimal mechanisms or optimizers are conditional on an optimum being attained. An optimum always exists when $\mathcal{A}$ is finite.

may be difficult to define, and the AI's success metric may not be perfectly aligned with the buyer's value-relevant notion of task completion. These concerns are central to the ongoing debate over how to price AI services that substitute for parts of firms' traditional workflows.[7]

Section 4 zooms in on these two contract classes and offers a focused comparison of their advantages and limitations to inform this practical debate. We show that PFS can outperform UBP when the tasks for which success is most valuable are also the tasks the AI is most likely to resolve. At the same time, PFS gives the seller an incentive to misalign the AI's success metric in order to increase profit, which is precisely the credibility concern that can hinder the adoption of PFS in practice. In Section 5, we return to the unrestricted problem in (3), where the seller may choose arbitrary success and failure payments, and show that full alignment is without loss for the seller.

4 Usage-Based Pricing vs. Pay-for-Success Pricing

We proceed in three steps. First, Section 4.1 characterizes optimal UBP and fully aligned PFS. Next, Section 4.2 compares the two pricing formats by separating the payment-base effect from the outcome-alignment effect. Finally, Section 4.3 applies the payment-base comparison to a customer-service queueing setting.

4.1 Optimal UBP and Fully Aligned PFS

In this section, we first characterize the optimal usage-based pricing mechanism, for which the AI-reported outcome is payoff-irrelevant, and the optimal pay-for-success mechanism under full outcome alignment, i.e., $\sigma_r^1 = 1$ and $\sigma_r^0 = 0$ for every r. For tractability, several results specialize the primitive functions to the following separable specification.

Definition 2 (Separable primitives)**.** *The private type enters multiplicatively in both the value function and the success-probability function:*

$$v_t(a) = \theta_t\, h(a), \qquad \gamma_t(a) = \alpha_t\, s(a), \qquad t \in \mathcal{T}, \qquad \theta_t > 0, \quad \alpha_t > 0. \tag{4}$$

Here $h(\cdot)$ is increasing, and $s(\cdot)$ is such that $\gamma_t(a) \in [0,1]$ for all $a \in \mathcal{A}$ and $t \in \mathcal{T}$. Consistent with the no-service normalization, $h(0) = s(0) = 0$. We also assume $s(a) > 0$ whenever $h(a) > 0$, so every positive-value allocation can generate a success-contingent payment under PFS.

This multiplicative form is standard in screening models because it separates type heterogeneity from the common allocation technology: θ_t and α_t capture buyer-specific heterogeneity, while $h(a)$ and $s(a)$ capture how AI capability affects value and success probability. Under Definition 2, two orderings will be useful. We use parentheses to denote the order of types by value θ_t, and brackets

[7] See, for example, Decagon, "Pricing the AI Agent Economy", Metronome, "What Is Outcome-Based Pricing?", and Siena, "Why resolutions aren't outcomes".

to denote the order of types by per-success value $\phi_t := \theta_t/\alpha_t$:

$$\theta_{(1)} \le \cdots \le \theta_{(n)}, \qquad \phi_{[1]} \le \cdots \le \phi_{[n]}.$$

Proposition 1 (Optimal UBP)**.** *Under* UBP*, the buyer never benefits from disputing, and the seller's success-metric design is payoff-irrelevant. Thus, finding the optimal* UBP *under the separable primitive functions in Definition 2 reduces to the standard one-dimensional screening problem (Mussa and Rosen, 1978): there exists an optimal* UBP *menu with* $a_{(1)}^{\mathsf{UBP}} \le \cdots \le a_{(n)}^{\mathsf{UBP}}$ *in the* θ*-order that solves*

$$R^*_{\mathsf{UBP}} = \max_{\substack{a_{(1)},\dots,a_{(n)}\in\mathcal{A} \\ a_{(1)}\le\cdots\le a_{(n)}}} \sum_{t=1}^{n} \pi_{(t)} \left[\widetilde{\theta}_{(t)} h(a_{(t)}) - c(a_{(t)})\right]. \tag{5}$$

where the UBP *virtual types are*

$$\widetilde{\theta}_{(t)} := \theta_{(t)} - \frac{\sum_{k=t+1}^{n} \pi_{(k)}}{\pi_{(t)}} \left(\theta_{(t+1)} - \theta_{(t)}\right), \quad t < n, \qquad \widetilde{\theta}_{(n)} := \theta_{(n)}.$$

The optimal prices satisfy

$$p_{(1)}^{\mathsf{UBP}} = \theta_{(1)} h\left(a_{(1)}^{\mathsf{UBP}}\right), \qquad p_{(t)}^{\mathsf{UBP}} = p_{(t-1)}^{\mathsf{UBP}} + \theta_{(t)} \left[h\left(a_{(t)}^{\mathsf{UBP}}\right) - h\left(a_{(t-1)}^{\mathsf{UBP}}\right)\right], \quad t \ge 2. \tag{6}$$

Under UBP, each item has the same success and failure payment, $p_r = q_r$. Thus, the AI-reported outcome does not affect either the buyer's payment or the seller's profit, and disputing can only add the buyer-side verification cost. Thus, the buyer never benefits from disputing, and the seller's outcome-alignment design is payoff-irrelevant. After removing the alignment and dispute layer, type t's utility from item r is $\theta_t h(a_r) - p_r$. The remaining problem is the standard nonlinear-pricing (screening) problem with scalar type θ_t, quality $h(a)$, and transfer p (Mussa and Rosen, 1978). The usual one-dimensional screening argument implies that incentive-compatible allocations are monotone in the value order and that the optimal allocation solves the virtual-surplus program in (5). As usual, the lowest-θ type's individual-rationality constraint binds, and the adjacent downward incentive constraints bind; these binding constraints give the payment recursion in (6). Because the payment is independent of the realized outcome, success heterogeneity does not enter the UBP screening index.

We next consider the benchmark PFS pricing format under full outcome alignment. This benchmark isolates the effect of charging for success rather than usage before introducing the seller's ability to adjust the AI success metric. It sets up the comparison between optimal PFS and optimal UBP in Section 4.2, where we decompose the profit difference into two parts: the effect of charging for success rather than usage when the AI is fully aligned, and the additional gain from allowing the seller to adjust the AI success metric. Throughout, we use the superscript "fa" to denote full alignment.

Proposition 2 (Optimal PFS under full alignment)**.** *Under the separable primitive functions in*

Definition 2 and *full alignment, there exists an optimal* PFS *menu whose allocation is monotone in the ϕ-order, $a_{[1]}^{\mathsf{PFS,fa}} \le \cdots \le a_{[n]}^{\mathsf{PFS,fa}}$. This allocation solves*

$$R_{\mathsf{PFS}}^{\mathrm{fa}} = \max_{\substack{a_{[1]},\ldots,a_{[n]}\in\mathcal{A} \\ a_{[1]}\le\cdots\le a_{[n]}}} \sum_{t=1}^{n} \pi_{[t]} \left[\alpha_{[t]}\widetilde{\phi}_{[t]} h(a_{[t]}) - c(a_{[t]})\right], \tag{7}$$

where the PFS *virtual types are*

$$\widetilde{\phi}_{[t]} := \phi_{[t]} - \frac{\sum_{k=t+1}^{n} \pi_{[k]}\alpha_{[k]}}{\pi_{[t]}\alpha_{[t]}} \left(\phi_{[t+1]} - \phi_{[t]}\right), \quad t < n, \qquad \widetilde{\phi}_{[n]} := \phi_{[n]}. \tag{8}$$

The optimal success prices satisfy

$$\begin{aligned} s\left(a_{[1]}^{\mathsf{PFS,fa}}\right) p_{[1]}^{\mathsf{PFS,fa}} &= \phi_{[1]} h\left(a_{[1]}^{\mathsf{PFS,fa}}\right), \\ s\left(a_{[t]}^{\mathsf{PFS,fa}}\right) p_{[t]}^{\mathsf{PFS,fa}} &= s\left(a_{[t-1]}^{\mathsf{PFS,fa}}\right) p_{[t-1]}^{\mathsf{PFS,fa}} + \phi_{[t]} \left[h\left(a_{[t]}^{\mathsf{PFS,fa}}\right) - h\left(a_{[t-1]}^{\mathsf{PFS,fa}}\right)\right], \qquad t \ge 2. \end{aligned} \tag{9}$$

Proposition 2 shows that, under full alignment, finding the optimal PFS menu reduces to a one-dimensional screening problem with screening index $\phi_t = \theta_t/\alpha_t$. To see why, define the transformed payment $m_r := s(a_r)p_r$. A type-t buyer who chooses item r obtains utility

$$\theta_t h(a_r) - \alpha_t s(a_r) p_r = \alpha_t \left[\phi_t h(a_r) - m_r\right].$$

Since $\alpha_t > 0$, multiplying utility by $1/\alpha_t$ does not change the buyer's choice among menu items. The buyer's IC and IR constraints are therefore the same as those in a one-dimensional screening problem with type ϕ_t, quality $h(a)$, and transformed payment m. Hence, an optimal fully aligned PFS allocation is monotone in the ϕ-order. The profit objective, however, differs from standard screening because the seller collects payment only upon success. The seller's expected revenue from the transformed payment m_t is therefore scaled by the success-probability shifter α_t. This is why the objective in (7) weights each virtual per-success value $\widetilde{\phi}_{[t]}$ by $\alpha_{[t]}$, and why the virtual type in (8) contains the corresponding α-adjustment. Finally, given a monotone allocation in the ϕ-order, the transformed payments are pinned down by the standard binding constraints: the lowest-ϕ type's IR constraint binds, and the adjacent downward IC constraints bind. Translating back from $m_t = s(a_t)p_t$ gives the success-payment recursion in (9).

The fact that the seller collects payment only upon success creates a distinctive feature of PFS relative to UBP: *a buyer may accept a high success-contingent payment while generating little expected revenue for the seller.* This occurs when the buyer's value conditional on success is high but the task is difficult for the AI to resolve. The high success payment is acceptable to the buyer because it is paid infrequently; for the same reason, it yields limited expected revenue to the seller. This wedge between the buyer's per-success willingness to pay and the seller's expected revenue has a sharp worst-case implication.

Proposition 3 (Asymmetric worst-case robustness)**.** *Recall that $n = |\mathcal{T}|$ is the number of buyer types and that R_{SB} denotes the unrestricted second-best profit defined in* (3)*.* UBP *satisfies the*

following worst-case robustness bound:

$$R^*_{\mathsf{UBP}} \;\geq\; \frac{1}{n}\, R_{\mathsf{SB}} \;\geq\; \frac{1}{n}\, R^{\mathrm{fa}}_{\mathsf{PFS}}. \tag{10}$$

By contrast, fully aligned PFS *can perform arbitrarily poorly relative to* UBP *in the worst case: even with* $n = 2$ *types,*

$$\sup \; \frac{R^*_{\mathsf{UBP}}}{R^{\mathrm{fa}}_{\mathsf{PFS}}} \;=\; +\infty, \tag{11}$$

where the supremum is over model primitives. The same unbounded worst-case gap holds even for optimal PFS *with an optimally chosen AI success metric, and for any fixed buyer- and seller-side dispute costs.*

Proposition 3 isolates an asymmetric robustness property. The bound in (10) follows from a simple targeting argument. Because the UBP payment is independent of the realized outcome, the seller can post a single usage payment aimed at the type with the largest first-best profit contribution. This guarantees at least a $1/n$ share of first-best profit, and therefore at least a $1/n$ share of the second-best benchmark. Fully aligned PFS lacks this guarantee because buyer acceptance and seller profitability need not move together. A buyer accepts a menu item based on her per-success value ϕ, but the seller collects that payment only when success occurs and still incurs the deployment cost $c(a)$. Hence, a type with a low success probability can have high ϕ, accept a high success-contingent payment, and nevertheless generate little or even negative expected profit for the seller. The worst-case statement in (11) formalizes this failure: under PFS, serving the profitable segment can require also serving a low-success segment, so the optimal fully aligned PFS policy can become nearly worthless because deployment costs are incurred whenever service is provided. By contrast, UBP can still profitably target the high-value segment. The same worst-case logic extends to optimal PFS with strategic outcome-alignment design.

4.2 UBP–PFS Comparison

Having characterized optimal UBP (Proposition 1) and optimal PFS under full alignment (Proposition 2), we now compare PFS and UBP when the seller can also choose how closely the AI success metric is aligned with the buyer's notion of task completion. Let $R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S)$ denote the optimal PFS profit with this alignment and dispute layer, that is, the value of (3) under the additional restriction $q_t = 0$ for all t. Under UBP, the seller's outcome-alignment design is payoff-irrelevant, so optimal UBP remains the one characterized in Proposition 1. The profit difference can be decomposed as

$$R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S) - R^*_{\mathsf{UBP}} = \underbrace{R^{\mathrm{fa}}_{\mathsf{PFS}} - R^*_{\mathsf{UBP}}}_{\text{fully aligned payment-base gap}} + \underbrace{R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S) - R^{\mathrm{fa}}_{\mathsf{PFS}}}_{\text{outcome-misalignment gain}}. \tag{12}$$

The first term reflects the *payment-base effect*: it compares charging for success with charging for usage when the AI is fully aligned. The second term indicates the *outcome-alignment* effect: it measures the additional profit from allowing the seller, under PFS, to choose how the AI's success criterion is aligned with the buyer's value-relevant notion of task completion. This second term is weakly positive because full alignment remains feasible. Let $R^{\text{fa}}_{\mathsf{LL}}$ denote the optimal limited-liability profit under full outcome alignment: the value of (3) when the AI-reported outcome coincides with the buyer-value-relevant true outcome, i.e., $\sigma^1_t = 1$ and $\sigma^0_t = 0$ for all t, and all success and failure payments are nonnegative, $p_t \geq 0$ and $q_t \geq 0$.

Theorem 1 (Comparison of UBP and PFS)**.** *The following statements hold.*

(i) (Outcome-alignment effect.) *For every* $\kappa_B, \kappa_S \geq 0$,

$$R^{\text{fa}}_{\mathsf{PFS}} \leq R^{*}_{\mathsf{PFS}}(\kappa_B, \kappa_S) \leq R^{\text{fa}}_{\mathsf{LL}}.$$

The optimal PFS *profit with outcome-alignment design,* $R^{*}_{\mathsf{PFS}}(\kappa_B, \kappa_S)$*, is weakly increasing in the buyer-side dispute cost* κ_B *and weakly decreasing in the seller-side dispute cost* κ_S*. Moreover:*

(a) If buyer disputes are costless, i.e., $\kappa_B = 0$*, then*

$$R^{*}_{\mathsf{PFS}}(0, \kappa_S) = R^{\text{fa}}_{\mathsf{PFS}}.$$

(b) There exists $\bar{\kappa}_B$ *such that, for every* $\kappa_B \geq \bar{\kappa}_B$*,*

$$R^{*}_{\mathsf{PFS}}(\kappa_B, \kappa_S) = R^{\text{fa}}_{\mathsf{LL}} \geq R^{*}_{\mathsf{UBP}}.$$

(ii) (Payment-base comparison under full alignment.) *Under the separable primitives in Definition 2, the following baseline comparisons hold.*

(a) (Tail-average success condition.) *Suppose the ordering by per-success value coincides with the ordering by value, i.e.,* $[t] = (t)$ *for every rank* t*, and define the value-order tail average*

$$\bar{\alpha}^{\text{tail}}_{(t)} := \frac{\sum_{k=t}^{n} \pi_{(k)} \alpha_{(k)}}{\sum_{k=t}^{n} \pi_{(k)}}, \qquad t = 1, \ldots, n.$$

If $\bar{\alpha}^{\text{tail}}_{(t)} \geq \alpha_{(t)}$ *for every* t*, then* $R^{\text{fa}}_{\mathsf{PFS}} \geq R^{*}_{\mathsf{UBP}}$*; if* $\bar{\alpha}^{\text{tail}}_{(t)} \leq \alpha_{(t)}$ *for every* t*, then* $R^{\text{fa}}_{\mathsf{PFS}} \leq R^{*}_{\mathsf{UBP}}$*.*

(b) (Power relation.) *In the power family* $\alpha_t = \theta_t^{\beta}$ *with* $\theta_t > 0$ *for all* t*,*[8] *the optimal* UBP *profit is independent of* β*, while* $R^{\text{fa}}_{\mathsf{PFS}}(\beta)$ *is weakly increasing for* $\beta \leq 1$ *and weakly decreasing for* $\beta \geq 1$*. In particular,* $R^{\text{fa}}_{\mathsf{PFS}}(\beta) \leq R^{*}_{\mathsf{UBP}}$ *for* $\beta < 0$*,* $R^{\text{fa}}_{\mathsf{PFS}}(0) = R^{*}_{\mathsf{UBP}}$*, and* $R^{\text{fa}}_{\mathsf{PFS}}(\beta) \geq R^{*}_{\mathsf{UBP}}$ *for* $0 < \beta \leq 1$*.*

Part (i) isolates the outcome-alignment effect in (12). Allowing the seller to choose the AI success metric weakly improves on fully aligned PFS, but the resulting profit cannot exceed the

[8] We restrict β to the range in which $\alpha_t(\beta)s(a) \in [0,1]$ on the relevant allocation set, so that γ_t remains a probability.

fully aligned limited-liability benchmark. The comparative statics reflect the costs of the dispute process on both sides: increasing κ_B makes reported successes harder for the buyer to challenge and expands the seller's scope to use a buyer-misaligned success metric, whereas increasing κ_S makes disputed reports more costly to the seller. In the extreme case in which $\kappa_B = 0$, any reported success that does not correspond to a true success can be challenged at no cost, so every misaligned PFS item is weakly dominated by a fully aligned PFS item with the same allocation and an appropriately scaled success payment.

For large κ_B, reported successes are hard for the buyer to challenge. Part (i)(b) shows that, once buyer verification is sufficiently costly, optimal PFS with a strategically chosen success metric reaches the fully aligned limited-liability two-price benchmark. The sharper message is that *a single success payment, combined with outcome-alignment design, can replicate the same profit as fully outcome-contingent pricing with a fully aligned metric.* Since UBP is the special case of a two-price contract with equal success and failure payments, and outcome-alignment design is payoff-irrelevant under UBP, PFS weakly dominates optimal UBP in this high-κ_B regime. The proof establishes the equivalence item by item. Take any fully aligned limited-liability item with nonnegative success and failure payments $(p^{\mathsf{LL}}, q^{\mathsf{LL}})$. The seller can set a single PFS success payment $p^{\mathsf{PFS}} = \max\{p^{\mathsf{LL}}, q^{\mathsf{LL}}\}$ and choose the AI reporting probabilities so that this payment is triggered with probability $\sigma^1 = p^{\mathsf{LL}}/p^{\mathsf{PFS}}$ after true success and probability $\sigma^0 = q^{\mathsf{LL}}/p^{\mathsf{PFS}}$ after true failure. If κ_B is at least as large as every constructed success payment p^{PFS} in the menu, the dispute condition cannot hold, so no type disputes a reported success. The induced expected payment is therefore exactly the same as under the original two-price item, for every success probability, which establishes the equivalence. The nonnegativity of p^{LL} and q^{LL} is what makes the construction feasible: because $p^{\mathsf{PFS}} = \max\{p^{\mathsf{LL}}, q^{\mathsf{LL}}\}$, both ratios $p^{\mathsf{LL}}/p^{\mathsf{PFS}}$ and $q^{\mathsf{LL}}/p^{\mathsf{PFS}}$ lie in $[0, 1]$. Thus, one success payment plus two reporting probabilities can replicate two nonnegative state-contingent payments, while preserving allocations, buyer costs, incentive constraints, and seller profit.

Part (ii) switches off the outcome-alignment layer and compares the two payment bases directly. Its main message is that fully aligned PFS becomes more attractive when heterogeneity in per-success values is small and when higher per-success values are associated with higher success probabilities. The first force weakens the screening distortion under PFS, because $\phi_t = \theta_t/\alpha_t$ is the relevant private-information dimension. The second force means that PFS charges on the right margin: high per-success willingness to pay comes together with a high probability that the seller collects the success payment.

The tail-average condition in Part (ii)(a) makes this intuition precise. When the ordering by per-success value coincides with the ordering by value, fully aligned PFS dominates UBP if each type's success-probability shifter $\alpha_{(t)}$ is no larger than the π-weighted tail average $\bar{\alpha}^{\text{tail}}_{(t)}$; the comparison reverses if each $\alpha_{(t)}$ is no smaller than its tail average. A simple sufficient condition for these tail-average inequalities is global monotonicity of α in this common order. In particular, if $\alpha_{(1)} \leq$

$\alpha_{(2)} \leq \cdots \leq \alpha_{(n)}$, then $\bar{\alpha}^{\text{tail}}_{(t)} \geq \alpha_{(t)}$ for every t. In this case, success-probability heterogeneity partly offsets value heterogeneity in the per-success-value dimension, and buyers with higher willingness to pay per successful completion are also those whose tasks the AI is more likely to complete. Hence fully aligned PFS outperforms UBP: $R^{\text{fa}}_{\mathsf{PFS}} \geq R^{*}_{\mathsf{UBP}}$. Conversely, if $\alpha_{(1)} \geq \alpha_{(2)} \geq \cdots \geq \alpha_{(n)}$, then $\bar{\alpha}^{\text{tail}}_{(t)} \leq \alpha_{(t)}$ for every t. In this case, success-probability heterogeneity reinforces value heterogeneity in the per-success-value dimension, while success probabilities are higher for buyers with lower willingness to pay per successful completion. Hence $R^{\text{fa}}_{\mathsf{PFS}} \leq R^{*}_{\mathsf{UBP}}$.

Part (ii)(b) gives a one-parameter illustration. In the power family $\alpha_t(\beta) = \theta_t^{\beta}$, β controls how strongly success probability is tied to value. The optimal UBP profit is unchanged as β varies because β changes only success probabilities, whereas UBP payments do not depend on task outcomes. Fully aligned PFS, however, screens on

$$\phi_t(\beta) = \frac{\theta_t}{\alpha_t(\beta)} = \theta_t^{1-\beta}.$$

Thus the per-success-value order coincides with the value order whenever $\beta \leq 1$. Within this range, the comparison with UBP depends on how success probabilities vary along this common order. If $\beta < 0$, then $\alpha_t(\beta) = \theta_t^{\beta}$ is decreasing in the common order, so the tail-average condition in Part (ii)(a) implies $R^{\text{fa}}_{\mathsf{PFS}}(\beta) \leq R^{*}_{\mathsf{UBP}}$. If $\beta = 0$, then $\alpha_t(\beta)$ is constant across types, so $R^{\text{fa}}_{\mathsf{PFS}}(0) = R^{*}_{\mathsf{UBP}}$. If $0 < \beta \leq 1$, then $\alpha_t(\beta)$ is increasing in the common order, so the tail-average condition implies $R^{\text{fa}}_{\mathsf{PFS}}(\beta) \geq R^{*}_{\mathsf{UBP}}$. As β rises toward one, heterogeneity in per-success values shrinks and success probabilities become more concentrated among higher per-success-value types, making fully aligned PFS more profitable. At $\beta = 1$, $\phi_t = \theta_t/\alpha_t = 1$ for all t. Buyers therefore no longer differ in the screening dimension relevant for fully aligned PFS, allowing the seller to choose the first-best allocation and extract the first-best surplus. For $\beta > 1$, the per-success-value order reverses relative to the value order, so the tail-average comparison in Part (ii)(a) no longer applies directly. In this region, $R^{\text{fa}}_{\mathsf{PFS}}(\beta)$ is weakly decreasing in β, but its level relative to R^{*}_{UBP} is not pinned down without additional conditions.

To summarize, Theorem 1 gives two distinct routes by which PFS can outperform UBP. The first is the fully aligned payment-base route: fully aligned PFS is strong when heterogeneity in the PFS screening index, the per-success value ϕ, is limited and when success probabilities are higher for buyers with higher ϕ. The second is the outcome-alignment route: strategic AI misalignment becomes more powerful as buyer-side verification becomes costly and less powerful as seller-side dispute costs rise. When $\kappa_B = 0$, metric design adds no value beyond fully aligned PFS. When κ_B is sufficiently large, misaligned PFS can implement the fully aligned limited-liability two-price benchmark, and therefore weakly dominates optimal UBP.

4.3 Application: Pricing AI in Customer Service

We now consider a customer-service setting in which the buyer is a downstream firm that deploys AI agents as the front end for incoming customer requests. Each request first reaches the AI layer, which attempts to resolve it immediately with no delay. If the AI succeeds, the request exits the system; if it fails, the request is escalated to the downstream human service queue. Thus, AI success determines both the outcome relevant for pay-for-success pricing and the reduction in workload faced by human staff. This is exactly the dynamic foundation introduced in Section 3: the ongoing stream of requests generates an infinite sequence of binary resolution outcomes, each request is billed on its reported outcome, and the buyer's per-task value is the steady-state expected reduction in congestion cost.

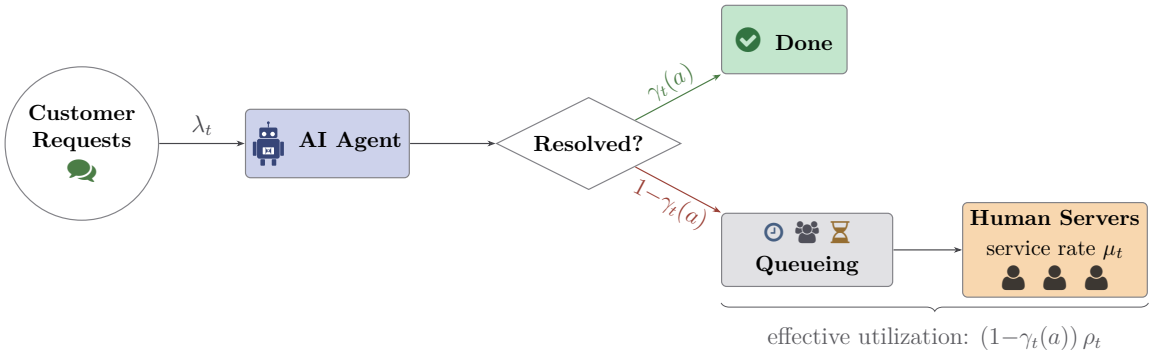


Figure 2: Customer-service flow with AI front end

Notes. Incoming customer requests first arrive at the AI agent. The AI layer can process all arrivals immediately, so it introduces no queue of its own. For a type-t buyer (the downstream firm), a request is resolved by the AI of capability a with probability $\gamma_t(a)$ and exits the system; otherwise, it is escalated to the downstream human queue. The buyer has arrival rate λ_t and N_t human servers, each serving at rate μ_t, so baseline human utilization is $\rho_t = \lambda_t/(N_t\mu_t)$. AI adoption reduces the fraction of requests handled by humans to $1-\gamma_t(a)$, and therefore reduces human utilization to $(1-\gamma_t(a))\rho_t$. The buyer's value is the resulting reduction in congestion cost.

Figure 2 summarizes the operational flow. For tractability, we restrict the application to two buyer types and one AI capability level, $n = 2$ and $|\mathcal{A}| = 1$.[9] Let c_0 denote the deployment cost of this capability, and let $\gamma_t \in (0,1)$ be type t's AI success probability. Before AI adoption, type t has arrival rate λ_t, N_t identical human servers, each with service rate μ_t, and baseline utilization $\rho_t = \lambda_t/(N_t\mu_t) < 1$. Since only unresolved requests enter the human queue, AI adoption reduces utilization to $(1-\gamma_t)\rho_t$.

Let $V(N,\lambda,\rho,\gamma)$ denote the value of AI adoption for a downstream firm with N human servers, arrival rate λ, baseline utilization ρ, and AI success probability γ. This value is measured as the reduction in congestion cost. We assume standard congestion monotonicity: holding N, λ, and γ fixed, V is continuous and strictly increasing in ρ, with $V \to 0$ as $\rho \downarrow 0$ and $V \to \infty$ as $\rho \uparrow 1$. Holding N, λ, and ρ fixed, V is weakly increasing in the AI success probability γ. Thus, type t's

[9] The same insights extend to a continuous capability set $\mathcal{A}$ under the binary-type assumption. We use a single capability here to keep the exposition simple.

value is $v_t = V(N_t, \lambda_t, \rho_t, \gamma_t)$. A concrete example satisfying these properties is an $M/M/N_t$ service system for a type-t buyer, with Poisson arrival rate λ_t, N_t exponential human servers, baseline utilization ρ_t, and per-unit-time waiting cost $h_0 > 0$. The buyer's value is the expected reduction in waiting cost from deploying the AI:

$$V(N_t, \lambda_t, \rho_t, \gamma_t) = h_0 \left[\frac{\mathsf{C}(N_t, \rho_t)\rho_t}{\lambda_t(1-\rho_t)} - \frac{\mathsf{C}(N_t, (1-\gamma_t)\rho_t)(1-\gamma_t)\rho_t}{\lambda_t[1-(1-\gamma_t)\rho_t]} \right],$$

where $\mathsf{C}(N, \rho)$ is the Erlang-C formula.

We focus on the payment-base comparison, so we impose full outcome alignment and abstract from the outcome-alignment force. By Theorem 1(i), allowing the seller to choose the AI success metric can only weakly increase PFS profit.

Corollary 1 (UBP–PFS Comparison in Customer Service)**.** *Suppose there are two buyer types and one AI capability. Relabel types so that $\gamma_2 > \gamma_1$. Fix all other primitives and vary only type 2's baseline utilization ρ_2. Then there is a cutoff $\rho^\dagger \in (0,1)$ such that*

$$\rho_2 > \rho^\dagger \quad \Longrightarrow \quad R^{\text{fa}}_{\mathsf{PFS}} \geq R^{*}_{\mathsf{UBP}},$$

and

$$\rho_2 \leq \rho^\dagger \quad \Longrightarrow \quad R^{*}_{\mathsf{UBP}} \geq R^{\text{fa}}_{\mathsf{PFS}}.$$

Corollary 1 translates the payment-base comparison in Theorem 1(ii) into queueing primitives. The theorem shows that fully aligned PFS is favored when high per-success willingness to pay is paired with a high probability of success. In the customer-service application, type 2 is the AI-easy type because $\gamma_2 > \gamma_1$. Holding the other primitives fixed, increasing ρ_2 raises type 2's value from AI adoption, since AI removes more congestion when the downstream human system is more heavily loaded. Thus, when ρ_2 is above the cutoff $\rho^\dagger$, the type that AI resolves more often also obtains sufficiently high value from the reduction in human congestion, and fully aligned PFS weakly outperforms UBP; below the cutoff, UBP weakly outperforms PFS. The proof gives the explicit cutoff $\rho^\dagger$.

Figure 3 visualizes the cutoff comparison in Corollary 1 in service-rate space. Holding γ_t, λ_t, and N_t fixed, the service rates μ_t determine utilizations through $\rho_t = \lambda_t/(N_t\mu_t)$; lower μ_t means higher utilization. For each value of μ_1, the cutoff in ρ_2 translates into a cutoff in μ_2: when type 2's human service rate is sufficiently low, PFS weakly outperforms UBP; otherwise, UBP weakly outperforms PFS. The color scale additionally reports the magnitude of the comparison as a profit ratio, not only its sign.

The figure highlights how the relative strengths of AI and human service—that is, human–AI substitutability and complementarity—shape the preferred pricing format. Since $\gamma_2 > \gamma_1$, type 2 is the task class on which the AI performs better. If humans are also relatively strong on type 2 tasks, the strengths of AI and human service overlap—AI and human servers are *substitutes*—and UBP is

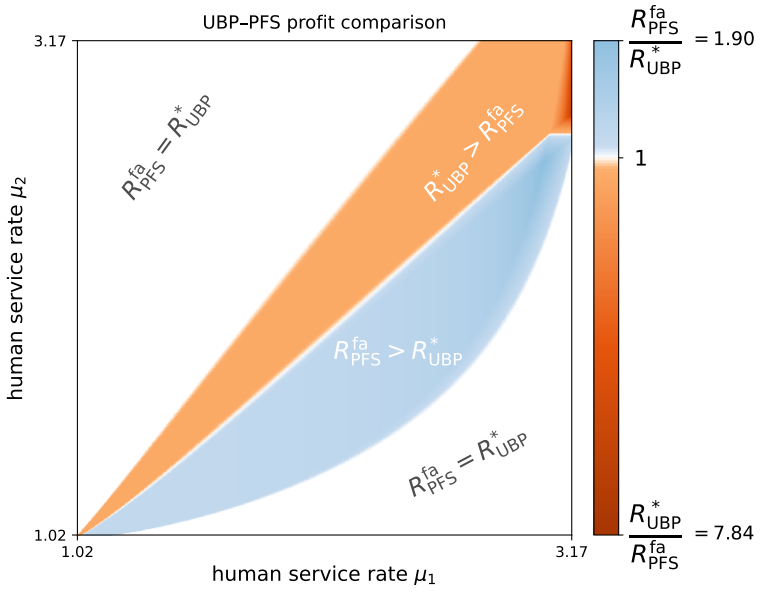


Figure 3: UBP–PFS comparison in service-rate space

Notes. The figure specializes the customer-service model to $M/M/1$ human queues. It fixes $\lambda_1 = \lambda_2 = 1$, $N_1 = N_2 = 1$, $h_0 = 1$, $(\gamma_1, \gamma_2) = (0.40, 0.80)$, $(\pi_1, \pi_2) = (0.90, 0.10)$, and $c_0 = 0.10$, and varies $(\mu_1, \mu_2) \in [1.02, 3.17]^2$. Colors report the multiplicative advantage of the better-performing format—the ratio $R^*_{\mathsf{UBP}}/R^{\mathrm{fa}}_{\mathsf{PFS}}$ on the orange side and $R^{\mathrm{fa}}_{\mathsf{PFS}}/R^*_{\mathsf{UBP}}$ on the blue side— white regions indicate equal optimal profits.

more attractive. If humans are relatively weak on type 2 tasks—AI and human servers are *complements*—the AI is effective precisely where the human system is more congested, making PFS more attractive. Beyond locating the dominance regions, the color scale also illustrates the asymmetric worst-case robustness in Proposition 3. Within the displayed parameters, the multiplicative advantage of UBP, $R^*_{\mathsf{UBP}}/R^{\mathrm{fa}}_{\mathsf{PFS}}$, reaches 7.84, whereas that of fully aligned PFS, $R^{\mathrm{fa}}_{\mathsf{PFS}}/R^*_{\mathsf{UBP}}$, reaches only 1.90.

5 Optimal Mechanisms

The previous section shows that strategic design of the AI success metric can increase the seller's profit under PFS. When buyer-side verification is sufficiently costly, PFS with an appropriately chosen metric can reproduce the fully aligned limited-liability two-price benchmark; see Theorem 1(i)(b). This possibility creates a central credibility concern for outcome-based pricing: the seller may profit from defining success differently from the buyer, making buyers wary of adopting such contracts in practice. This section returns to the unrestricted problem (3). In Section 5.1, we show that the misalignment motive disappears once the seller can charge separate success and failure fees, equivalently a baseline fee plus a success premium. Any feasible menu can be replaced, item by item, by a fully aligned two-price menu that preserves each assigned type's expected cost, makes deviations weakly less attractive, and eliminates dispute waste. Full alignment is therefore without loss in the unrestricted problem. In Section 5.2, we characterize the optimal unrestricted

contract.

5.1 Two-Price Mechanisms Eliminate the Misalignment Motive

Theorem 2 (Full Outcome Alignment is Without Loss)**.** *Problem* (3) *admits an optimal solution with full outcome alignment: for every menu item* $t \in \mathcal{T}$*,* $\sigma_t^1 = 1$ *and* $\sigma_t^0 = 0$*.*

Theorem 2 resolves the credibility concern raised by the PFS analysis. Under PFS, a buyer-misaligned success metric can increase profit when verification is costly, suggesting that seller-defined success metrics are inherently difficult to trust. The theorem shows that this motive arises from PFS's restriction to a single success payment, rather than from outcome-based pricing itself. Misalignment allows the seller to impose indirectly a payment after some true failures. Once the contract can charge separate success and failure payments, the same incentive effects can be achieved under full alignment. The replacement also eliminates the disputes through which misalignment operates. Thus, in the unrestricted contract space, metric misalignment is weakly dominated by transparent two-price contracting: it creates no additional contracting power and wastes resources. The broader takeaway is that the apparent outcome-alignment problem is, in this setting, a consequence of an insufficiently rich contract. Explicit outcome-contingent prices make the success metric credible without reducing the seller's profit.[10]

It is useful to distinguish this result from the revelation principle. The revelation principle (Myerson, 1979) concerns the buyer's message space: without loss of generality, the mechanism can ask the buyer to report her private type directly. Here, by contrast, the issue is the seller's own rule for mapping true task outcomes into AI-reported outcomes. Theorem 2 shows that, in the unrestricted two-price contract space, the seller does not need to distort the AI success metric.

Proof Sketch. Fix any feasible menu $\mathcal{M} = (a_t, p_t, q_t, \sigma_t^1, \sigma_t^0)_{t\in\mathcal{T}}$, possibly with an outcome-alignment rule that differs from the buyer's one. We construct a fully aligned menu $\widetilde{\mathcal{M}} = (a_t, \widetilde{p}_t, \widetilde{q}_t, 1, 0)_{t\in\mathcal{T}}$ item by item, keeping the allocation profile fixed. The goal is to choose success and failure payments that keep the assigned type's expected buyer cost unchanged, make deviations weakly more costly, and weakly increase the seller's profit.

For item t, fix the allocation a_t. Let $C_t(\gamma)$, defined in (1), be the expected cost of choosing item t in the original menu $\mathcal{M}$ for a buyer whose success probability at a_t is γ, after she optimally decides whether to pay or dispute each AI-reported outcome. For any fixed pay-or-dispute rule, the expected cost is affine in γ. Since $C_t(\cdot)$ is the lower envelope of these affine costs, it is concave. Hence there exists an affine line that coincides with C_t at the assigned type's success probability and lies weakly above C_t everywhere:

$$\ell_t(\gamma_t(a_t)) = C_t(\gamma_t(a_t)), \qquad \ell_t(\gamma) \geq C_t(\gamma) \quad \forall \gamma \in [0,1].$$

[10] The conclusion is unchanged under limited liability, i.e., when problem (3) additionally requires all payments to be nonnegative, $p_t, q_t \geq 0$ for all t.

Write this line as

$$\ell_t(\gamma) = \widetilde{q}_t + (\widetilde{p}_t - \widetilde{q}_t)\gamma, \qquad \widetilde{q}_t := \ell_t(0), \quad \widetilde{p}_t := \ell_t(1).$$

In the replacement menu $\widetilde{\mathcal{M}}$, item t keeps allocation a_t, uses the fully aligned report $\widehat{\omega} = \omega$, charges $\widetilde{q}_t$ after failure, and charges $\widetilde{p}_t$ after success.

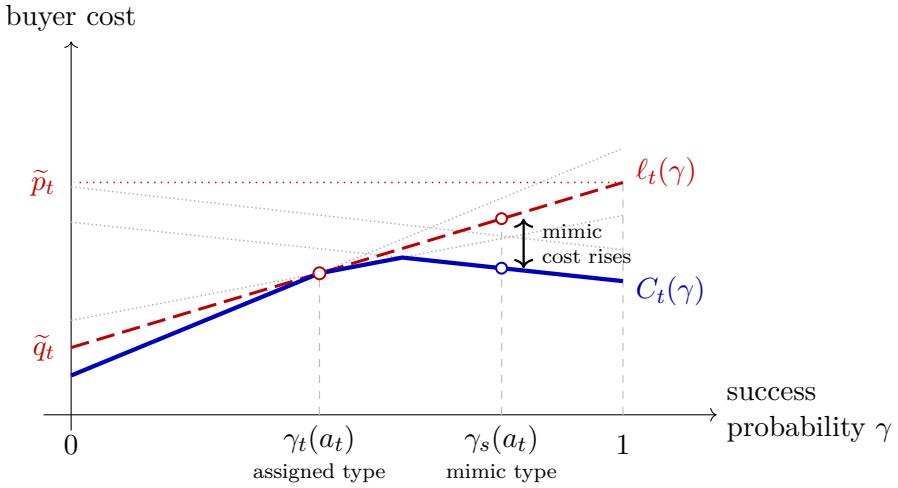


Figure 4: Buyer cost from choosing item t under $\mathcal{M}$ and $\widetilde{\mathcal{M}}$

Notes. The thin gray dotted lines are the buyer-cost functions generated by fixed pay-or-dispute rules; each is affine in the success probability γ. The blue curve $C_t(\gamma)$ is their lower envelope. The red supporting line ℓ_t coincides with C_t at $\gamma_t(a_t)$ and lies weakly above it elsewhere. Its value at $\gamma = 0$ and $\gamma = 1$ define the failure and success payments, $\widetilde{q}_t$ and $\widetilde{p}_t$, in the replacement menu $\widetilde{\mathcal{M}}$.

Figure 4 illustrates the replacement. The blue curve is the original buyer-cost function $C_t(\gamma)$, and the red dashed line is the supporting affine line ℓ_t. At the assigned success probability $\gamma_t(a_t)$, the two curves coincide; hence type t's expected cost is unchanged. At a mimicking success probability $\gamma_s(a_t)$, the red point lies weakly above the blue point. Thus, deviating to item t is weakly more costly for type s than under the original menu. Applying this construction to every item preserves IR and IC. Because the replacement uses direct outcome-contingent payments under a fully aligned AI, the buyer never disputes: buyer-side dispute costs from the original menu are converted into payments, and seller-side dispute costs disappear. Hence the seller's profit is weakly higher item by item. □

5.2 Characterizing the Optimal Mechanism

By Theorem 2, we can restrict attention to menus with full outcome alignment without loss of optimality. In such menus, the buyer never disputes, and the seller's problem becomes

$$\max_{\{a_t,p_t,q_t\}_{t\in\mathcal{T}}} R_{\mathsf{SB}} = R^{\mathrm{fa}}_{\mathsf{SB}} := \sum_{t\in\mathcal{T}} \pi_t \Big[(1-\gamma_t(a_t))q_t + \gamma_t(a_t)p_t - c(a_t)\Big] \tag{13}$$

$$\text{s.t.} \quad v_t(a_t) - \big[(1-\gamma_t(a_t))q_t + \gamma_t(a_t)p_t\big] \geq v_t(a_r) - \big[(1-\gamma_t(a_r))q_r + \gamma_t(a_r)p_r\big], \quad \forall t, r \in \mathcal{T}, \quad (\mathsf{IC})$$

$$v_t(a_t) - \big[(1-\gamma_t(a_t))q_t + \gamma_t(a_t)p_t\big] \geq 0, \quad \forall t \in \mathcal{T}. \tag{IR}$$

Problem (13) differs from a standard screening problem because each menu item includes not only an allocation but also two outcome-contingent prices. We first reduce these three decision vectors to two: allocations and buyer rents. For each type t, define the buyer rent as the expected utility type t obtains when choosing her assigned item t:

$$u_t := v_t(a_t) - \big[(1-\gamma_t(a_t))q_t + \gamma_t(a_t)p_t\big]. \tag{14}$$

Thus, the seller's expected payment from type t is $v_t(a_t) - u_t$. Let $\big((a_t, u_t)\big)_{t\in\mathcal{T}}$ denote the allocation-rent profile. In terms of this allocation-rent profile, problem (13) becomes

$$\max_{\substack{a_t\in\mathcal{A},\, u_t\geq 0 \\ t\in\mathcal{T}}} \quad \sum_{t\in\mathcal{T}} \pi_t\big[v_t(a_t) - u_t - c(a_t)\big] \tag{15}$$

$$\text{s.t.} \quad \big((a_t, u_t)\big)_{t\in\mathcal{T}} \text{ is implementable.} \tag{16}$$

Here, $\big((a_t, u_t)\big)_{t\in\mathcal{T}}$ is implementable if there exist success and failure prices $(p_t, q_t)_{t\in\mathcal{T}}$ that generate the rents $(u_t)_{t\in\mathcal{T}}$ from the allocations $(a_t)_{t\in\mathcal{T}}$ through (14) and satisfy the IC constraints (IC). The IR constraints (IR) are captured by the restrictions $u_t \geq 0$ for all $t \in \mathcal{T}$.

We next translate implementability into geometric conditions in the γ-payment plane. We say that a point (x_t, y_t) in a finite set $\{(x_r, y_r) : r \in \mathcal{T}\} \subset \mathbb{R}^2$ is *upper-supported* if there exists an affine line that passes through (x_t, y_t) and lies weakly above every point in the set.[11]

Lemma 1 (Allocation-rent implementability)**.** *An allocation-rent profile* $\big((a_t, u_t)\big)_{t\in\mathcal{T}} \in (\mathcal{A}\times\mathbb{R}_+)^n$ *is implementable* if and only if *the following condition holds: for every item* $t \in \mathcal{T}$*, the intended point* $(\gamma_t(a_t),\, v_t(a_t) - u_t)$ *is* upper-supported *in the point set* $\{(\gamma_r(a_t),\, v_r(a_t) - u_r) : r \in \mathcal{T}\}$.

Lemma 1 shows that implementability is equivalent to an upper-support condition. To see why, fix the menu item intended for type t. Its success and failure prices generate an affine expected-payment line in the success probability, $\ell_t(\gamma) = q_t + (p_t - q_t)\gamma$. The rent defined in (14) pins down the expected payment of type t when she chooses her assigned item t, so the payment line must satisfy

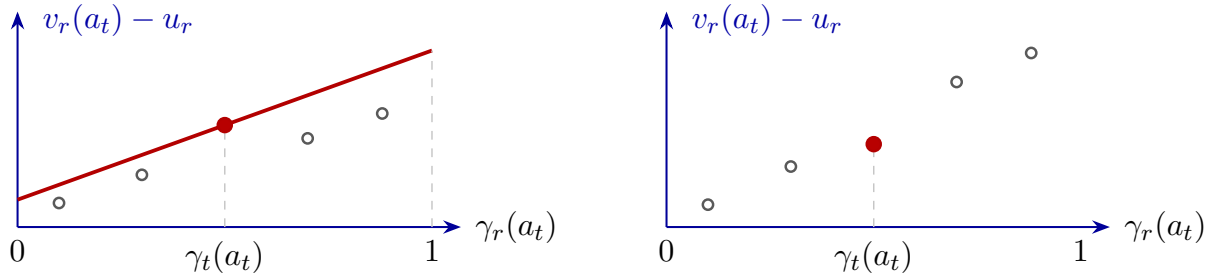
$$\ell_t(\gamma_t(a_t)) = v_t(a_t) - u_t.$$

Equivalently, the line passes through the intended point $(\gamma_t(a_t), v_t(a_t) - u_t)$. If type r instead chooses item t, her deviation payoff is $v_r(a_t) - \ell_t(\gamma_r(a_t))$, where $\ell_t(\gamma_r(a_t))$ is the expected payment she faces from item t. Thus, the IC constraint that prevents type r from deviating to item t, $u_r \geq v_r(a_t) - \ell_t(\gamma_r(a_t))$, is equivalent to

$$\ell_t(\gamma_r(a_t)) \geq v_r(a_t) - u_r. \tag{17}$$

[11] Formally, there exist $w, d \in \mathbb{R}$ such that $w + dx_t = y_t$ and $w + dx_r \geq y_r$ for all $r \in \mathcal{T}$.

Hence, implementability requires an affine line that passes through the intended point $(\gamma_t(a_t), v_t(a_t) - u_t)$ and lies weakly above every point $(\gamma_r(a_t), v_r(a_t) - u_r)$ generated by a possible mimicking type r. This is exactly the upper-support condition. Conversely, any such supporting line can be implemented by setting $q_t = \ell_t(0)$ and $p_t = \ell_t(1)$. Figure 5 illustrates the geometry: panel (a) is implementable for item t because such a line exists, whereas panel (b) is not.



(a) $(\gamma_t(a_t), v_t(a_t) - u_t)$ is upper-supported

(b) $(\gamma_t(a_t), v_t(a_t) - u_t)$ is not upper-supported

Figure 5: Upper-support condition for item t

Notes. The red point is the intended point for type t. In panel (a), the red line passes through this point and lies weakly above the points corresponding to all mimicking types; hence the red point is upper-supported. In panel (b), no affine line through the red point can lie weakly above all such points.

Remark 1 (Any positive-success allocation profile is implementable)**.** *In standard one-dimensional screening—*UBP *in our setting—the main implementability result is on allocations. Under single crossing, an allocation profile* $(a_t)_{t\in\mathcal{T}} \in \mathcal{A}^n$ *is implementable* if and only if *it is monotone in the buyer's valuation parameter (Mussa and Rosen, 1978; Maskin and Riley, 1984). Also, in standard one-dimensional screening, characterizing implementable allocations is enough: once a monotone allocation profile is fixed, the incentive-binding pattern is known, i.e., the lowest type's IR constraint and the local downward IC constraints binding, and pins down the optimal rents.*

Our setting differs in both respects. Unlike in standard screening, the binding pattern of the constraints is not known and can be intricate. Thus, we need to characterize the implementability of allocation-rent profiles, not only of the allocations themselves, since the seller's objective function depends on both; see (15)*. In addition, in contrast to standard screening without outcomes, in our setting, under mild conditions (see Lemma A-3 in the appendix),* any *positive-success allocation profile can be supported by some nonnegative rents and outcome-contingent prices and is therefore implementable.*

Define the first-best profit as

$$R_{\mathsf{FB}} := \max_{\{a_t\}_{t\in\mathcal{T}}} \sum_{t\in\mathcal{T}} \pi_t \big[v_t(a_t) - c(a_t)\big],$$

and let $(a_t^{\mathsf{FB}})_{t\in\mathcal{T}}$ denote a first-best allocation profile. We next characterize when the seller can achieve this first-best benchmark.

Corollary 2 (First-best extraction)**.** *The following statements hold.*

(i) The first-best profit is achieved, i.e., $R_{\mathsf{SB}} = R_{\mathsf{FB}}$*,* if and only if *there exists a first-best allocation profile* $(a_t^{\mathsf{FB}})_{t\in\mathcal{T}}$ *such that, for every type* $t \in \mathcal{T}$*, the point* $(\gamma_t(a_t^{\mathsf{FB}}), v_t(a_t^{\mathsf{FB}}))$ *is upper-supported in the set* $\{(\gamma_r(a_t^{\mathsf{FB}}), v_r(a_t^{\mathsf{FB}})) : r \in \mathcal{T}\}$*.*

(ii) Under the separable primitive functions in Definition 2, with $\alpha_1 < \cdots < \alpha_n$*, if* θ *is discretely concave in* α*,*[12] *then* $R_{\mathsf{SB}} = R_{\mathsf{FB}}$*.*

Part (i) applies the support-line characterization in Lemma 1 to the candidate allocation-rent profile $((a_t^{\mathsf{FB}}, 0))_{t\in\mathcal{T}}$. First-best profit is achievable when some first-best allocation profile can be implemented with zero rents. In the upper-support geometry, this means that, for each item t, an expected-payment line passes through $(\gamma_t(a_t^{\mathsf{FB}}), v_t(a_t^{\mathsf{FB}}))$ and lies weakly above the corresponding points for all other types.

Part (ii) gives a primitive sufficient condition for the allocation-rent profile $((a_t^{\mathsf{FB}}, 0))_{t\in\mathcal{T}}$ to be implementable. Under Definition 2, the point set $\{(\gamma_r(a_t^{\mathsf{FB}}), v_r(a_t^{\mathsf{FB}})) : r \in \mathcal{T}\}$ becomes $\{(\alpha_r s(a_t^{\mathsf{FB}}), \theta_r h(a_t^{\mathsf{FB}})) : r \in \mathcal{T}\}$. Up to positive horizontal and vertical rescaling, this set has the same geometry as the primitive type set $\{(\alpha_r, \theta_r) : r \in \mathcal{T}\}$. Because positive rescaling preserves upper support, the support condition reduces to a condition on the primitive type set. Discrete concavity of θ in α is exactly that condition: by the supporting-hyperplane theorem, every point of the concave type set $\{(\alpha_r, \theta_r) : r \in \mathcal{T}\}$ admits a supporting line that lies weakly above all other type points; after rescaling, the upper-support condition therefore holds for every menu item t. Thus, under discrete concavity, outcome-contingent prices can separate all types without leaving rents, so the seller achieves first-best profit.

Theorem 3 (Optimal mechanism)**.** *Suppose* $\gamma_t(a) = \alpha_t s(a)$*, where* $s(0) = 0$ *and* $s(a) > 0$ *for every* $a \in \mathcal{A} \setminus \{0\}$*, and label types so that* $\alpha_1 < \cdots < \alpha_n$*.*

(i) (**Endpoint allocations**) *Assume* $\max_{a\in\mathcal{A}}\{v_t(a) - c(a)\} > 0$ *for* $t \in \{1, n\}$*. At any attained optimum of problem (13), the lowest- and highest-success-probability types receive first-best allocations:*

$$a_1^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}}\{v_1(a) - c(a)\}, \qquad a_n^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}}\{v_n(a) - c(a)\}.$$

(ii) Suppose, in addition, $v_t(a) = \theta_t h(a)$ *with* h *strictly increasing. Define the capability index* $z := h(a) \in \mathcal{Z} := h(\mathcal{A})$*, the reduced cost* $\tilde{c}(z) := c(h^{-1}(z))$*, and, for every triple* $r < t < \tau$*,*

$$\lambda_t^{r\tau} := \frac{\alpha_\tau - \alpha_t}{\alpha_\tau - \alpha_r}, \qquad \Delta_t^{r\tau} := \theta_t - \lambda_t^{r\tau}\theta_r - (1 - \lambda_t^{r\tau})\theta_\tau.$$

(a) (**Convex optimization reformulation**) *The problem of finding an optimal contract un-*

[12] For $\alpha_1 < \cdots < \alpha_n$, discrete concavity means that $(\theta_t - \theta_r)/(\alpha_t - \alpha_r) \geq (\theta_\tau - \theta_t)/(\alpha_\tau - \alpha_t)$ for every $r < t < \tau$.

der full alignment, i.e., problem (13), can be written as

$$\max_{\substack{(z_t,u_t)\in\{(0,0)\}\cup\left((\mathcal{Z}\setminus\{0\})\times\mathbb{R}_+\right)\\ t\in\mathcal{T}}} \quad \sum_{t\in\mathcal{T}} \pi_t[\theta_t z_t - u_t - \tilde{c}(z_t)] \tag{18}$$

$$s.t. \quad u_t - \lambda_t^{r\tau} u_r - (1-\lambda_t^{r\tau})u_\tau \le z_t \Delta_t^{r\tau}, \qquad \forall\, r<t<\tau. \tag{19}$$

If $\mathcal{Z}$ and $\tilde{c}$ are convex, this is a convex optimization problem.[13]

(b) (**Interior curvature virtual types**) *Suppose $\mathcal{Z}$ and $\tilde{c}$ are convex. If problem (18) admits a primal–dual saddle point with nonnegative multipliers $m_t^{r\tau} \ge 0$ on the chord constraints (19), then each interior type $t = 2,\dots,n-1$ has virtual value*

$$\widehat{\theta}_t = \theta_t + \frac{1}{\pi_t} \sum_{r<t<\tau} m_t^{r\tau} \Delta_t^{r\tau},$$

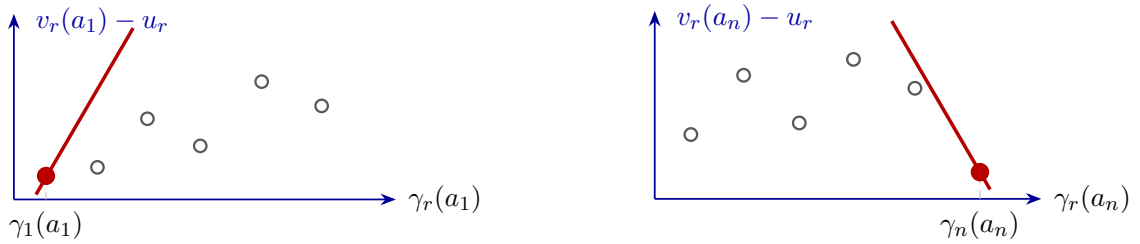
and its optimal allocation solves

$$a_t^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}} \{\widehat{\theta}_t h(a) - c(a)\}.$$

Theorem 3(i) contrasts with standard one-dimensional screening. There, under the usual single-crossing condition, the highest-*value* buyer receives the first-best allocation, while other types are distorted downward (Mussa and Rosen, 1978). Our setting differs in two ways. First, the relevant ordering is by success probability rather than by value. Second, buyers with the highest *and* lowest *success probabilities*—those whose tasks are easiest and hardest for the AI to resolve—receive their first-best allocations at the optimum. That is, each of these two types is assigned the *efficient* capability level: the one that maximizes the total surplus $v_t(a) - c(a)$ she generates and that would be chosen under complete information. The reason comes from the upper-support characterization in Lemma 1. As illustrated in Figure 6, an endpoint item faces mimicking types on only one side of the success-probability axis. A sufficiently steep line can therefore upper-support the point associated with the endpoint type, for any positive-service allocation and any rent of that item. This means that outcome-contingent pricing alone can deter other types from choosing an endpoint buyer's item. For example, for the lowest-success-probability buyer's item, the seller can set a sufficiently high success fee so that buyers with higher success probabilities do not want to deviate to that item. The seller can therefore choose the endpoint allocations independently of the implementability constraints, which yields first-best allocations for types 1 and n. Endpoint rents, however, need not be zero: changing an endpoint rent shifts that type's mimicking point in the upper-support conditions for other items.

Theorem 3(ii)(a) shows that, under separable primitives, finding the optimal mechanism admits a finite-dimensional convex formulation and thus is computationally tractable. The original formulation (13) is non-convex because allocations and outcome-contingent prices interact multi-

[13] A primitive sufficient condition is that $\mathcal{A}$ is an interval, c is increasing and convex, and h is concave.



(a) Lowest success probability

(b) Highest success probability

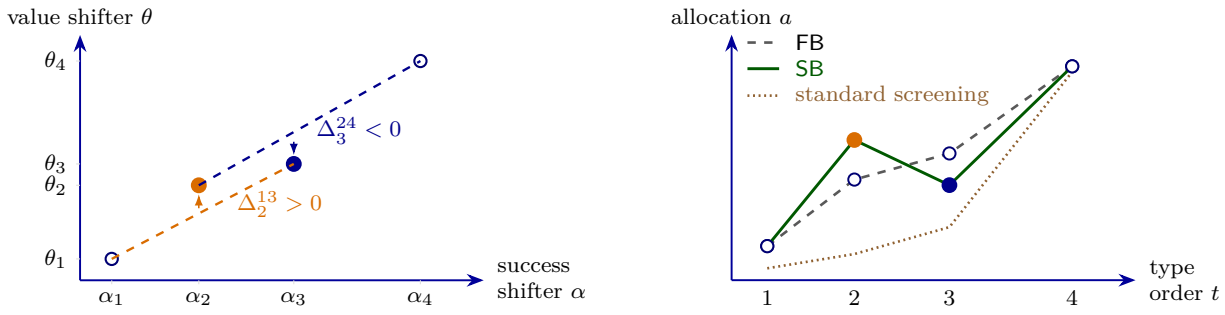
Figure 6: Endpoint items are always upper-supported

Notes. Regardless of allocations and rents, the items assigned to the lowest- and highest-success-probability types are always upper-supported: each faces mimics on only one side of the success-probability axis, so a line with a sufficiently steep slope, positive for the lowest-success item and negative for the highest-success item, supports the assigned type's point.

plicatively. Under separable primitives, the change of variables $z_t = h(a_t)$ converts the problem into the allocation-rent program (18), which is convex whenever $\mathcal{Z}$ and $\tilde{c}$ are convex. The domain restriction $(z_t,\, u_t) \in \{(0,0)\} \cup ((\mathcal{Z} \setminus \{0\}) \times \mathbb{R}_+)$ in (18) reflects exclusion: a no-service item offers every type the same success probability and zero gross value, so incentive compatibility caps its rent at the minimum rent across types, and an optimal menu sets that rent to zero. The result builds on the implementability characterization in Lemma 1. The key step is to convert the geometric upper-support condition into algebraic constraints on allocations and rents; see Lemma A-2 in Section A.2. As Figure 5(a) illustrates, upper support requires *the intended point to lie weakly above every chord connecting one point to its left and one point to its right.* Under separable primitives, this chord condition becomes the linear constraints in (19) after the change of variables $z_t = h(a_t)$. The weight $\lambda_t^{r\tau}$ writes the middle success shifter as a convex combination of the two outer shifters, $\alpha_t = \lambda_t^{r\tau}\alpha_r + (1 - \lambda_t^{r\tau})\alpha_\tau$. The term $\Delta_t^{r\tau}$ is the *chord gap*: the vertical distance at α_t between the middle point (α_t, θ_t) and the chord through (α_r, θ_r) and $(\alpha_\tau, \theta_\tau)$. A positive gap means the middle point lies above the chord, so the triple is locally concave; a negative gap means it lies below the chord, so the triple is locally convex. Figure 7(a) illustrates this geometry.

Theorem 3(ii)(b), built on the convex formulation (18), shows that each interior allocation is chosen as if the type had a curvature-adjusted value parameter $\widehat{\theta}_t$. We call $\widehat{\theta}_t$ the *curvature virtual type* because its adjustment is governed by the chord gaps $\Delta_t^{r\tau}$, which capture the curvature of θ in α.[14] The direction of the screening distortion depends on the sign of $\Delta_t^{r\tau}$ for the chord constraints. If $\Delta_t^{r\tau} > 0$, then type t's value shifter lies above the chord connecting the two surrounding success-probability types r and τ. Increasing $z_t = h(a_t)$ then raises type t's value more than it raises the interpolated value of those surrounding types. Geometrically, the intended point for item t moves up relative to the chord through the mimicking types, making it easier to deter surrounding

[14] The multipliers $m_t^{r\tau}$ are endogenous objects of the optimum. They admit no closed form for general type sets, but they are computationally tractable since (18) is a convex program. For special cases, such as two or three types, we obtain closed-form solutions; see Corollary 4.



(a) Type geometry

(b) Allocation comparison

Figure 7: Chord gaps $\Delta_t^{r\tau}$ and optimal allocation distortions

Notes. In this example, the value and success parameters θ and α have the *same* ordering. Panel (a) plots the type set in the (α, θ) plane and illustrates the chord gaps $\Delta_t^{r\tau}$ that enter the chord constraints. Panel (b) compares the resulting second-best allocations with the first-best allocations and with standard one-dimensional screening.

types from mimicking item t. Thus, a higher allocation has a screening benefit: by making item t's IC constraints easier to satisfy, it can reduce the rents needed to keep types separated. When the corresponding chord constraint binds, this screening benefit raises $\widehat{\theta}_t$ and can push a_t^{SB} above a_t^{FB}. If instead $\Delta_t^{r\tau} < 0$, increasing z_t tightens the support constraint, lowers $\widehat{\theta}_t$, and pushes the allocation downward.

Figure 7(b) illustrates the allocation implications. The dashed curve is the first-best allocation $(a_t^{\mathsf{FB}})_{t\in\mathcal{T}}$, the solid curve is the unrestricted second-best allocation $(a_t^{\mathsf{SB}})_{t\in\mathcal{T}}$, and the dotted curve is the optimal allocation under standard one-dimensional value screening (UBP in our setting). In the example plotted, θ and α are comonotone: types are ordered identically by valuation and by success probability. Under standard screening, only the highest-valuation type receives the first-best allocation and all lower types are distorted downward, so the dotted curve lies weakly below the dashed curve (Mussa and Rosen, 1978). In our setting, consistent with Theorem 3, the buyers with the lowest and highest success probabilities—types 1 and 4—both receive their first-best allocations, while interior distortions can go in either direction. These interior distortions are governed by the chord gaps $\Delta_t^{r\tau}$ through the endogenous multipliers on the chord constraints: positive chord gaps can create an upward screening force, whereas negative chord gaps can create a downward screening force.

We next examine more closely how the curvature of θ in α shapes the optimal contract. Corollary 2 shows that when θ is discretely concave in α, the seller attains the first best. We now turn to the opposite case, in which θ is discretely convex in α.

Corollary 3 (Convex type geometry)**.** *Maintain the assumptions of Theorem 3(ii)(b), and suppose in addition that $\tilde{c}$ is differentiable. Assume that θ is discretely convex in α, equivalently, $\Delta_t^{r\tau} \leq 0$ for all $r < t < \tau \in \mathcal{T}$.*

(i) Every interior type has a weakly lower curvature virtual type: $\widehat{\theta}_t \leq \theta_t$ for $t = 2, \ldots, n-1$.

Consequently,[15] $a_t^{\mathsf{SB}} \le a_t^{\mathsf{FB}}$ *for* $t = 2, \dots, n-1$.

(ii) The optimal buyer-rent profile $(u_t^{\mathsf{SB}})_{t\in\mathcal{T}}$ *is discretely convex in the* α*-order. Hence, the zero-rent types, i.e., the types with binding IR constraints, form a contiguous interval, and positive rents, if any, can occur only in the two tails. If, in addition,* $\Delta_t^{r\tau} < 0$ *for all* $r < t < \tau$ *and* $h(a_t^{\mathsf{SB}}) > 0$ *for all interior types, then at most two IR constraints bind; if two bind, they are adjacent.*

Discrete convexity means that every interior point (α_t, θ_t) lies weakly below the chord joining any lower-success type r and higher-success type τ; equivalently, $\Delta_t^{r\tau} \le 0$ for all $r < t < \tau$. For an interior item t, the chord constraint (19) bounds the rent gap $u_t - \lambda_t^{r\tau} u_r - (1-\lambda_t^{r\tau})u_\tau$ from above by $z_t \Delta_t^{r\tau} \le 0$. Raising z_t, or equivalently $a_t = h^{-1}(z_t)$, therefore weakly lowers this upper bound in every such constraint, making item t harder to support against mimics from lower- and higher-success types. This pushes down the curvature virtual type $\widehat{\theta}_t$. Hence $\widehat{\theta}_t \le \theta_t$, and every interior allocation is weakly distorted downward: $a_t^{\mathsf{SB}} \le a_t^{\mathsf{FB}}$.

The rent statement in Corollary 3(ii) is the dual side of the same geometry. With $\Delta_t^{r\tau} \le 0$, the support constraint (19) implies $u_t^{\mathsf{SB}} \le \lambda_t^{r\tau} u_r^{\mathsf{SB}} + (1-\lambda_t^{r\tau}) u_\tau^{\mathsf{SB}}$ for all $r < t < \tau$. Thus, each interior rent lies weakly below the chord connecting the rents of any lower-success type r and higher-success type τ; equivalently, $(u_t^{\mathsf{SB}})_{t\in\mathcal{T}}$ is discretely convex in α. Combined with $u_t^{\mathsf{SB}} \ge 0$ for every $t \in \mathcal{T}$, this implies that any type between two zero-rent types must also earn zero rent. Hence, the zero-rent types are consecutive in the α-order, and positive rents can appear only in the two tails. The final claim follows from the additional assumptions that $\Delta_t^{r\tau} < 0$ for all $r < t < \tau$ and $h(a_t^{\mathsf{SB}}) > 0$ for all interior types. If three types $i < j < k$ all earned zero rent, then the chord constraint for item j would require $0 \le h(a_j^{\mathsf{SB}})\Delta_j^{ik} < 0$, a contradiction. Hence, at most two IR constraints bind; if two bind, the corresponding types are adjacent.

Finally, we zoom in on the two- and three-type cases, which can be solved in closed form and provide additional insight.

Corollary 4 (Two- and three-type cases)**.** *Suppose* $\gamma_t(a) = \alpha_t s(a)$*, with* $\alpha_1 < \cdots < \alpha_n$*. Also, assume* $s(0) = 0$*,* $s(a) > 0$ *for every* $a \neq 0$*, and* $\max_{a\in\mathcal{A}}\{v_t(a) - c(a)\} > 0$ *for* $t \in \{1, n\}$.

(i) If $n = 2$*, then the first-best profit is achieved:* $R_{\mathsf{SB}} = R_{\mathsf{FB}}$.

(ii) Suppose $n = 3$ *and, in addition,* $v_t(a) = \theta_t h(a)$*,* $t \in \mathcal{T}$*, where* $\theta_t > 0$ *and* h *is strictly increasing. Define*

$$\lambda := \frac{\alpha_3 - \alpha_2}{\alpha_3 - \alpha_1}, \qquad \chi := \theta_2 - \lambda\theta_1 - (1-\lambda)\theta_3.$$

(a) If $\chi \ge 0$*, then* $R_{\mathsf{SB}} = R_{\mathsf{FB}}$.

(b) If $\chi < 0$*, then the endpoint types receive first-best allocations, while the middle type's*

[15] If the first-best maximizer is not unique, take a_t^{FB} to be the largest first-best maximizer.

allocation solves

$$a_2^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}} \left\{ \left(\theta_2 + \frac{\xi\chi}{\pi_2} \right) h(a) - c(a) \right\}, \qquad \xi := \min\left\{ \frac{\pi_1}{\lambda}, \frac{\pi_3}{1-\lambda} \right\}.$$

Moreover, the middle type earns zero rent, $u_2^{\mathsf{SB}} = 0$*, and the endpoint rents are given by:*

$$(u_1^{\mathsf{SB}}, u_3^{\mathsf{SB}}) = \begin{cases} (-h(a_2^{\mathsf{SB}})\chi/\lambda,\, 0), & \pi_1/\lambda < \pi_3/(1-\lambda), \\ (0,\, -h(a_2^{\mathsf{SB}})\chi/(1-\lambda)), & \pi_3/(1-\lambda) < \pi_1/\lambda, \\ \textit{any split with } \lambda u_1 + (1-\lambda)u_3 = -h(a_2^{\mathsf{SB}})\chi, & \pi_1/\lambda = \pi_3/(1-\lambda). \end{cases}$$

With two types, there is no interior type and hence no triple $r < t < \tau$. Problem (18) therefore has no constraints: both types are endpoints, and the seller achieves the first best with zero rents. With three types, there is a single chord constraint in (18): the middle item must be defended against the two endpoint types. If the middle point is locally concave, $\chi \geq 0$, this constraint is satisfied with zero rents, so the first-best profit is achieved. If the middle point is locally convex, $\chi < 0$, the middle-item chord constraint cannot be satisfied with zero rents. The endpoint allocations remain first best, while the middle allocation is distorted downward, $a_2^{\mathsf{SB}} \leq a_2^{\mathsf{FB}}$, by the effective rent price $-\xi\chi/\pi_2 > 0$. The middle type earns zero rent, and the seller provides the rent needed to satisfy the chord constraint through whichever endpoint has the lower expected cost per unit of support, $\min\{\pi_1/\lambda, \pi_3/(1-\lambda)\}$. Accordingly, the single chord constraint binds, so $\mathsf{IC}_{1\to 2}$, $\mathsf{IC}_{3\to 2}$, and IR_2 bind. The endpoint that does not receive this rent earns zero rent, so its IR constraint binds; if the two unit support costs are tied, any rent split satisfying the chord constraint is optimal. This contrasts sharply with standard screening, where the binding IC constraints form a downward chain in the valuation order: $\mathsf{IC}_{(2)\to(1)}$ and $\mathsf{IC}_{(3)\to(2)}$. Here the IC-binding pattern is instead *concentric* in the success-probability dimension.

6 Conclusion

AI as a Service shifts the object of exchange from access to performance, making both the allocation of failure risk and the measurement of success central to pricing. Neither UBP nor PFS is universally superior: their relative appeal depends on how buyer value relates to AI performance, while PFS additionally requires a credible definition of success. Our central takeaway is that this credibility problem is not inherent to outcome-based pricing; it arises from restricting payment to a single success fee. With richer, transparent outcome-contingent contracts, the seller can preserve an aligned success metric without sacrificing profit or incurring costly disputes. Under the optimal contract, the buyers with the easiest and the hardest tasks receive efficient, first-best allocations.

This alignment-for-free conclusion, however, has clear boundaries. It relies on the seller's freedom to choose *both* outcome-contingent prices: the replacement construction operates on expected payments—presuming a risk-neutral buyer—and may require a failure fee that exceeds the success

fee, or even negative fees. Bounds on fees or buyer risk aversion may likely block the construction. Alignment for free is thus a property of rich, transparent outcome-contingent contracting; when such practical constraints bind, strategic metric design and costly disputes re-emerge. Effective AI pricing therefore requires contract design and performance measurement to be considered jointly.

References

Saeed Alaei, Alexandre Belloni, Ali Makhdoumi, and Azarakhsh Malekian. Optimal auction design with deferred inspection and reward. *Operations Research*, 72(6):2413–2429, 2024.

Saeed Alaei, Ali Daei Naby, Ali Makhdoumi, and Azarakhsh Malekian. Optimal mechanism design with post-allocation signal. *SSRN Electronic Journal*, Feb 2026. URL https://ssrn.com/abstract=6220718.

Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and Matei Zaharia. A view of cloud computing. *Communications of the ACM*, 53(4):50–58, April 2010.

Kursad Asdemir, Nanda Kumar, and Varghese S. Jacob. Pricing models for online advertising: CPM vs. CPC. *Information Systems Research*, 23(3):804–822, 2012. doi: 10.1287/isre.1110.0391.

Martino Banchio and Giacomo Mantegazza. Artificial intelligence and spontaneous collusion. *arXiv preprint arXiv:2202.05946*, 2022.

Hamsa Bastani and Gerard P. Cachon. The human-AI contracting paradox, December 2025. Working paper.

Alexandre Belloni, Xuanjie Li, and Ali Makhdoumi. Approximately optimal online mechanism for multiunit demand buyers with post-allocation inspection. In *Proceedings of the 26th ACM Conference on Economics and Computation*, pages 354–354, 2025.

Dirk Bergemann and Stephen Morris. Bayes correlated equilibrium and the comparison of information structures in games. *Theoretical Economics*, 11(2):487–522, 2016. doi: 10.3982/TE1808.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *Journal of Economic Literature*, 57 (1):44–95, 2019. doi: 10.1257/jel.20181489.

Dirk Bergemann, Marek Bojko, Paul Dütting, Renato Paes Leme, Haifeng Xu, and Song Zuo. Data-driven mechanism design: Jointly eliciting preferences and information. *arXiv:2412.16132v2*, 2025a.

Dirk Bergemann, Alessandro Bonatti, and Alex Smolin. Menu pricing of large language models. *arXiv preprint arXiv:2502.07736*, 2025b.

Boston Consulting Group. Rethinking B2B software pricing in the agentic AI era, 2025. URL https://www.bcg.com/publications/2025/rethinking-b2b-software-pricing-in-the-era-of-ai. Accessed Jul. 2026.

Emilio Calvano, Giacomo Calzolari, Vincenzo Denicolò, and Sergio Pastorello. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review*, 110(10):3267–3297, 2020. doi: 10.1257/aer.20190623.

Ozan Candogan. Information design in operations. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, Tutorials in Operations Research, pages 176–201. INFORMS, 2020. doi: 10.1287/educ.2020.0217.

Shengyu Cao and Ming Hu. LLM collusion. *arXiv preprint arXiv:2601.01279*, 2026.

Francisco Castro, Jian Gao, and Sébastien Martin. Human-ai interactions and societal pitfalls. In *Proceedings of the 25th ACM Conference on Economics and Computation (EC '24), New Haven, CT, USA, July 8–11, 2024*, page 205. ACM, 2024. doi: 10.1145/3670865.3673482.

Jacques Crémer and Richard P McLean. Optimal selling strategies under uncertainty for a discriminating monopolist when demands are interdependent. *Econometrica*, 53(2):345–361, 1985. doi: 10.2307/1911240.

Jacques Crémer and Richard P McLean. Full extraction of the surplus in Bayesian and dominant strategy auctions. *Econometrica: Journal of the Econometric Society*, pages 1247–1257, 1988.

Chrysanthos Dellarocas. Double marginalization in performance-based advertising: Implications and solutions. *Man-*

agement Science, 58(6):1178–1195, 2012.

Deloitte Insights. SaaS meets AI agents: Transforming budgets, customer experience, and workforce dynamics, November 2025. URL https://www.deloitte.com/us/en/insights/industry/technology/technology-media-and-telecom-predictions/2026/saas-ai-agents.html. Technology, Media, and Telecommunications Predictions 2026 (accessed Jul. 2026).

Mert Demirer, Andrey Fradkin, Nadav Tadelis, and Sida Peng. The emerging market for intelligence: Pricing, supply, and demand for LLMs. NBER Working Paper 34608, National Bureau of Economic Research, 2025.

Paul Dütting, Michal Feldman, Inbal Talgam-Cohen, et al. Algorithmic contract theory: A survey. *Foundations and Trends® in Theoretical Computer Science*, 16(3-4):211–412, 2024a.

Paul Dütting, Vahab Mirrokni, Renato Paes Leme, Haifeng Xu, and Song Zuo. Mechanism design for large language models. In *Proceedings of the ACM Web Conference (WWW)*, 2024b. doi: 10.1145/3589334.3645511. Also available as arXiv:2310.10826.

Benjamin Edelman, Michael Ostrovsky, and Michael Schwarz. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American economic review*, 97(1):242–259, 2007.

Pnina Feldman, Ella Segev, and Shreyas Sekar. Pricing human-AI services: The role of observability and congestion. Working paper, SSRN 6849501, 2026.

Sara Fish, Yannai A. Gonczarowski, and Ran I. Shorrer. Algorithmic collusion by large language models. *arXiv preprint arXiv:2404.00806*, 2024.

Caner Gocmen, Thodoris Lykouris, Deeksha Sinha, and Wentao Weng. Scheduling with uncertain holding costs and its application to content moderation. *arXiv:2505.21331*, 2025.

Sanford J Grossman and Oliver D Hart. An analysis of the principal-agent problem. *Econometrica*, 51(1):7–45, 1983.

Jose A. Guajardo, Morris A. Cohen, Sang-Hyun Kim, and Serguei Netessine. Impact of performance-based contracting on product reliability: An empirical analysis. *Management Science*, 58(5):961–979, 2012. doi: 10.1287/mnsc.1110.1465.

Shivam Gupta, Shouqiang Wang, Milind Dawande, and Ganesh Janakiraman. Procurement with cost and noncost attributes: Cost-sharing mechanisms. *Operations Research*, 69(5):1349–1367, 2021.

Yonatan Gur, Gregory Macnamara, and Daniela Saban. Sequential procurement with contractual and experimental learning. *Management Science*, 68(4):2714–2731, 2022.

Huseyin Gürkan and Francis de Véricourt. Contracting, pricing, and data collection under the AI flywheel effect. *Management Science*, 68(12):8791–8808, 2022. doi: 10.1287/mnsc.2022.4333.

Nima Haghpanah and Ron Siegel. Screening two types. In *Proceedings of the 26th ACM Conference on Economics and Computation*, EC '25, page 921, 2025.

Bengt Holmström. Moral hazard and observability. *The Bell Journal of Economics*, 10(1):74–91, 1979. doi: 10.2307/3003320.

Bengt Holmström and Paul Milgrom. Aggregation and linearity in the provision of intertemporal incentives. *Econometrica*, 55(2):303–328, 1987. doi: 10.2307/1913238.

Ming Hu, Xiaofei Li, Quan Yuan, and Han Zhu. How ai may impact customer-intensive services. *Working paper*, 2025.

Krishnamurthy Iyer, Alec Sun, Haifeng Xu, and You Zu. How to sell a service with uncertain outcomes. *arXiv preprint arXiv:2502.13334*, 2025.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *American Economic Review*, 101(6):2590–2615, 2011. doi: 10.1257/aer.101.6.2590.

Polyna Khudyakov, Paul D. Feigin, and Avishai Mandelbaum. Designing a call center with an IVR (interactive voice response). *Queueing Systems*, 66(3):215–237, 2010. doi: 10.1007/s11134-010-9193-y.

Sang-Hyun Kim, Morris A. Cohen, and Serguei Netessine. Performance contracting in after-sales service supply

chains. *Management Science*, 53(12):1843–1858, 2007. doi: 10.1287/mnsc.1070.0741.

Jiung Lee, Hongseok Namkoong, and Yibo Zeng. Design and scheduling of an ai-based queueing system. *arXiv preprint arXiv:2406.06855*, 2024.

Xuanjie Li, Ali Makhdoumi, and Saša Pekeč. Revenue maximization with exclusion. Working paper, SSRN 5139835, 2025.

Giuseppe Lopomo, Luca Rigotti, and Chris Shannon. Detectability, duality, and surplus extraction. *Journal of Economic Theory*, 204:105465, 2022. doi: 10.1016/j.jet.2022.105465.

Thodoris Lykouris and Wentao Weng. Learning to defer in congested systems: The ai-human interplay. *arXiv:2402.12237*, 2025.

Rafid Mahmood. Pricing and competition for generative AI. In *Advances in Neural Information Processing Systems*, 2024. Also available as arXiv:2411.02661.

Eric Maskin and John Riley. Monopoly with incomplete information. *The RAND Journal of Economics*, 15(2): 171–196, 1984.

R Preston McAfee and Philip J Reny. Correlated information and mechanism design. *Econometrica: Journal of the Econometric Society*, pages 395–421, 1992.

McKinsey & Company. The state of AI: How organizations are rewiring to capture value. Technical report, McKinsey & Company, March 2025.

Peter Mell and Timothy Grance. The NIST definition of cloud computing. NIST Special Publication 800-145, National Institute of Standards and Technology, September 2011. URL https://nvlpubs.nist.gov/nistpubs/legacy/sp/nistspecialpublication800-145.pdf.

Claudio Mezzetti. Mechanism design with interdependent valuations: Efficiency. *Econometrica*, 72(5):1617–1626, 2004.

Michael Mussa and Sherwin Rosen. Monopoly and product quality. *Journal of Economic theory*, 18(2):301–317, 1978.

Roger B Myerson. Incentive compatibility and the bargaining problem. *Econometrica: journal of the Econometric Society*, pages 61–73, 1979.

Roger B Myerson. Optimal auction design. *Mathematics of operations research*, 6(1):58–73, 1981.

William Overman and Mohsen Bayati. The oversight game: Learning to cooperatively balance an ai agent's safety and autonomy. *arXiv preprint arXiv:2510.26752*, 2025.

Michael H Riordan and David EM Sappington. Optimal contracts with public ex post information. *Journal of Economic Theory*, 45(1):189–199, 1988.

Guillaume Roels, Uday S Karmarkar, and Scott Carr. Contracting for collaborative services. *Management Science*, 56(5):849–863, 2010.

Ludvig Sinander. Topics in mechanism design, 2026. URL https://www.ludvigsinander.net/lecture_notes.html. Lecture notes, University of Oxford.

Daniel Sun. How to implement AI agents to transform business models, February 2025. URL https://www.gartner.com/en/articles/ai-agents. Published February 27, 2025 (accessed Jan. 2026).

Appendix

Appendix Contents

A	Additional Mechanism Results	App-1
A.1	Dispute Structure under PFS	App-1
A.2	Further Results and Discussion of the Implementability Result	App-2
A.3	Beyond Binary Outcomes	App-3
B	Proofs for the Main Text	App-6
B.1	Proofs for Section 4	App-6
B.2	Proofs for Section 5	App-16
C	Proofs for Appendix Results	App-24
C.1	Proofs for Appendix A.1	App-24
C.2	Proofs for Appendix A.2	App-27

Appendix A collects auxiliary results: dispute structure under PFS, additional implementability details for the unrestricted mechanism, and the extension to many outcomes. Appendix B proves the main-text results. Appendix C collects proofs for appendix results.

A Additional Mechanism Results

A.1 Dispute Structure under PFS

This subsection records how a buyer responds to AI-recorded billable metric outcomes under pay-for-success (PFS) and collects the dispute structure that the comparison in Section 4 relies on. Proofs appear in Section C.1.

For an item s, write

$$\eta_s(\gamma) := \gamma\sigma_s^1 + (1-\gamma)\sigma_s^0$$

for the probability of a reported success when the true success probability is γ.

Lemma A-1 (PFS reported-success disputes)**.** *Under* PFS*, the buyer never disputes after reported failure $\widehat{\omega} = 0$. The only possible dispute is after reported success $\widehat{\omega} = 1$. After observing $\widehat{\omega} = 1$, the buyer disputes if and only if*

$$(1-\gamma)p_s\sigma_s^0 > \kappa_B\eta_s(\gamma).$$

The buyer's expected cost from item s, including both the bill and any dispute cost, when her success

probability at this item is γ, is

$$C_s(\gamma) = \min\Bigg\{ \underbrace{p_s \Pr(\widehat{\omega}=1)}_{\substack{\text{no dispute}\\ \text{reported-outcome bill}}} , \underbrace{p_s \Pr(\omega=1,\widehat{\omega}=1)}_{\substack{\text{dispute}\\ \text{verified bill}}} + \underbrace{\kappa_B \Pr(\widehat{\omega}=1)}_{\substack{\text{dispute}\\ \text{buyer cost}}} \Bigg\} \tag{A-1}$$
$$= \min\left\{ p_s\eta_s(\gamma), \gamma p_s \sigma_s^1 + \kappa_B \eta_s(\gamma) \right\}.$$

The seller's gross revenue is $p_s\eta_s(\gamma)$ in the no-dispute branch and $\gamma p_s\sigma_s^1 - \kappa_S\eta_s(\gamma)$ in the dispute branch.

The buyer's cost (A-1) is the lower envelope of two affine functions of the success probability—a no-dispute bill and a dispute bill. The next result shows that there exists an optimal PFS such that at equilibrium, no buyer will dispute.

Proposition A-1 (No on-path disputes in optimal PFS)**.** *Assume $v_t(a) = 0$ whenever $\gamma_t(a) = 0$. There exists an optimal PFS menu in which no buyer disputes on her assigned item.*

The next result shows that there exists an optimal PFS mechanism in which only lower-success-probability types may dispute when they deviate to items targeted at higher-success-probability types.

Proposition A-2. *Assume that $v_t(a) = 0$ whenever $\gamma_t(a) = 0$, and also types are ordered by success probability:*

$$\gamma_1(a) \le \gamma_2(a) \le \cdots \le \gamma_n(a) \qquad \forall a \in \mathcal{A}.$$

There exists an optimal PFS menu such that, if a type-t buyer chooses item $s \le t$, she does not dispute.

A.2 Further Results and Discussion of the Implementability Result

We now provide further results and discussion of the key implementability result, Lemma 1. The next lemma gives an algebraic representation of the upper-support geometry.

Lemma A-2 (Algebraic Representation of Upper-Support Geometry)**.** *The point (x_t, y_t), $t \in \mathcal{T}$, in a finite set $\{(x_r, y_r) : r \in \mathcal{T}\} \subset \mathbb{R}^2$ is* upper-supported if and only if

$$y_t \ge y_r \qquad \forall r \in \mathcal{T} \text{ with } x_r = x_t,$$

and

$$y_t \ge \lambda_t^{r\tau} y_r + (1 - \lambda_t^{r\tau}) y_\tau \qquad \forall r, \tau \in \mathcal{T} \text{ with } x_r < x_t < x_\tau,$$

where $\lambda_t^{r\tau} := \frac{x_\tau - x_t}{x_\tau - x_r}$.

The weight $\lambda_t^{r\tau} := (x_\tau - x_t)/(x_\tau - x_r)$ is the interpolation weight that writes $x_t = \lambda_t^{r\tau} x_r + (1 - \lambda_t^{r\tau}) x_\tau$. Thus $\lambda_t^{r\tau} y_r + (1 - \lambda_t^{r\tau}) y_\tau$ is the height of the chord connecting (x_r, y_r) and (x_τ, y_τ) at the

intended point x_t. The lemma therefore says that upper support requires two things: points with the same horizontal coordinate must lie weakly below the intended point, and *for any point to the left of the intended point and any point to its right, the chord connecting them must pass weakly below the intended point.* These conditions are also sufficient because they guarantee a supporting line through (x_t, y_t) that lies weakly above every point in the finite set.

Applying Lemma A-2 to the unrestricted contract-design problem (15) gives the corollary.

Corollary A-1. *The unrestricted contract-design problem* (13) *is equivalent to*

$$R^{\text{fa}}_{\mathsf{SB}} = \sup_{\substack{a_t \in \mathcal{A},\ u_t \geq 0 \\ t \in \mathcal{T}}} \sum_{t \in \mathcal{T}} \pi_t \big[v_t(a_t) - c(a_t) - u_t \big] \tag{A-2}$$

$$\text{s.t.} \quad v_t(a_t) - u_t \geq v_r(a_t) - u_r, \quad \forall t, r \in \mathcal{T} \text{ with } \gamma_r(a_t) = \gamma_t(a_t),$$

$$v_t(a_t) - u_t \geq \lambda_t^{r\tau}(a) \big[v_r(a_t) - u_r \big] + \big(1 - \lambda_t^{r\tau}(a)\big) \big[v_\tau(a_t) - u_\tau \big], \ \forall t, r, \tau \in \mathcal{T} \text{ with } \gamma_r(a_t) < \gamma_t(a_t) < \gamma_\tau(a_t),$$

where

$$\lambda_t^{r\tau}(a) := \frac{\gamma_\tau(a_t) - \gamma_t(a_t)}{\gamma_\tau(a_t) - \gamma_r(a_t)}. \tag{A-3}$$

This corollary reduces the problem (13) from three decision vectors—allocations and two outcome-contingent prices—to two decision vectors: allocations and rents. This formulation is powerful because, under certain additional structure, e.g., under separable primitives, the same upper-support constraints become linear, leading to the convex formulation in Theorem 3.

Our next result supports Remark 1: under mild conditions on the success-probability function, any positive-success allocation profile is implementable.

Lemma A-3 (Any positive-success allocation is supportable with enough rents)**.** *Suppose the success probabilities are separable, $\gamma_t(a) = \alpha_t s(a)$, with distinct parameters α_t. Then, for any fixed allocation profile $(a_t)_{t \in \mathcal{T}} \in \mathcal{A}^n$ satisfying $s(a_t) > 0$ for every t, there exists a rent profile $(u_t)_{t \in \mathcal{T}} \in \mathbb{R}^n_+$ such that the allocation-rent profile $\big((a_t, u_t)\big)_{t \in \mathcal{T}}$ satisfies support-line implementability. That is, any positive-success allocation profile is implementable.*

Intuitively, given an allocation profile and ignoring rents, the seller still has two outcome-contingent prices for each report, which are flexible enough to implement the allocation. Formally, for any allocation profile $(a_t)_{t \in \mathcal{T}}$ satisfying $s(a_t) > 0$ for every t, we construct a rent profile $(u_t)_{t \in \mathcal{T}}$, chosen large enough, such that $\big((a_t, u_t)\big)_{t \in \mathcal{T}}$ is implementable.

A.3 Beyond Binary Outcomes

We now relax the binary-outcome assumption and show that the main results about optimal unrestricted mechanisms in Section 5 continue to hold.

Let the buyer-value-relevant outcome set be $\Omega = \{\omega_1, \ldots, \omega_K\}$. At allocation a, a type-t task

induces the true outcome distribution

$$\boldsymbol{\gamma}_t(a) := \big(\gamma_{t,\omega_1}(a), \ldots, \gamma_{t,\omega_K}(a)\big) \in \Delta(\Omega), \qquad \gamma_{t,\omega_k}(a) := \mathbb{P}(\omega = \omega_k \mid t, a), \tag{A-4}$$

where $\Delta(\Omega)$ is the simplex over Ω. As in the main model, the seller commits to an outcome-alignment rule that maps the true outcome ω into the AI-recorded billable metric outcome $\widehat{\omega} \in \Omega$. With many outcomes, this rule is a row-stochastic matrix

$$\Sigma_t = (\sigma_t^{ij})_{i,j=1}^K, \qquad \sigma_t^{ij} := \mathbb{P}(\widehat{\omega} = \omega_j \mid \omega = \omega_i, t), \qquad \sum_{j=1}^K \sigma_t^{ij} = 1,$$

whose off-diagonal entries capture success-metric misalignment. Full outcome alignment is $\sigma_t^{ij} = \mathbf{1}\{i = j\}$. A menu item t specifies an allocation a_t, an outcome-alignment rule Σ_t, and outcome-contingent bills $\boldsymbol{p}_t := (p_{t,\omega_1}, \ldots, p_{t,\omega_K}) \in \mathbb{R}^K$. After observing $\widehat{\omega}$, the buyer pays the corresponding bill or disputes to verify the true outcome ω at cost κ_B to herself and κ_S to the seller, similar to the binary-outcome setting.

Theorem A-1 (Many-outcome full outcome alignment)**.** *In the many-outcome setting, full outcome alignment is still without loss of optimality, i.e.,* $\sigma_t^{ij} = \mathbf{1}\{i = j\}$.

The argument extends Theorem 2. Fix a menu item t and its allocation a_t. For each fixed pay-or-dispute rule, the buyer's expected cost is affine in the true outcome distribution $\boldsymbol{\gamma}$, because the reporting matrix maps this distribution linearly into true and reported outcomes. Because the buyer optimally chooses whether to pay or dispute, the buyer's expected cost $C_t(\boldsymbol{\gamma})$ is the minimum of finitely many affine functions and is therefore *concave*. Choose an affine function attaining this minimum at $\boldsymbol{\gamma}_t(a_t)$. It equals C_t there and lies weakly above C_t elsewhere due to concavity. Under full alignment, its values at the simplex vertices define new outcome-contingent bills. These bills preserve the assigned type's cost, weakly raise every mimic's cost, eliminate dispute waste, and weakly increase profit. Thus, unrestricted outcome-contingent bills make success-metric misalignment redundant.

For the optimal-contract extension, impose a separability condition on the many-outcome distribution: each type has an outcome-distribution-relevant parameter $\boldsymbol{\alpha}_t \in \mathbb{R}^d$ for some positive integer $d \geq 1$, and the whole outcome-distribution vector is affine in that vector. We use $\mathrm{co}(\cdot)$ for convex hulls and $\mathrm{ext}(\cdot)$ for extreme points.

Theorem A-2 (Many-outcome optimal contract)**.** *Suppose each type $t \in \mathcal{T}$ has a vector index $\boldsymbol{\alpha}_t \in \mathbb{R}^d$ and, for each allocation a, the outcome-distribution vector admits the affine representation*

$$\boldsymbol{\gamma}_t(a) = \boldsymbol{b}(a) + B(a)\boldsymbol{\alpha}_t, \tag{A-5}$$

where $\boldsymbol{b}(a) \in \mathbb{R}^K$, $B(a) \in \mathbb{R}^{K \times d}$ with $B(0) = \mathbf{0}_{K \times d}$, and the columns of $B(a)$ are linearly independent for every $a \in \mathcal{A} \setminus \{0\}$. Assume also $\max_{a \in \mathcal{A}}\{v_t(a) - c(a)\} > 0$ for every t with

$\boldsymbol{\alpha}_t \notin \mathrm{co}\{\boldsymbol{\alpha}_r : r \neq t\}$.

(i) (**Extreme-point allocations**) *If* $\boldsymbol{\alpha}_t \notin \mathrm{co}\{\boldsymbol{\alpha}_r : r \neq t\}$*—equivalently,* $\boldsymbol{\alpha}_t \in \mathrm{ext}(\mathrm{co}\{\boldsymbol{\alpha}_r : r \in \mathcal{T}\})$*—then type* t *receives a first-best allocation, i.e.,* $a_t^{\mathsf{SB}} \in \arg\max_{a \in \mathcal{A}}\{v_t(a) - c(a)\}$.

(ii) (**Convex characterization**) *Suppose, in addition,* $v_t(a) = \theta_t h(a)$ *with* h *strictly increasing, and write* $z := h(a) \in \mathcal{Z} := h(\mathcal{A})$ *and* $\tilde{c} := c \circ h^{-1}$. *Let* e_t *denote the unit vector in* $\Delta(\mathcal{T})$ *that puts probability one on type* t. *With*

$$\Lambda_t := \Big\{\lambda \in \Delta(\mathcal{T}) : \textstyle\sum_{r\in\mathcal{T}} \lambda_r \boldsymbol{\alpha}_r = \boldsymbol{\alpha}_t\Big\}, \quad \mathcal{E}_t := \mathrm{ext}(\Lambda_t) \setminus \{e_t\}, \quad \Delta_t^\lambda := \theta_t - \textstyle\sum_{r\in\mathcal{T}} \lambda_r \theta_r,$$

the second-best problem is equivalent to

$$\max_{\substack{(z_t,u_t)\in\{(0,0)\}\cup\left((\mathcal{Z}\setminus\{0\})\times\mathbb{R}_+\right)\\ t\in\mathcal{T}}} \quad \sum_{t\in\mathcal{T}} \pi_t[\theta_t z_t - \tilde{c}(z_t) - u_t] \tag{A-6}$$

$$s.t. \quad u_t - \sum_{r\in\mathcal{T}} \lambda_r u_r \leq z_t\, \Delta_t^\lambda, \qquad \forall t \in \mathcal{T},\ \lambda \in \mathcal{E}_t. \tag{A-7}$$

When $\mathcal{Z}$ *and* $\tilde{c}$ *are convex, this is a convex program. If it admits a primal–dual saddle point with nonnegative multipliers* $m_t^\lambda \geq 0$ *on the constraints* (A-7)*, each optimal allocation solves* $a_t^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}}\{\widehat{\theta}_t h(a) - c(a)\}$ *with curvature virtual type* $\widehat{\theta}_t = \theta_t + \pi_t^{-1} \sum_{\lambda\in\mathcal{E}_t} m_t^\lambda \Delta_t^\lambda$.

Part (i) is the many-outcome form of the endpoint result. In the binary model the lowest and highest success shifters are the endpoints of a line; here an "endpoint" is any index $\boldsymbol{\alpha}_t$ lying outside the convex hull of the other indices. Such a type can be strictly separated from the rest by a hyperplane, so a sufficiently steep bill vector supports its incentive constraints whatever its allocation. Its allocation therefore need not be distorted for screening, and it receives a first-best allocation at any attained optimum. For such an extreme type, $\Lambda_t = \{e_t\}$ and hence $\mathcal{E}_t = \varnothing$, so (A-7) is automatically satisfied.

The linear-independence condition is what lets outcome-contingent bills act as supporting hyperplanes in the type-index space. For item t, the bills $\boldsymbol{p}_t$ generate the expected payment $\boldsymbol{p}_t^\top \boldsymbol{\gamma}_r(a_t) = \boldsymbol{p}_t^\top \boldsymbol{b}(a_t) + \boldsymbol{p}_t^\top B(a_t)\boldsymbol{\alpha}_r$, affine in the mimic's index $\boldsymbol{\alpha}_r$ with slope $B(a_t)^\top \boldsymbol{p}_t$. When the columns of $B(a_t)$ with $a_t \neq 0$ are linearly independent, the map $\boldsymbol{p}_t \mapsto B(a_t)^\top \boldsymbol{p}_t$ is onto $\mathbb{R}^d$, so the seller can set this slope in any direction, while a common shift of all K bills sets the intercept. Outcome bills can therefore trace out an arbitrary affine function of $\boldsymbol{\alpha}_r$—the vector form of the support line in Lemma 1. The base model is the special case $d = 1$ and $K = 2$. Ordering failure first and success second, its outcome distribution is

$$\boldsymbol{\gamma}_t(a) = \begin{pmatrix} 1 - \alpha_t s(a) \\ \alpha_t s(a) \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \begin{pmatrix} -s(a) \\ s(a) \end{pmatrix} \alpha_t.$$

Thus $B(a) = (-s(a), s(a))^\top$. Since $s(a) > 0$ whenever $a \neq 0$, $B(a)$ is a nonzero one-column matrix and therefore has full column rank.

Part (ii) extends the base-model convex program in Theorem 3(ii)(a) to the many-outcome setting. The set Λ_t collects the nonnegative weights λ that sum to one and reproduce $\boldsymbol{\alpha}_t$ as $\sum_r \lambda_r \boldsymbol{\alpha}_r = \boldsymbol{\alpha}_t$, while $\mathcal{E}_t$ excludes the unit vector e_t, which places all weight on type t, and retains the nontrivial extreme weights. For each $\lambda \in \mathcal{E}_t$, the gap $\Delta_t^\lambda = \theta_t - \sum_r \lambda_r \theta_r$ compares type t's value shifter with the corresponding weighted average. After setting $z_t = h(a_t)$, these conditions become (A-7), whose multipliers generate the curvature virtual types. When $d = 1$, the hull's extreme points are the lowest and highest α-types, and (A-7) reduces to the scalar chord constraints (19).

B Proofs for the Main Text

B.1 Proofs for Section 4

B.1.1 Proofs for Section 4.1

Proof of Proposition 1. Under UBP, each item has the same success and failure bill, $p_r = q_r$. Thus, the AI-recorded billable metric outcome does not change the buyer's payment or the seller's profit, and disputing can only add to the buyer's dispute cost. Replacing any outcome-alignment rule by full alignment, inducing no dispute, is therefore without loss.

After removing the reporting layer, type t's utility from item r is

$$\theta_t h(a_r) - p_r.$$

The remaining problem is the standard finite-type nonlinear-pricing problem with scalar type θ_t, quality $h(a)$, and transfer p (Mussa and Rosen, 1978). The usual one-dimensional screening argument gives monotone allocations in the value order, the virtual-surplus program in (5), and the payment recursion in (6). □

Proof of Proposition 2. Under PFS with full outcome alignment, item r has failure bill zero and success bill p_r. Let

$$m_r := s(a_r) p_r.$$

Then type t's utility from item r can be written as

$$\theta_t h(a_r) - \alpha_t s(a_r) p_r = \alpha_t [\phi_t h(a_r) - m_r], \qquad \phi_t := \frac{\theta_t}{\alpha_t}.$$

Since $\alpha_t > 0$, type t's pairwise IC constraint between any two items r and s is equivalent to

$$\theta_t h(a_r) - \alpha_t m_r \geq \theta_t h(a_s) - \alpha_t m_s \quad \Longleftrightarrow \quad \phi_t h(a_r) - m_r \geq \phi_t h(a_s) - m_s.$$

Thus, fully aligned PFS has exactly the standard one-dimensional IC system with type ϕ_t, quality $h(a)$, and transformed transfer m. IC therefore requires the allocation to be monotone in the per-success order $\phi_{[1]} \leq \cdots \leq \phi_{[n]}$.

Fix any allocation monotone in this order, and define transformed utility $u_t := \phi_t h(a_t) - m_t$. By the standard one-dimensional screening result, the profit-maximizing payments for this fixed allocation minimize transformed rents: the lowest-type IR constraint binds, and the adjacent downward IC constraints bind. Hence

$$u_{[1]} = 0, \qquad u_{[t]} = u_{[t-1]} + (\phi_{[t]} - \phi_{[t-1]})h(a_{[t-1]}), \quad t = 2, \dots, n.$$

Expected profit is

$$\sum_{t=1}^{n} \pi_{[t]} \big[\alpha_{[t]} m_{[t]} - c(a_{[t]})\big] = \sum_{t=1}^{n} \pi_{[t]} \big[\alpha_{[t]} \phi_{[t]} h(a_{[t]}) - c(a_{[t]}) - \alpha_{[t]} u_{[t]}\big].$$

Substituting the transformed rents yields

$$\begin{aligned} &\sum_{t=1}^{n} \pi_{[t]} \big[\alpha_{[t]} \phi_{[t]} h(a_{[t]}) - c(a_{[t]})\big] - \sum_{t=1}^{n-1} \left(\sum_{k=t+1}^{n} \pi_{[k]} \alpha_{[k]} \right) (\phi_{[t+1]} - \phi_{[t]}) h(a_{[t]}) \\ &\qquad = \sum_{t=1}^{n} \pi_{[t]} \big[\alpha_{[t]} \widetilde{\phi}_{[t]} h(a_{[t]}) - c(a_{[t]})\big], \end{aligned}$$

where $\widetilde{\phi}_{[t]}$ is the PFS virtual per-success value from Proposition 2. Finally, $u_{[1]} = 0$ gives $m_{[1]} = \phi_{[1]} h(a_{[1]})$. For $t \geq 2$, the adjacent downward IC constraint binds:

$$\phi_{[t]} h(a_{[t]}) - m_{[t]} = \phi_{[t]} h(a_{[t-1]}) - m_{[t-1]},$$

so

$$m_{[t]} = m_{[t-1]} + \phi_{[t]} \big(h(a_{[t]}) - h(a_{[t-1]})\big), \qquad t \geq 2.$$

If $h(a_{[t]}) > 0$, then $s(a_{[t]}) > 0$ and we set $p_{[t]} = m_{[t]}/s(a_{[t]})$. If $h(a_{[t]}) = 0$, monotonicity and the recursion give $m_{[t]} = 0$, so we set $p_{[t]} = 0$. Since $m_{[t]} = s(a_{[t]}) p_{[t]}$, this is exactly the success-price recursion in the proposition. □

Proof of Proposition 3. Let

$$S_t^{\mathsf{FB}} := \max_{a \in \mathcal{A}} \{v_t(a) - c(a)\}$$

be type t's first-best surplus. The no-service allocation implies $S_t^{\mathsf{FB}} \geq 0$. Choose $t^* \in \arg\max_t \pi_t S_t^{\mathsf{FB}}$, and let $a_{t^*}^{\mathsf{FB}}$ attain $S_{t^*}^{\mathsf{FB}}$. A UBP seller can offer the single item $(a_{t^*}^{\mathsf{FB}}, p)$ with use price $p = v_{t^*}(a_{t^*}^{\mathsf{FB}})$. Type t^* accepts and yields profit $S_{t^*}^{\mathsf{FB}}$. Any other type who also accepts pays the same price, so the seller's profit from that buyer is $p - c(a_{t^*}^{\mathsf{FB}}) = S_{t^*}^{\mathsf{FB}} \geq 0$. Hence

$$R_{\mathsf{UBP}}^* \geq \pi_{t^*} S_{t^*}^{\mathsf{FB}} \geq \frac{1}{n} \sum_{t=1}^{n} \pi_t S_t^{\mathsf{FB}} = \frac{1}{n} R_{\mathsf{FB}} \geq \frac{1}{n} R_{\mathsf{SB}} \geq \frac{1}{n} R_{\mathsf{PFS}}^*.$$

It remains to show that fully aligned PFS has no analogous guarantee. We give a two-type

example with separable primitives as Definition 2. Fix $\delta \in (0,1]$. Let

$$\mathcal{A} = [0, 1/\delta], \qquad h(a) = a, \qquad s(a) = a, \qquad c(a) = \frac{1}{2}a^2.$$

For $\varepsilon > 0$, set

$$\pi_2 = \varepsilon, \quad \pi_1 = 1 - \varepsilon, \qquad \theta_2 = 1, \quad \theta_1 = \varepsilon^2, \qquad \alpha_2 = \delta, \quad \alpha_1 = \frac{\delta}{2}\varepsilon^2.$$

For small ε, all success probabilities $\gamma_t(a) = \alpha_t a$ lie in $[0,1]$ on $\mathcal{A}$.

By Proposition 1, the low-value UBP virtual type is

$$\widetilde{\theta}_{(1)} = \varepsilon^2 - \frac{\varepsilon}{1-\varepsilon}(1-\varepsilon^2) = -\varepsilon < 0,$$

while $\widetilde{\theta}_{(2)} = 1$. Thus optimal UBP excludes type 1, assigns $a_2 = 1$, and obtains

$$R^*_{\mathsf{UBP}} = \pi_2 \left(1 - \frac{1}{2}\right) = \frac{\varepsilon}{2}.$$

For fully aligned PFS, the per-success values are

$$\phi_2 = \frac{\theta_2}{\alpha_2} = \frac{1}{\delta} < \frac{2}{\delta} = \frac{\theta_1}{\alpha_1} = \phi_1.$$

Hence $[1] = 2$ and $[2] = 1$. By Proposition 2,

$$\widetilde{\phi}_{[1]} = \phi_2 - \frac{\pi_1\alpha_1}{\pi_2\alpha_2}(\phi_1 - \phi_2) = \frac{1}{\delta} - \frac{(1-\varepsilon)\varepsilon}{2\delta}, \qquad \widetilde{\phi}_{[2]} = \phi_1 = \frac{2}{\delta}.$$

Therefore the value-unit coefficients in (7) are

$$\alpha_{[1]}\widetilde{\phi}_{[1]} = 1 - \frac{(1-\varepsilon)\varepsilon}{2}, \qquad \alpha_{[2]}\widetilde{\phi}_{[2]} = \varepsilon^2.$$

The first coefficient is larger than the second, while the PFS allocation must be monotone in the ϕ-order, $a_{[1]} \le a_{[2]}$. Since the objective in (7) is concave, the optimal fully aligned PFS allocation pools the two types: $a_{[1]} = a_{[2]} = a$. The pooled objective is

$$\left[\varepsilon\left(1 - \frac{(1-\varepsilon)\varepsilon}{2}\right) + (1-\varepsilon)\varepsilon^2\right] a - \frac{1}{2}a^2.$$

Thus the optimal pooled allocation is

$$a = \varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{2},$$

which is feasible for small ε, and

$$R^{\mathrm{fa}}_{\mathsf{PFS}} = \frac{1}{2}\left(\varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{2}\right)^2.$$

Consequently,

$$\frac{R^*_{\mathsf{UBP}}}{R^{\mathrm{fa}}_{\mathsf{PFS}}} = \frac{\varepsilon/2}{\frac{1}{2}\left(\varepsilon + \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{2}\right)^2} = \frac{1}{\varepsilon\left(1 + \frac{\varepsilon}{2} - \frac{\varepsilon^2}{2}\right)^2} \longrightarrow +\infty.$$

Thus fully aligned PFS can be arbitrarily worse than UBP.

Finally, we show that the same conclusion holds for optimal PFS with an optimally chosen success metric, for every fixed pair of dispute costs $(\kappa_B, \kappa_S) \in [0,\infty)^2$:

$$\sup \frac{R^*_{\mathsf{UBP}}}{R^*_{\mathsf{PFS}}(\kappa_B,\kappa_S)} = +\infty, \tag{B-8}$$

where $R^*_{\mathsf{PFS}}(\kappa_B,\kappa_S)$ is the optimal PFS profit with success-metric design and the supremum is over the remaining primitives. The idea is a monetary rescaling of the example above: multiplying all values and costs by a large constant leaves the UBP–PFS comparison unchanged when $\kappa_B = 0$, while making any fixed dispute cost negligible relative to the stakes.

Fix $(\kappa_B, \kappa_S) \in [0,\infty)^2$ and $\varepsilon \in (0, 1/2]$, and set

$$M := \frac{\max\{1, \kappa_B\}}{\varepsilon^4}, \qquad \text{so that} \qquad \frac{\kappa_B}{M} \le \varepsilon^4.$$

Consider the same primitives as above, except that all monetary quantities are scaled by M:

$$\theta_1 = M\varepsilon^2, \qquad \theta_2 = M, \qquad c(a) = \frac{M}{2}a^2,$$

with π, α, $\mathcal{A}$, h, and s unchanged. Since the UBP virtual types and the deployment cost scale by M, the optimal UBP menu is unchanged and

$$R^*_{\mathsf{UBP}} = \frac{M\varepsilon}{2}.$$

We now bound $R^*_{\mathsf{PFS}}(\kappa_B,\kappa_S)$ from above. Fix any feasible PFS menu with success-metric design, $\{(a_t, p_t, 0, \sigma_t^1, \sigma_t^0)\}_{t\in\{1,2\}}$ with $p_t \ge 0$, and define type t's on-path expected cost and rent,

$$P_t := C_t(\gamma_t(a_t)) \ge 0, \qquad u_t := v_t(a_t) - P_t.$$

$P_t \ge 0$ holds because PFS payments and buyer dispute costs are nonnegative. Four observations deliver the bound.

First, the seller's expected payoff from each assigned item is at most P_t: the buyer's expected cost P_t equals the expected transfer plus the buyer's expected dispute cost, while the seller's expected payoff equals the expected transfer minus the seller's expected dispute cost. Hence, for every $\kappa_S \ge 0$,

$$R^*_{\mathsf{PFS}}(\kappa_B,\kappa_S) \le \sup\ (1-\varepsilon)\big[P_1 - c(a_1)\big] + \varepsilon\big[P_2 - c(a_2)\big],$$

where the supremum is over feasible menus. Second, IR and $P_t \ge 0$ give

$$0 \le u_t \le v_t(a_t), \qquad \text{so} \qquad P_1 \le M\varepsilon^2 a_1, \qquad P_2 \le M a_2.$$

Third, by Lemma A-1, for a buyer with success probability γ at item 2,

$$C_2(\gamma) = \min\{p_2\eta_2(\gamma),\ \gamma p_2\sigma_2^1 + \kappa_B\,\eta_2(\gamma)\}, \qquad \eta_2(\gamma) := \gamma\sigma_2^1 + (1-\gamma)\sigma_2^0.$$

Both candidate costs are at least $\gamma p_2\sigma_2^1$, so $P_2 = C_2(\gamma_2(a_2)) \ge \gamma_2(a_2)\, p_2\sigma_2^1$. Moreover, choosing the

dispute branch and using $\eta_2(\cdot) \le 1$,

$$C_2(\gamma_1(a_2)) \le \gamma_1(a_2)\, p_2\sigma_2^1 + \kappa_B = \frac{\alpha_1}{\alpha_2}\, \gamma_2(a_2)\, p_2\sigma_2^1 + \kappa_B \le \frac{\varepsilon^2}{2} P_2 + \kappa_B \le \frac{\varepsilon^2}{2} M a_2 + \kappa_B,$$

where we used $\alpha_1/\alpha_2 = \varepsilon^2/2$. This is the key step: success-metric misalignment can raise the mimicking cost of the low-success-probability type by at most κ_B, because that type can always dispute every reported success, paying only for verified true successes while incurring an expected dispute cost of at most κ_B. Fourth, type 1's IC constraint toward item 2, together with $u_1 \le v_1(a_1) = M\varepsilon^2 a_1$, yields

$$M\varepsilon^2 a_1 \ge u_1 \ge v_1(a_2) - C_2(\gamma_1(a_2)) \ge M\varepsilon^2 a_2 - \frac{\varepsilon^2}{2} M a_2 - \kappa_B = \frac{M\varepsilon^2}{2} a_2 - \kappa_B,$$

so that

$$a_2 \le 2a_1 + \frac{2\kappa_B}{M\varepsilon^2} \le 2a_1 + 2\varepsilon^2.$$

Combining these observations, any feasible menu earns at most

$$\begin{aligned} \frac{R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S)}{M} &\le \sup_{\substack{a_1, a_2 \ge 0 \\ a_2 \le 2a_1 + 2\varepsilon^2}} (1-\varepsilon)\Big(\varepsilon^2 a_1 - \frac{1}{2} a_1^2\Big) + \varepsilon\Big(a_2 - \frac{1}{2} a_2^2\Big) \\ &\le \sup_{a_1 \ge 0} \big[(1-\varepsilon)\varepsilon^2 + 2\varepsilon\big] a_1 - \frac{1-\varepsilon}{2} a_1^2 + 2\varepsilon^3 = \frac{\big[(1-\varepsilon)\varepsilon^2 + 2\varepsilon\big]^2}{2(1-\varepsilon)} + 2\varepsilon^3, \end{aligned}$$

where the second inequality drops $-\frac{1}{2}a_2^2 \le 0$ and substitutes $a_2 \le 2a_1 + 2\varepsilon^2$. For $\varepsilon \le 1/2$, we have $(1-\varepsilon)\varepsilon^2 + 2\varepsilon \le \varepsilon(2+\varepsilon) \le \frac{5}{2}\varepsilon$ and $2(1-\varepsilon) \ge 1$, so

$$R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S) \le M\Big(\frac{25}{4}\varepsilon^2 + 2\varepsilon^3\Big).$$

Therefore,

$$\frac{R^*_{\mathsf{UBP}}}{R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S)} \ge \frac{M\varepsilon/2}{M\big(\frac{25}{4}\varepsilon^2 + 2\varepsilon^3\big)} = \frac{1}{\frac{25}{2}\varepsilon + 4\varepsilon^2} \xrightarrow[\varepsilon \downarrow 0]{} +\infty.$$

Since (κ_B, κ_S) was arbitrary and held fixed along the sequence, this proves (B-8). Note that the construction scales the monetary primitives with $1/\varepsilon^4$: what drives the result is that the dispute cost κ_B becomes small relative to the monetary stakes, not that κ_B is small in absolute terms. □

B.1.2 Proofs for Section 4.2

Proof of Theorem 1. We prove each result separately.

Proof of (i): bounds.

We first prove the chain of inequalities. Fully aligned PFS is nested in PFS with success-metric design by setting $\sigma_r^1 = 1$ and $\sigma_r^0 = 0$ for every r, so

$$R^{\mathrm{fa}}_{\mathsf{PFS}} \le R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S).$$

To prove $R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S) \le R^{\mathrm{fa}}_{\mathsf{LL}}$, fix an arbitrary feasible PFS menu with success-metric design,

$$\mathcal{M} = \{(a_r, p_r, q_r, \sigma_r^1, \sigma_r^0)\}_{r\in\mathcal{T}}, \qquad q_r = 0, \quad p_r \ge 0, \quad \sigma_r^1, \sigma_r^0 \in [0,1].$$

We construct another menu,

$$\widetilde{\mathcal{M}} = \{(a_r, \widetilde{p}_r, \widetilde{q}_r, 1, 0)\}_{r\in\mathcal{T}}, \qquad \widetilde{p}_r, \widetilde{q}_r \ge 0,$$

with full outcome alignment, outcome-contingent prices, and the same allocations. We show that this new menu preserves IC and IR and weakly increases the seller's profit.

Fix an item r. For a buyer whose success probability at allocation a_r is γ, the success-metric rule in $\mathcal{M}$ generates reported success with probability

$$\eta_r(\gamma) := \gamma\sigma_r^1 + (1-\gamma)\sigma_r^0.$$

By Lemma A-1, item r induces the following buyer cost for a buyer whose success probability at allocation a_r is γ:

$$C_r(\gamma) = \min\left\{p_r\eta_r(\gamma), \gamma p_r\sigma_r^1 + \kappa_B\eta_r(\gamma)\right\}.$$

The two terms inside the minimum are affine in γ. Denote these two candidate cost lines by

$$f_r^0(\gamma) := p_r\eta_r(\gamma), \qquad f_r^1(\gamma) := \gamma p_r\sigma_r^1 + \kappa_B\eta_r(\gamma).$$

Thus $C_r(\gamma) = \min\{f_r^0(\gamma), f_r^1(\gamma)\}$. At the intended success probability $\gamma_r(a_r)$, choose one line $\ell_r \in \{f_r^0, f_r^1\}$ that attains this minimum. Then

$$\ell_r(\gamma_r(a_r)) = C_r(\gamma_r(a_r)), \qquad \ell_r(\gamma) \ge C_r(\gamma) \quad \forall\gamma \in [0,1],$$

because C_r is the pointwise minimum of the two candidate lines.

We now use ℓ_r as the fully aligned expected-payment line for item r. An outcome-contingent fully aligned item with failure bill $\widetilde{q}_r$ and success bill $\widetilde{p}_r$ generates expected payment

$$(1-\gamma)\widetilde{q}_r + \gamma\widetilde{p}_r = \widetilde{q}_r + (\widetilde{p}_r - \widetilde{q}_r)\gamma.$$

Therefore set

$$\widetilde{q}_r := \ell_r(0), \qquad \widetilde{p}_r := \ell_r(1).$$

These bills are nonnegative because the two candidate lines have endpoint values

$$p_r\sigma_r^0, \quad p_r\sigma_r^1, \quad \kappa_B\sigma_r^0, \quad (p_r+\kappa_B)\sigma_r^1,$$

all of which are nonnegative. Hence the replacement item $(a_r, \widetilde{p}_r, \widetilde{q}_r, 1, 0)$ satisfies limited liability.

Doing this for every r gives $\widetilde{\mathcal{M}}$. For each intended report r, fully aligned buyer cost is unchanged:

$$\ell_r(\gamma_r(a_r)) = C_r(\gamma_r(a_r)).$$

For any deviation by true type t to report r, buyer cost weakly increases:

$$\ell_r(\gamma_t(a_r)) \geq C_r(\gamma_t(a_r)).$$

Therefore fully aligned utility is unchanged and every deviation becomes weakly less attractive, so IC and IR are preserved.

It remains to compare seller payoff. Under the original item r, buyer cost equals the expected transfer plus the buyer's expected dispute cost, while seller payoff equals the expected transfer minus the seller's expected dispute cost. Hence the original seller payoff from item r is weakly below $C_r(\gamma_r(a_r))$ due to the potential dispute cost. The replacement item has full outcome alignment, no dispute waste, and expected payment $\ell_r(\gamma_r(a_r)) = C_r(\gamma_r(a_r))$. Thus seller payoff weakly increases for every intended item, while allocation costs are unchanged. Hence every feasible PFS menu with success-metric design is weakly dominated by a fully outcome-aligned limited-liability menu.

Proof of (i)(a).

When $\kappa_B = 0$, the buyer can freely dispute any reported success. Formally, by Lemma A-1 with $\kappa_B = 0$, a buyer with success probability γ at item r has cost $C_r(\gamma) = \min\{p_r\eta_r(\gamma), \gamma p_r\sigma_r^1\} = \gamma p_r\sigma_r^1$, since $\eta_r(\gamma) \geq \gamma\sigma_r^1$. Replace item r by the fully aligned PFS item $(a_r,\, p_r\sigma_r^1,\, 0,\, 1,\, 0)$: every type's cost line is unchanged at $\gamma p_r\sigma_r^1$, so IC and IR are preserved. Expected transfers are also unchanged, no disputes occur under full alignment, and the seller saves the dispute cost κ_S whenever the original item was disputed. Applying this replacement item by item weakly raises profit, so $R^*_{\mathsf{PFS}}(0, \kappa_S) = R^{\text{fa}}_{\mathsf{PFS}}$.

Proof of (i): comparative statics.

Fix $\kappa_B' > \kappa_B$. If $\kappa_B = 0$, part (i) and the nesting of fully aligned PFS in PFS with success-metric design imply

$$R^*_{\mathsf{PFS}}(0, \kappa_S) = R^{\text{fa}}_{\mathsf{PFS}} \leq R^*_{\mathsf{PFS}}(\kappa_B', \kappa_S).$$

Now suppose $\kappa_B > 0$. Fix any feasible PFS with success-metric-design menu for (κ_B, κ_S). For each item with $p_r = 0$, leave the item unchanged. For each item with $p_r > 0$, set

$$\lambda := \frac{\kappa_B'}{\kappa_B} \geq 1, \qquad p_r' := \lambda p_r, \qquad \sigma_r^{1\prime} := \frac{\sigma_r^1}{\lambda}, \qquad \sigma_r^{0\prime} := \frac{\sigma_r^0}{\lambda}.$$

The effective bills are unchanged:

$$p_r'\sigma_r^{1\prime} = p_r\sigma_r^1, \qquad p_r'\sigma_r^{0\prime} = p_r\sigma_r^0,$$

and the reported-success probability is scaled down:

$$\eta_r'(\gamma) = \gamma\sigma_r^{1\prime} + (1-\gamma)\sigma_r^{0\prime} = \frac{\eta_r(\gamma)}{\lambda}.$$

Therefore Lemma A-1 gives, for every γ,

$$C'_r(\gamma) = \min\{p'_r\eta'_r(\gamma), \gamma p'_r\sigma_r^{1\prime} + \kappa'_B\eta'_r(\gamma)\} = \min\{p_r\eta_r(\gamma), \gamma p_r\sigma_r^1 + \kappa_B\eta_r(\gamma)\} = C_r(\gamma).$$

Thus all IR and IC constraints are unchanged. Seller payoff is weakly higher: in the no-dispute case, the expected reported-outcome bill $p_r\eta_r(\gamma)$ is unchanged; in the dispute case, the verified expected bill $\gamma p_r\sigma_r^1$ is unchanged, while the seller pays the dispute cost with probability $\eta'_r(\gamma) = \eta_r(\gamma)/\lambda \le \eta_r(\gamma)$. Taking suprema proves weak monotonicity in κ_B.

For monotonicity in the seller dispute cost, fix $\kappa'_S \ge \kappa_S$. The feasible set and the buyer-optimal dispute set do not depend on κ_S. For any fixed PFS with success-metric-design menu and any buyer-optimal dispute rule, increasing κ_S only subtracts additional seller dispute cost from the objective, item by item. Hence, the seller's profit is weakly lower at κ'_S than at κ_S, and taking suprema gives

$$R^*_{\mathsf{PFS}}(\kappa_B, \kappa'_S) \le R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S).$$

Proof of (i)(b). Take any fully outcome-aligned limited-liability item with success payment p^{LL} and failure payment q^{LL}. If $p^{\mathsf{LL}} = q^{\mathsf{LL}} = 0$, implement it by the zero PFS item. Otherwise set

$$p^{\mathsf{PFS}} := \max\{p^{\mathsf{LL}}, q^{\mathsf{LL}}\}, \qquad \sigma^1 := \frac{p^{\mathsf{LL}}}{p^{\mathsf{PFS}}}, \qquad \sigma^0 := \frac{q^{\mathsf{LL}}}{p^{\mathsf{PFS}}}.$$

Then the expected PFS bill without disputes is

$$p^{\mathsf{PFS}}[\gamma\sigma^1 + (1-\gamma)\sigma^0] = \gamma p^{\mathsf{LL}} + (1-\gamma)q^{\mathsf{LL}},$$

which is exactly the expected bill of the fully aligned limited-liability item.

For sufficiently large κ_B, no type disputes reported successes. Indeed, by Lemma A-1, the dispute condition is

$$(1-\gamma)p^{\mathsf{PFS}}\sigma^0 > \kappa_B[\gamma\sigma^1 + (1-\gamma)\sigma^0].$$

If $\kappa_B \ge p^{\mathsf{PFS}}$, the right-hand side is at least

$$p^{\mathsf{PFS}}[\gamma\sigma^1 + (1-\gamma)\sigma^0] \ge (1-\gamma)p^{\mathsf{PFS}}\sigma^0,$$

so the strict dispute condition cannot hold. Thus the constructed PFS item implements the same buyer payments and the same seller profit as the fully aligned limited-liability item.

Applying this construction to every item of an optimal limited-liability menu, let

$$\bar\kappa_B := \max_t \max\{p_t^{\mathsf{LL}}, q_t^{\mathsf{LL}}\}.$$

For every $\kappa_B \ge \bar\kappa_B$, no constructed item is disputed, so

$$R^*_{\mathsf{PFS}}(\kappa_B, \kappa_S) = R^{\mathrm{fa}}_{\mathsf{LL}}.$$

Finally, UBP is a fully aligned limited-liability subclass, so

$$R^{\text{fa}}_{\mathsf{LL}} \geq R^{*}_{\mathsf{UBP}}.$$

Proof of (ii)(a).

Because the value order and the per-success order coincide, write the common order as $(1), \ldots, (n)$. Fix any allocation profile $(a_{(t)})_{t=1}^{n}$ that is monotone in this order, and set $h(a_{(0)}) = 0$.

By Proposition 1, the UBP reduced objective at this allocation is

$$\sum_{t=1}^{n} \pi_{(t)} \left[\widetilde{\theta}_{(t)} h(a_{(t)}) - c(a_{(t)})\right].$$

By Proposition 2, the fully aligned PFS reduced objective at the same allocation is

$$\sum_{t=1}^{n} \pi_{(t)} \left[\alpha_{(t)} \widetilde{\phi}_{(t)} h(a_{(t)}) - c(a_{(t)})\right].$$

The cost terms are identical, so we compare only the virtual-value terms. Using

$$h(a_{(t)}) = \sum_{j=1}^{t} \left[h(a_{(j)}) - h(a_{(j-1)})\right],$$

and swapping the order of summation, the UBP virtual-value term is

$$\begin{aligned}
\sum_{t=1}^{n} \pi_{(t)} \widetilde{\theta}_{(t)} h(a_{(t)}) &= \sum_{j=1}^{n} \left(\sum_{t=j}^{n} \pi_{(t)} \widetilde{\theta}_{(t)} \right) \left[h(a_{(j)}) - h(a_{(j-1)})\right] \\
&= \sum_{j=1}^{n} \left(\sum_{t=j}^{n} \pi_{(t)} \right) \theta_{(j)} \left[h(a_{(j)}) - h(a_{(j-1)})\right],
\end{aligned}$$

where the last equality follows from the definition of the UBP virtual type $\tilde{\theta}_{(t)}$. Similarly, using the truthful-PFS virtual type,

$$\sum_{t=1}^{n} \pi_{(t)} \alpha_{(t)} \widetilde{\phi}_{(t)} h(a_{(t)}) = \sum_{j=1}^{n} \left(\sum_{t=j}^{n} \pi_{(t)} \alpha_{(t)} \right) \phi_{(j)} \left[h(a_{(j)}) - h(a_{(j-1)})\right]. \tag{B-9}$$

Therefore fully aligned PFS minus UBP at this fixed allocation equals

$$\sum_{j=1}^{n} \left[\left(\sum_{t=j}^{n} \pi_{(t)} \alpha_{(t)} \right) \phi_{(j)} - \left(\sum_{t=j}^{n} \pi_{(t)} \right) \theta_{(j)} \right] \left[h(a_{(j)}) - h(a_{(j-1)})\right].$$

If

$$\frac{\sum_{t=j}^{n} \pi_{(t)} \alpha_{(t)}}{\sum_{t=j}^{n} \pi_{(t)}} \geq \alpha_{(j)} \qquad \forall j,$$

then, since $\theta_{(j)} = \alpha_{(j)} \phi_{(j)}$, every bracketed coefficient above is nonnegative. Monotonicity of the allocation gives $h(a_{(j)}) - h(a_{(j-1)}) \geq 0$. Hence fully aligned PFS weakly dominates UBP at every

common monotone allocation. Evaluating this at the UBP-optimal allocation gives

$$R^{\text{fa}}_{\mathsf{PFS}} \geq R^{*}_{\mathsf{UBP}}.$$

The case with the reverse tail-average inequalities is identical, with all inequalities reversed.

Proof of (ii)(b).

By Proposition 1, R^{*}_{UBP} is independent of success probability, and thus independent of β. It remains to study $R^{\text{fa}}_{\mathsf{PFS}}(\beta)$.

Fix β, under the power-relation $\alpha_t = \theta_t^{\beta}$, we have

$$\phi_t(\beta) = \frac{\theta_t}{\alpha_t(\beta)} = \theta_t^{1-\beta}.$$

By the same summation-by-parts argument in (B-9), the coefficient on the increment $h(a_{[t]}) - h(a_{[t-1]})$ in the fully aligned PFS reduced objective is

$$\theta_{[t]}^{1-\beta} \sum_{k=t}^{n} \pi_{[k]} \theta_{[k]}^{\beta}.$$

For fixed t, its logarithmic derivative with respect to β is

$$-\log \theta_{[t]} + \frac{\sum_{k=t}^{n} \pi_{[k]} \theta_{[k]}^{\beta} \log \theta_{[k]}}{\sum_{k=t}^{n} \pi_{[k]} \theta_{[k]}^{\beta}}.$$

If $\beta < 1$, the ϕ-order is the same as the value order, so $\theta_{[k]} \geq \theta_{[t]}$ for every $k \geq t$. Hence this derivative is nonnegative. If $\beta > 1$, the ϕ-order is the reverse value order, so $\theta_{[k]} \leq \theta_{[t]}$ for every $k \geq t$, and the derivative is nonpositive. Since all increments $h(a_{[t]}) - h(a_{[t-1]})$ are nonnegative, the fully aligned PFS reduced objective is weakly increasing on the feasible part of $(-\infty, 1]$ and weakly decreasing on the feasible part of $[1, \infty)$. At $\beta = 1$, $\phi_t(1) \equiv 1$, i.e., all types have the same per-success value. Thus, fully aligned PFS can extract the first-best surplus. Therefore $R^{\text{fa}}_{\mathsf{PFS}}(1) \geq R^{*}_{\mathsf{UBP}}$.

At $\beta = 0$, $\alpha_t = 1$ and $\phi_t = \theta_t$ for every type. Thus, fully aligned PFS and UBP screen buyers on the same index and, crucially, also generate the same expected transfers. Indeed, for a PFS item (a_r, p_r), let $m_r := s(a_r)p_r$. A type-t buyer's utility is $\theta_t h(a_r) - m_r$, while the seller's expected transfer is m_r, exactly as under a UBP item with usage price m_r. Conversely, any positive-service UBP item with price m_r can be implemented under PFS by setting $p_r = m_r/s(a_r)$. Hence $R^{\text{fa}}_{\mathsf{PFS}}(0) = R^{*}_{\mathsf{UBP}}$. The monotonicity established above then immediately implies $R^{\text{fa}}_{\mathsf{PFS}}(\beta) \leq R^{*}_{\mathsf{UBP}}$ for $\beta < 0$, and $R^{\text{fa}}_{\mathsf{PFS}}(\beta) \geq R^{*}_{\mathsf{UBP}}$ for $0 < \beta \leq 1$. □

B.1.3 Proofs for Section 4.3

Proof of Corollary 1. Write $v_t := V(N_t, \lambda_t, \rho_t, \gamma_t)$ and $\phi_t := v_t/\gamma_t$. It is easy to show that

$$R^*_{\mathsf{UBP}} = \max\left\{0, \max_{i\in\{1,2\}} \sum_{j:v_j\geq v_i} \pi_j(v_i - c_0)\right\}, \tag{B-10}$$

and

$$R^{\mathrm{fa}}_{\mathsf{PFS}} = \max\left\{0, \max_{i\in\{1,2\}} \sum_{j:\phi_j\geq \phi_i} \pi_j(\gamma_j\phi_i - c_0)\right\}. \tag{B-11}$$

Assume $\gamma_2 > \gamma_1$. Define

$$\bar{\gamma} := \pi_1\gamma_1 + \pi_2\gamma_2, \qquad v^\dagger := \max\left\{\frac{\gamma_2}{\bar{\gamma}}v_1, \frac{\gamma_2}{\gamma_1}\min\{v_1, c_0\}\right\}.$$

By the congestion property, $V(N_2, \lambda_2, \rho, \gamma_2)$ is continuous and strictly increasing in ρ, with limit 0 as $\rho \downarrow 0$ and limit ∞ as $\rho \uparrow 1$. Hence there is a unique $\rho^\dagger \in (0,1)$ satisfying

$$V(N_2, \lambda_2, \rho^\dagger, \gamma_2) = v^\dagger.$$

Since $v_2 = V(N_2, \lambda_2, \rho_2, \gamma_2)$ is strictly increasing in ρ_2, it remains only to verify the following value-threshold comparison:

$$v_2 > v^\dagger \quad\Longrightarrow\quad R^{\mathrm{fa}}_{\mathsf{PFS}} \geq R^*_{\mathsf{UBP}},$$

and

$$v_2 \leq v^\dagger \quad\Longrightarrow\quad R^*_{\mathsf{UBP}} \geq R^{\mathrm{fa}}_{\mathsf{PFS}}.$$

This can be verified case by case by comparing (B-10) and (B-11). We omit the details here. □

B.2 Proofs for Section 5

B.2.1 Proofs for Section 5.1

Proof of Theorem 2. Consider any feasible menu

$$\mathcal{M} = \{(a_s, p_s, q_s, \sigma^1_s, \sigma^0_s)\}_{s\in\mathcal{T}}$$

in (3) and keep the allocation profile fixed. Consider one item t. If a buyer chooses this item and has success probability γ at a_t, let $C_t(\gamma)$, see (1), denote her minimum expected cost, including any buyer dispute cost. For a fixed pay-or-dispute rule $\delta = (\delta_0, \delta_1) \in \{0,1\}^2$, define the corresponding fixed-rule cost by $\underline{C}_{t,\delta}(\gamma)$. The joint probabilities of $(\omega, \widehat{\omega})$ under item t are

$$\begin{aligned} \mathbb{P}(1,1) &= \gamma\sigma^1_t, & \mathbb{P}(1,0) &= \gamma(1-\sigma^1_t),\\ \mathbb{P}(0,1) &= (1-\gamma)\sigma^0_t, & \mathbb{P}(0,0) &= (1-\gamma)(1-\sigma^0_t), \end{aligned}$$

so $\underline{C}_{t,\delta}(\gamma)$ is

$$\begin{aligned}\underline{C}_{t,\delta}(\gamma) = \; & \underbrace{\gamma\sigma_t^1}_{\mathbb{P}(\omega=1,\widehat{\omega}=1)} \underbrace{\big[(1-\delta_1)p_t + \delta_1(p_t+\kappa_B)\big]}_{\substack{\text{reported success;}\\ \text{no dispute: } p_t\\ \text{dispute: } p_t+\kappa_B}} \\ & + \underbrace{\gamma(1-\sigma_t^1)}_{\mathbb{P}(\omega=1,\widehat{\omega}=0)} \underbrace{\big[(1-\delta_0)q_t + \delta_0(p_t+\kappa_B)\big]}_{\substack{\text{reported failure;}\\ \text{no dispute: } q_t\\ \text{dispute: } p_t+\kappa_B}} \\ & + \underbrace{(1-\gamma)\sigma_t^0}_{\mathbb{P}(\omega=0,\widehat{\omega}=1)} \underbrace{\big[(1-\delta_1)p_t + \delta_1(q_t+\kappa_B)\big]}_{\substack{\text{reported success;}\\ \text{no dispute: } p_t\\ \text{dispute: } q_t+\kappa_B}} \\ & + \underbrace{(1-\gamma)(1-\sigma_t^0)}_{\mathbb{P}(\omega=0,\widehat{\omega}=0)} \underbrace{\big[(1-\delta_0)q_t + \delta_0(q_t+\kappa_B)\big]}_{\substack{\text{reported failure;}\\ \text{no dispute: } q_t\\ \text{dispute: } q_t+\kappa_B}}.\end{aligned}$$

Each bracketed term is independent of γ, and each probability coefficient is affine in γ. Hence every $\underline{C}_{t,\delta}(\gamma)$ is affine in γ, and

$$C_t(\gamma) = \min_{\delta\in\{0,1\}^2} \underline{C}_{t,\delta}(\gamma)$$

is finite, piecewise affine, and concave. Therefore, at the assigned success probability $\gamma_t(a_t)$, there is an affine branch ℓ_t such that

$$\ell_t(\gamma_t(a_t)) = C_t(\gamma_t(a_t)), \qquad \ell_t(\gamma) \geq C_t(\gamma) \quad \forall\gamma\in[0,1].$$

Writing

$$\ell_t(\gamma) = \widetilde{q}_t + (\widetilde{p}_t - \widetilde{q}_t)\gamma, \qquad \widetilde{q}_t := \ell_t(0), \quad \widetilde{p}_t := \ell_t(1),$$

gives the fully outcome-aligned replacement: replace item t by $(a_t, \widetilde{p}_t, \widetilde{q}_t, 1, 0)$. Applying this construction item by item gives

$$\widetilde{\mathcal{M}} = \{(a_s, \widetilde{p}_s, \widetilde{q}_s, 1, 0)\}_{s\in\mathcal{T}}.$$

We now check the constraints and objective. For each true type t, her truthful buyer cost is unchanged:

$$v_t(a_t) - \ell_t(\gamma_t(a_t)) = v_t(a_t) - C_t(\gamma_t(a_t)) \geq 0,$$

so IR is preserved. For IC, fix true type t and deviation report s. Since the original menu is IC,

$$v_t(a_t) - C_t(\gamma_t(a_t)) \geq v_t(a_s) - C_s(\gamma_t(a_s)).$$

Moreover, by construction, $\ell_t(\gamma_t(a_t)) = C_t(\gamma_t(a_t))$, and because ℓ_s is a supporting affine branch of

the lower envelope C_s, $\ell_s(\gamma) \geq C_s(\gamma)$ for all γ. Therefore

$$\begin{aligned} v_t(a_t) - \ell_t(\gamma_t(a_t)) &= v_t(a_t) - C_t(\gamma_t(a_t)) \\ &\geq v_t(a_s) - C_s(\gamma_t(a_s)) \\ &\geq v_t(a_s) - \ell_s(\gamma_t(a_s)). \end{aligned}$$

Thus every deviation is weakly less attractive under $\widetilde{\mathcal{M}}$.

Finally, compare the type-t term in the seller's objective. Under the original item t, let D_t be the probability that type t's buyer disputes under her optimal dispute rule. Since buyer cost equals expected payment plus $\kappa_B D_t$, expected payment equals $C_t(\gamma_t(a_t)) - \kappa_B D_t$. The seller also pays dispute cost $\kappa_S D_t$. Thus the original type-t objective term is

$$C_t(\gamma_t(a_t)) - (\kappa_B + \kappa_S) D_t - c(a_t) \leq C_t(\gamma_t(a_t)) - c(a_t).$$

Under $\widetilde{\mathcal{M}}$, full outcome alignment makes disputes unprofitable, so the type-t objective term is

$$\ell_t(\gamma_t(a_t)) - c(a_t) = C_t(\gamma_t(a_t)) - c(a_t).$$

Thus the seller's objective is weakly higher for every true type, while IC and IR are preserved. Therefore $\widetilde{\mathcal{M}}$ weakly dominates $\mathcal{M}$. This completes the proof. □

B.2.2 Proofs for Section 5.2

Proof of Lemma 1. *Necessity.* Suppose the allocation-rent profile $\big((a_t, u_t)\big)_{t \in \mathcal{T}}$ is implementable. Then there exist prices $(p_t, q_t)_{t \in \mathcal{T}}$ such that the rent identity (14) and the IC constraints (IC) hold. For each intended type t, write its expected-payment line as

$$\ell_t(x) := q_t + (p_t - q_t)x.$$

The rent identity gives

$$\ell_t(\gamma_t(a_t)) = v_t(a_t) - u_t,$$

so ℓ_t passes through the intended point $(\gamma_t(a_t), v_t(a_t) - u_t)$. If another buyer r deviates to item t, incentive compatibility gives

$$u_r \geq v_r(a_t) - \ell_t(\gamma_r(a_t)),$$

which rearranges to

$$\ell_t(\gamma_r(a_t)) \geq v_r(a_t) - u_r.$$

That is, ℓ_t lies weakly above every mimicking type's point.

Sufficiency. Suppose $u_t \geq 0$ for all t and each intended point is upper-supported. For each intended type t, choose a supporting affine line and set $q_t := \ell_t(0)$ and $p_t := \ell_t(1)$. Then, by the definition

of upper support, as illustrated in Figure 5a, the rent identity is satisfied and IC is preserved. IR follows from $u_t \geq 0$. □

Proof of Corollary 2. We prove each part separately.

Proof of (i).

Suppose a first-best allocation profile $(a_t^{\mathsf{FB}})_{t\in\mathcal{T}}$ satisfies the stated type-by-type upper-support condition. Setting $u_t = 0$ for every type, Lemma 1 implements $(a_t^{\mathsf{FB}})_{t\in\mathcal{T}}$ with zero buyer rents, so profit equals R_{FB}.

Conversely, suppose $R_{\mathsf{SB}} = R_{\mathsf{FB}}$, and let $((a_t, u_t))_{t\in\mathcal{T}}$ be an implemented allocation-rent profile attaining it. Then

$$R_{\mathsf{FB}} = R_{\mathsf{SB}} = \sum_t \pi_t[v_t(a_t) - c(a_t) - u_t] \leq \sum_t \pi_t[v_t(a_t) - c(a_t)] \leq R_{\mathsf{FB}}.$$

Both inequalities are equalities. The first forces $\sum_t \pi_t u_t = 0$, and since $\pi_t > 0$ and $u_t \geq 0$, it follows that $u_t = 0$ for every t; the second says the allocation profile $(a_t)_{t\in\mathcal{T}}$ is first-best. Applying Lemma 1 with $u_t = 0$ for every $t \in \mathcal{T}$ yields the upper-support condition.

Proof of (ii).

Under the separable primitives $\gamma_r(a) = \alpha_r s(a)$ and $v_r(a) = \theta_r h(a)$, the zero-rent point set for intended report t at the selected first-best allocation is

$$\{(\alpha_r s(a_t^{\mathsf{FB}}),\, \theta_r h(a_t^{\mathsf{FB}})) : r \in \mathcal{T}\}.$$

If $h(a_t^{\mathsf{FB}}) = 0$, all vertical coordinates vanish, so the intended point is trivially upper-supported. If $h(a_t^{\mathsf{FB}}) > 0$, then $s(a_t^{\mathsf{FB}}) > 0$ by assumption to avoid corner cases. Therefore the support condition in part (i) is just the same support condition in the primitive (α, θ)-space, up to positive horizontal and vertical rescaling, and is independent of the allocation level a_t^{FB}.

If θ is discretely concave in α, then the primitive points $\{(\alpha_r, \theta_r) : r \in \mathcal{T}\}$ lie on a concave curve. Hence, by the supporting-hyperplane theorem for concave functions, every point (α_t, θ_t) admits a supporting line crossing it and lying weakly above all the other primitive points. Rescaling that line back gives the required report-specific supporting line for every t. Thus the condition in part (i) holds, and $R_{\mathsf{SB}} = R_{\mathsf{FB}}$. □

Proof of Theorem 3. By Corollary A-1, the truthful second-best problem can be written in the allocation and rent variables. We first dispose of no-service items. If $a_t = 0$, then $\gamma_r(0) = v_r(0) = 0$ for every r, so all types share the same success probability and gross value at item t—their adjusted points $(0, -u_r)$ differ only through rents—and the same-success constraints in Corollary A-1 reduce to $u_t \leq u_r$ for every $r \in \mathcal{T}$. If $t \in \{1, n\}$, such an item cannot arise at an attained optimum: replace item t by (a^*, u_t) with $a^* \in \arg\max_{a\in\mathcal{A}}\{v_t(a) - c(a)\}$. Since $\max_{a\in\mathcal{A}}\{v_t(a) - c(a)\} > 0 = v_t(0) - c(0)$, we have $a^* \neq 0$ and hence $s(a^*) > 0$; item t then faces mimicking types on only one

side of the success-probability axis, so a sufficiently steep payment line upper-supports it for any rent, while the constraints of all other items are unchanged because type t's rent is unchanged. The objective strictly increases by $\pi_t \max_{a\in\mathcal{A}}\{v_t(a) - c(a)\} > 0$, a contradiction. Hence endpoint items are positive-service at any attained optimum. If instead t is interior with $a_t = 0$, then $u_t = 0$ without loss at an attained optimum: u_t is a minimum rent, and if $u_t > 0$, subtracting u_t from every rent preserves nonnegativity, leaves all same-success and chord constraints unchanged, and strictly increases profit. With $u_t = 0$, item t's same-success constraints hold and its chord constraints below are automatic. We record this boundary condition by restricting the domain to $(a_t,\, u_t) \in \{(0,0)\} \cup ((\mathcal{A} \setminus \{0\}) \times \mathbb{R}_+)$, which is a convex set, and, for positive-service items, use $s(a_t) > 0$ in what follows. Under $\gamma_j(a_t) = \alpha_j s(a_t)$, for any $r < t < \tau$,

$$\lambda_t^{r\tau}(a) = \frac{\gamma_\tau(a_t) - \gamma_t(a_t)}{\gamma_\tau(a_t) - \gamma_r(a_t)} = \frac{(\alpha_\tau - \alpha_t)s(a_t)}{(\alpha_\tau - \alpha_r)s(a_t)} = \frac{\alpha_\tau - \alpha_t}{\alpha_\tau - \alpha_r} =: \lambda_t^{r\tau}.$$

Thus the support constraints no longer depend on the allocation through $\lambda_t^{r\tau}$ for positive-service items; for a no-service item, the boundary condition $u_t = 0$ replaces its constraints. The reduced problem is

$$\begin{aligned} R_{\mathsf{SB}} = \max_{\substack{(a_t,\, u_t)\in\{(0,0)\}\cup\left((\mathcal{A}\setminus\{0\})\times\mathbb{R}_+\right)\\ t\in\mathcal{T}}} & \sum_{t\in\mathcal{T}} \pi_t\big[v_t(a_t) - c(a_t) - u_t\big] \qquad \text{(B-12)}\\ \text{s.t.}\quad v_t(a_t) - u_t \geq \lambda_t^{r\tau}\big[v_r(a_t) - u_r\big] & + \big(1 - \lambda_t^{r\tau}\big)\big[v_\tau(a_t) - u_\tau\big], \qquad \forall r < t < \tau. \end{aligned}$$

Proof of (i).

In the reduced formulation, each allocation a_t appears only in the support constraints for which t is the intended type:

$$v_t(a_t) - u_t \geq \lambda_t^{r\tau}\big[v_r(a_t) - u_r\big] + (1 - \lambda_t^{r\tau})\big[v_\tau(a_t) - u_\tau\big], \qquad r < t < \tau.$$

Thus a_1 can appear only in the constraints for intended type 1, and a_n can appear only in the constraints for intended type n. But type 1 has no left bracketing type $r < 1$, and type n has no right bracketing type $\tau > n$. Hence, there is no constraint for intended reports 1 and n. Therefore, a_1 and a_n can be chosen freely. (By the boundary step, endpoint items are positive-service at any attained optimum, so the boundary restriction at $a_t = 0$ does not bind for them; and since $\max_{a\in\mathcal{A}}\{v_t(a) - c(a)\} > 0$, every maximizer of $v_t(a) - c(a)$ is a positive-service allocation, so the replacement below never activates it.) The a_1 enters the reduced problem only through the objective term

$$\pi_1[v_1(a_1) - c(a_1)],$$

and a_n enters only through

$$\pi_n[v_n(a_n) - c(a_n)].$$

Therefore, at any attained optimum, replacing a_1 by a maximizer of $v_1(a) - c(a)$, or replacing a_n by a maximizer of $v_n(a) - c(a)$, leaves feasibility unchanged and weakly increases the objective. Hence

$$a_1^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}}\{v_1(a) - c(a)\}, \qquad a_n^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}}\{v_n(a) - c(a)\}.$$

Proof of (ii)(a).

Now impose $v_t(a) = \theta_t h(a)$. Let $z_t = h(a_t)$, let $\mathcal{Z} = h(\mathcal{A})$, and define $\tilde{c}(z) = c(h^{-1}(z))$. Substituting $v_j(a_t) = \theta_j z_t$ into (B-12), the constraint for $r < t < \tau$ becomes

$$\theta_t z_t - u_t \geq \lambda_t^{r\tau}(\theta_r z_t - u_r) + (1 - \lambda_t^{r\tau})(\theta_\tau z_t - u_\tau).$$

Rearranging gives linear constraints

$$u_t - \lambda_t^{r\tau} u_r - (1 - \lambda_t^{r\tau})u_\tau \leq z_t\Big[\theta_t - \lambda_t^{r\tau}\theta_r - (1 - \lambda_t^{r\tau})\theta_\tau\Big].$$

By the definition of $\Delta_t^{r\tau}$, this is exactly (19). The objective becomes

$$\sum_{t\in\mathcal{T}} \pi_t[\theta_t z_t - \tilde{c}(z_t) - u_t].$$

Because h is strictly increasing and $h(0) = 0$, the domain restriction on (a_t, u_t) is equivalent to $(z_t,\, u_t) \in \{(0,0)\} \cup ((\mathcal{Z} \setminus \{0\}) \times \mathbb{R}_+)$. Hence the problem is equivalent to (18). If $\mathcal{Z}$ and $\tilde{c}$ are convex, problem (18) is a convex program: the objective is concave because $-\tilde{c}$ is concave, the chord constraints are affine, and the maximization domain is convex in each (z_t, u_t)-plane: if a convex combination of two feasible points has $z_t = 0$, then $z_t = 0$ at both points because $z_t \geq 0$, so $u_t = 0$ at both, and hence at the combination. The feasible set is therefore an intersection of convex sets.

Proof of (ii)(b).

Suppose the program admits a primal–dual saddle point with nonnegative multipliers $m_t^{r\tau} \geq 0$ on the chord constraints (19). Dualizing only these constraints and retaining each (z_t, u_t) in the primal domain $\{(0,0)\} \cup ((\mathcal{Z} \setminus \{0\}) \times \mathbb{R}_+)$, the Lagrangian is

$$\mathcal{L} = \sum_{t\in\mathcal{T}} \pi_t[\theta_t z_t - \tilde{c}(z_t) - u_t] - \sum_{r<t<\tau} m_t^{r\tau}\left[u_t - \lambda_t^{r\tau} u_r - (1 - \lambda_t^{r\tau})u_\tau - z_t\Delta_t^{r\tau}\right].$$

Fix an interior type t. Collecting the terms involving z_t, the z_t-dependent part of the Lagrangian is

$$\pi_t\big[\widehat{\theta}_t z_t - \tilde{c}(z_t)\big],$$

where

$$\widehat{\theta}_t := \theta_t + \frac{1}{\pi_t}\sum_{r<t<\tau} m_t^{r\tau}\Delta_t^{r\tau}.$$

The remaining dependence on u_t is linear, and its coefficient must be nonpositive at the saddle point; if it were positive, fixing any nonzero z_t and letting $u_t \to \infty$ would make the Lagrangian unbounded above.

The boundary condition $z_t = 0 \Longrightarrow u_t = 0$ prevents us from separating the maximizations over z_t and u_t immediately. However, $(z, 0)$ is admissible for every $z \in \mathcal{Z}$. Comparing the saddle-point choice $(z_t^{\mathsf{SB}}, u_t^{\mathsf{SB}})$ with $(z_t^{\mathsf{SB}}, 0)$, saddle-point maximality implies that the u_t-term at $(z_t^{\mathsf{SB}}, u_t^{\mathsf{SB}})$ is nonnegative. Because its coefficient is nonpositive and $u_t^{\mathsf{SB}} \geq 0$, that term is also nonpositive and hence equals zero.

Now compare $(z_t^{\mathsf{SB}}, u_t^{\mathsf{SB}})$ with $(z, 0)$ for any $z \in \mathcal{Z}$. The u_t-term equals zero at both choices, so saddle-point maximality gives

$$\pi_t\big[\widehat{\theta}_t z_t^{\mathsf{SB}} - \tilde{c}(z_t^{\mathsf{SB}})\big] \geq \pi_t\big[\widehat{\theta}_t z - \tilde{c}(z)\big] \qquad \forall z \in \mathcal{Z}.$$

Since $\pi_t > 0$,

$$z_t^{\mathsf{SB}} \in \arg\max_{z \in \mathcal{Z}} \{\widehat{\theta}_t z - \tilde{c}(z)\}.$$

Translating back through $z = h(a)$ gives

$$a_t^{\mathsf{SB}} \in \arg\max_{a \in \mathcal{A}} \{\widehat{\theta}_t h(a) - c(a)\}.$$

This completes the proof. □

Proof of Corollary 3. We prove each result separately.

Proof of (i).

Discrete convexity of θ in α means each interior point (α_t, θ_t) lies weakly below the chord between its bracketing types, i.e. $\Delta_t^{r\tau} \leq 0$ for all $r < t < \tau$. The virtual-value formula in Theorem 3 adds $\pi_t^{-1} \sum_{r<t<\tau} m_t^{r\tau} \Delta_t^{r\tau}$, which is nonpositive because $\pi_t > 0$, $m_t^{r\tau} \geq 0$, and $\Delta_t^{r\tau} \leq 0$. Hence $\widehat{\theta}_t \leq \theta_t$. Since $\tilde{c}$ is convex and differentiable, the maximizer correspondence of $\vartheta z - \tilde{c}(z)$ is nondecreasing in ϑ, which gives $a_t^{\mathsf{SB}} \leq a_t^{\mathsf{FB}}$.

Proof of (ii).

The constraint (19) gives

$$u_t^{\mathsf{SB}} - \lambda_t^{r\tau} u_r^{\mathsf{SB}} - (1 - \lambda_t^{r\tau}) u_\tau^{\mathsf{SB}} \leq z_t^{\mathsf{SB}} \Delta_t^{r\tau} \leq 0.$$

By definition,

$$\lambda_t^{r\tau} = \frac{\alpha_\tau - \alpha_t}{\alpha_\tau - \alpha_r}, \qquad \alpha_t = \lambda_t^{r\tau} \alpha_r + (1 - \lambda_t^{r\tau}) \alpha_\tau.$$

Thus $\lambda_t^{r\tau}$ is exactly the chord weight that locates α_t between α_r and α_τ. Hence the preceding inequality is precisely the chord inequality

$$u_t^{\mathsf{SB}} \leq \lambda_t^{r\tau} u_r^{\mathsf{SB}} + (1 - \lambda_t^{r\tau}) u_\tau^{\mathsf{SB}}, \qquad r < t < \tau,$$

so the rent profile $(u_t^{\mathsf{SB}})_{t\in\mathcal{T}}$ is discretely convex in α.

If $i<j<k$ with $u_i^{\mathsf{SB}}=u_k^{\mathsf{SB}}=0$, discrete convexity gives

$$u_j^{\mathsf{SB}} \le \lambda_j^{ik} u_i^{\mathsf{SB}} + (1-\lambda_j^{ik}) u_k^{\mathsf{SB}} = 0.$$

Since $u_j^{\mathsf{SB}} \ge 0$, we get $u_j^{\mathsf{SB}}=0$. Thus the zero-rent set is a *contiguous* interval, and positive rents can occur only in the two tails.

Under the strict conditions $\Delta_t^{r\tau} < 0$ and $h(a_t^{\mathsf{SB}}) > 0$ for interior types, three zero-rent types $i<j<k$ would make the constraint for j read

$$0 = u_j^{\mathsf{SB}} - \lambda_j^{ik} u_i^{\mathsf{SB}} - (1-\lambda_j^{ik}) u_k^{\mathsf{SB}} \le z_j^{\mathsf{SB}} \Delta_j^{ik} < 0,$$

a contradiction. Hence at most two IR constraints bind, and by the interval property they are adjacent. □

Proof of Corollary 4. We prove each result separately.

Proof of (i).

When there are only two types, both types are endpoints. By Theorem 3(i), both receive first-best allocations. With no interior type, there are no chord constraints, so the zero-rent profile $u_1=u_2=0$ is implementable; hence $R_{\mathsf{SB}}=R_{\mathsf{FB}}$.

Proof of (ii).

Use the ordering, λ, and χ from the statement. With three types, types 1 and 3 are endpoints, so there is no support constraint targeted at those types. The only allocation-relevant support constraint is the one targeted at type 2. Specifically, the problem (18) reduces to

$$R_{\mathsf{SB}} = \max_{\substack{z_i\in\mathcal{Z},\ u_i\ge 0\\ i=1,2,3}} \ \sum_{i=1}^{3} \pi_i[\theta_i z_i - \tilde{c}(z_i)] - \sum_{i=1}^{3} \pi_i u_i \qquad \text{(SBThree)}$$
$$\text{s.t.} \quad u_2 - \lambda u_1 - (1-\lambda) u_3 \le z_2 \chi.$$

(a) The case $\chi \ge 0$. Since $u_1=u_2=u_3=0$ satisfies the only constraint for every z_2, all rents can be set to zero and each allocation can be chosen to maximize its own first-best surplus. Therefore $R_{\mathsf{SB}}=R_{\mathsf{FB}}$.

(b) The case $\chi<0$. By Theorem 3(i), types 1 and 3 are endpoint types and therefore receive first-best allocations. It remains to characterize the middle type's allocation and the rent profile.

Fix z_2. The rent-minimization problem is

$$\min_{u_1,u_2,u_3\ge 0} \ \pi_1 u_1 + \pi_2 u_2 + \pi_3 u_3 \quad \text{s.t.} \quad \lambda u_1 + (1-\lambda) u_3 - u_2 \ge -z_2\chi,$$

where $-z_2\chi \ge 0$ because $\chi<0$ and $z_2\ge 0$. Since u_2 enters the constraint with coefficient -1, lowering u_2 weakly relaxes the constraint and strictly lowers the objective whenever $u_2>0$. Thus

every rent-minimizing solution has $u_2^{\mathsf{SB}} = 0$, i.e., IR_2 binds. The remaining problem is

$$\min_{u_1,u_3\geq 0} \ \pi_1 u_1 + \pi_3 u_3 \quad \text{s.t.} \quad \lambda u_1 + (1-\lambda)u_3 \geq -z_2\chi.$$

One unit of the left-hand side costs π_1/λ if it is produced through u_1, and $\pi_3/(1-\lambda)$ if it is produced through u_3. Hence the minimum rent cost is

$$-\xi z_2 \chi, \qquad \xi := \min\left\{\frac{\pi_1}{\lambda}, \frac{\pi_3}{1-\lambda}\right\}.$$

It is attained by assigning the required rent to the cheaper endpoint:

$$(u_1^{\mathsf{SB}}, u_3^{\mathsf{SB}}) = \begin{cases} (-z_2\chi/\lambda,\, 0), & \pi_1/\lambda < \pi_3/(1-\lambda), \\ (0,\, -z_2\chi/(1-\lambda)), & \pi_3/(1-\lambda) < \pi_1/\lambda, \\ \text{any split with } \lambda u_1 + (1-\lambda)u_3 = -z_2\chi, & \pi_1/\lambda = \pi_3/(1-\lambda). \end{cases}$$

Substituting the minimized rent cost into (SBThree), the part of the objective involving z_2 becomes

$$\pi_2 \left[\left(\theta_2 + \frac{\xi\chi}{\pi_2}\right) z_2 - \tilde{c}(z_2)\right].$$

Thus the middle type solves

$$a_2^{\mathsf{SB}} \in \arg\max_{a\in\mathcal{A}} \left\{\left(\theta_2 + \frac{\xi\chi}{\pi_2}\right) h(a) - c(a)\right\}.$$

Since $z_2^{\mathsf{SB}} = h(a_2^{\mathsf{SB}})$, the rent formula above is exactly the stated one.

By the analysis above, the single support constraint binds. That is, the supporting line for report 2 also touches the two endpoint points, so $\mathsf{IC}_{1\to 2}$ and $\mathsf{IC}_{3\to 2}$ bind. The positive rent is assigned to the cheaper endpoint, so the other endpoint's IR binds; if the unit costs are tied, any split is optimal. □

C Proofs for Appendix Results

C.1 Proofs for Appendix A.1

Proof of Lemma A-1. Fix item s and a buyer whose success probability at this item is γ.

First consider a reported failure. Under PFS, the reported-failure bill is zero. If the buyer disputes a reported failure, verification can only leave the bill at zero or raise it to the verified success bill, and the buyer also pays the dispute cost. Hence the buyer never disputes after $\widehat{\omega} = 0$.

Now consider a reported success. If the buyer does not dispute, she pays p_s. If she disputes, her expected verified bill is

$$p_s \Pr(\omega = 1 \mid \widehat{\omega} = 1) = p_s \frac{\gamma\sigma_s^1}{\eta_s(\gamma)},$$

and she also pays the dispute cost κ_B. Therefore, when $\eta_s(\gamma) > 0$, she disputes after a reported

success if and only if

$$p_s \frac{\gamma \sigma_s^1}{\eta_s(\gamma)} + \kappa_B < p_s.$$

Equivalently,

$$(1-\gamma) p_s \sigma_s^0 > \kappa_B \eta_s(\gamma).$$

If $\eta_s(\gamma) = 0$, reported success never occurs, and the condition is false.

The buyer's expected cost therefore compares two branches. If reported successes are not disputed, the expected bill is

$$p_s \eta_s(\gamma).$$

If reported successes are disputed, the expected verified bill is

$$\gamma p_s \sigma_s^1,$$

and the expected buyer dispute cost is

$$\kappa_B \eta_s(\gamma).$$

Thus

$$C_s(\gamma) = \min \left\{ p_s \eta_s(\gamma), \gamma p_s \sigma_s^1 + \kappa_B \eta_s(\gamma) \right\}.$$

The seller's expected payoff is $p_s \eta_s(\gamma)$ in the no-dispute branch. In the dispute branch, the seller receives the expected verified bill $\gamma p_s \sigma_s^1$ and pays dispute cost with probability $\eta_s(\gamma)$, so seller expected payoff is $\gamma p_s \sigma_s^1 - \kappa_S \eta_s(\gamma)$. □

Proof of Proposition A-1. By Lemma A-1, at a PFS item (p, σ^1, σ^0) a buyer never disputes a reported failure, and a buyer with success probability $\tilde{\gamma}$ has expected cost $\min\{N(\tilde{\gamma}), D(\tilde{\gamma})\}$ from the item, where

$$N(\tilde{\gamma}) := p[\tilde{\gamma}\sigma^1 + (1-\tilde{\gamma})\sigma^0], \qquad D(\tilde{\gamma}) := \tilde{\gamma}\,(p+\kappa_B)\sigma^1 + \kappa_B(1-\tilde{\gamma})\sigma^0$$

are the no-dispute and dispute branches; she strictly disputes reported successes iff $D(\tilde{\gamma}) < N(\tilde{\gamma})$. The seller's payoff from an assigned type γ is $N(\gamma)$ if she does not dispute and $\gamma(p-\kappa_S)\sigma^1 - \kappa_S(1-\gamma)\sigma^0$ if she does. The key observation is that *the dispute branch of the buyer's cost depends on the item only through σ^0 and the product $(p+\kappa_B)\sigma^1$.*

Costless disputes ($\kappa_B = \kappa_S = 0$). For every type $\tilde{\gamma}$, observe that $D(\tilde{\gamma}) = \tilde{\gamma} p \sigma^1 \le N(\tilde{\gamma})$. If the inequality is strict, the buyer disputes reported successes, in this case, both her expected cost and the seller's payoff equal $\tilde{\gamma} p \sigma^1$; if equality holds, she does not dispute and the same conclusion holds because both sides' dispute costs are zero. Hence (p, σ^1, σ^0) is payoff-equivalent to the fully aligned item $(p\sigma^1, 1, 0)$ for all buyers, and replacing each item accordingly yields an optimal menu with no on-path disputes.

Costly disputes ($\kappa_B + \kappa_S > 0$). We show no optimal menu has a disputing assigned type. By Lemma A-1, a disputing assigned type t, with $\gamma := \gamma_t(a_t)$, must satisfy $(1-\gamma)p\sigma^0 > \kappa_B\eta(\gamma)$. This implies $\gamma < 1$ and $\sigma^0 > 0$. Since $\eta(\gamma) \geq (1-\gamma)\sigma^0$, we further have $(1-\gamma)p\sigma^0 > \kappa_B(1-\gamma)\sigma^0$; dividing by $(1-\gamma)\sigma^0 > 0$ yields $p > \kappa_B$.

We first rule out the corner $\gamma = 0$. If $\kappa_B > 0$, a disputing buyer with $\gamma = 0$ incurs $D(0) = \kappa_B\sigma^0 > 0$; since $\gamma_t(a_t) = 0$ implies $v_t(a_t) = 0$, her IR constraint would require $-\kappa_B\sigma^0 \geq 0$, a contradiction. If $\kappa_B = 0$, then $\kappa_S > 0$. Replace her item (p, σ^1, σ^0) by $(p, \sigma^1, 0)$, keeping the allocation fixed. Before the replacement, every buyer type $\widetilde{\gamma}$ has $D(\widetilde{\gamma}) = \widetilde{\gamma}p\sigma^1 \leq N(\widetilde{\gamma})$, so her optimal expected cost is $\widetilde{\gamma}p\sigma^1$. Afterward, $N'(\widetilde{\gamma}) = D'(\widetilde{\gamma}) = \widetilde{\gamma}p\sigma^1$, so every buyer's expected cost remains unchanged. Since the allocation is fixed, all IC and IR constraints are preserved. The replacement nevertheless strictly benefits the seller when serving the assigned type t. Under the original item, $\gamma = 0$ means that a reported success is necessarily false and occurs with probability $\sigma^0 > 0$. Type t disputes each such report, so the charge p is cancelled and the seller incurs κ_S, yielding the expected payoff $-\kappa_S\sigma^0$. Under the replacement, the false-positive probability is zero, so no charge or dispute occurs and the seller's expected payoff is 0. Thus, without loss, we can eliminate the on-path dispute at $\gamma = 0$ and strictly increase the seller's payoff. Hence every on-path disputer in an optimal menu has $\gamma \in (0,1)$, which is our focus in the remaining proof. Suppose now that assigned type t with $\gamma \in (0,1)$ strictly disputes her item, i.e. $D(\gamma) < N(\gamma)$. Perturb only her item, keeping σ^0 fixed.

Case $\sigma^1 > 0$. Let $I := (p + \kappa_B)\sigma^1 > 0$. Raise the success bill to any $p' > p$ and set $\sigma'^1 := I/(p' + \kappa_B) \in (0, \sigma^1)$, so that $(p' + \kappa_B)\sigma'^1 = I$. Then $D(\widetilde{\gamma})$ is unchanged for every type, while $N(\widetilde{\gamma})$ weakly increases, since $p'\sigma'^1 = Ip'/(p' + \kappa_B)$ and $p'\sigma^0$ are both nondecreasing in p'. Hence every type's optimal expected cost $\min\{N, D\}$ weakly rises, while type t, which remains strictly on the D branch, keeps her cost $D(\gamma)$ exactly. Thus type t's IR and IC constraints are unchanged, and the item becomes weakly less attractive to every other type, so the menu remains feasible. The allocation is unchanged, and the seller's payoff from type t is

$$\gamma(p' - \kappa_S)\sigma'^1 - \kappa_S(1-\gamma)\sigma^0 = \gamma I \frac{p' - \kappa_S}{p' + \kappa_B} - \kappa_S(1-\gamma)\sigma^0.$$

This expression is strictly increasing in p' because $\frac{d}{dp'}\frac{p'-\kappa_S}{p'+\kappa_B} = \frac{\kappa_B+\kappa_S}{(p'+\kappa_B)^2} > 0$ and $\gamma I > 0$. Any $p' > p$ therefore strictly raises the objective, contradicting optimality.

Case $\sigma^1 = 0$. For every buyer type $\widetilde{\gamma}$, the dispute and no-dispute branches are $D(\widetilde{\gamma}) = \kappa_B(1 - \widetilde{\gamma})\sigma^0$ and $N(\widetilde{\gamma}) = p(1-\widetilde{\gamma})\sigma^0$. Since type t strictly disputes and $(1-\gamma)\sigma^0 > 0$, $D(\gamma) < N(\gamma)$ implies $p > \kappa_B$. Replace the old item $(p, 0, \sigma^0)$ by $(p', 0, \sigma^0)$, keeping the allocation fixed, where $p' := \kappa_B$. Under the new item, $N'(\widetilde{\gamma}) = D'(\widetilde{\gamma}) = D(\widetilde{\gamma}) = \kappa_B(1 - \widetilde{\gamma})\sigma^0$, so every buyer's optimal expected cost remains unchanged and all IC and IR constraints remain satisfied. By the seller-favorable tie-break, type t no longer disputes. The seller's expected payoff from type t rises from $-\kappa_S(1-\gamma)\sigma^0$ to

$\kappa_B(1-\gamma)\sigma^0$, a strict increase because $(\kappa_B+\kappa_S)(1-\gamma)\sigma^0 > 0$. Thus the new item preserves feasibility and strictly improves upon the supposedly optimal old item, again contradicting optimality.

In both cases the assigned type cannot dispute her item, so an optimal PFS menu has no on-path disputes. □

Proof of Proposition A-2. Take an optimal PFS menu with no on-path disputes, which exists by Proposition A-1. By Lemma A-1, a buyer never disputes a reported failure, so it suffices to consider reported successes. For a fixed item s, Lemma A-1 gives the dispute condition

$$(1-\gamma)p_s\sigma_s^0 > \kappa_B[\gamma\sigma_s^1 + (1-\gamma)\sigma_s^0].$$

The set of γ's satisfying this strict inequality is a lower interval: rearranged, the condition reads $(1-\gamma)(p_s-\kappa_B)\sigma_s^0 > \kappa_B\gamma\sigma_s^1$. If $p_s \le \kappa_B$ or $\sigma_s^0 = 0$, its left side is nonpositive and its right side nonnegative, so the set is empty; otherwise $(p_s-\kappa_B)\sigma_s^0 > 0$, and collecting the γ-terms shows the condition is equivalent to

$$\gamma < \frac{(p_s-\kappa_B)\sigma_s^0}{(p_s-\kappa_B)\sigma_s^0 + \kappa_B\sigma_s^1}.$$

Since type s does not dispute item s in this menu, $\gamma_s(a_s)$ lies weakly above this threshold whenever the dispute set is nonempty; hence any type with success probability weakly above $\gamma_s(a_s)$ also does not dispute item s. Under the common ordering, $t \ge s$ implies $\gamma_t(a_s) \ge \gamma_s(a_s)$, proving the claim. □

C.2 Proofs for Appendix A.2

Proof of Lemma A-2. We prove the two directions separately.

Necessity. Suppose first that (x_t, y_t) is upper-supported. Then there is an affine line (with slope denoted as d) through the focal point,

$$\ell(x) = y_t + d(x - x_t),$$

such that $\ell(x_j) \ge y_j$ for every j. If another point has the same horizontal coordinate as the focal point, say $x_r = x_t$, then the line has the same height at the two points:

$$\ell(x_r) = \ell(x_t) = y_t.$$

Therefore support of point r requires $y_r \le y_t$. This is exactly the same-success inequality.

Now take one point to the left and one point to the right of the focal point:

$$x_r < x_t < x_\tau.$$

Because $x_t = \lambda_t^{r\tau}x_r + (1-\lambda_t^{r\tau})x_\tau$, linearity of ℓ gives

$$\lambda_t^{r\tau}\ell(x_r) + (1-\lambda_t^{r\tau})\ell(x_\tau) = \ell(x_t) = y_t.$$

Since support means $y_r \le \ell(x_r)$ and $y_\tau \le \ell(x_\tau)$, we obtain

$$\lambda_t^{r\tau} y_r + (1 - \lambda_t^{r\tau}) y_\tau \le y_t.$$

Thus every chord connecting a left point and a right point must pass weakly below the focal point.

Sufficiency. Now suppose the same-success and chord inequalities hold. We need to construct one slope d such that the line

$$\ell(x) = y_t + d(x - x_t)$$

lies weakly above every point.

For a right-side point $x_\tau > x_t$, the support requirement is

$$y_t + d(x_\tau - x_t) \ge y_\tau.$$

Because $x_\tau - x_t > 0$, this is equivalent to the lower bound

$$d \ge \frac{y_\tau - y_t}{x_\tau - x_t}.$$

For a left-side point $x_r < x_t$, the support requirement is

$$y_t + d(x_r - x_t) \ge y_r.$$

Because $x_r - x_t < 0$, dividing by this negative number reverses the inequality, giving the upper bound

$$d \le \frac{y_r - y_t}{x_r - x_t}.$$

Thus right-side points require the slope to be high enough, while left-side points require the slope not to be too high.

Collect the strongest of these lower and upper bounds:

$$L_t := \max_{\tau : x_\tau > x_t} \frac{y_\tau - y_t}{x_\tau - x_t}, \qquad U_t := \min_{r : x_r < x_t} \frac{y_r - y_t}{x_r - x_t},$$

with the usual conventions $L_t = -\infty$ if the first set is empty and $U_t = +\infty$ if the second set is empty. A slope satisfying all side-point constraints exists if and only if $L_t \le U_t$.

The chord inequalities give exactly this compatibility. Indeed, for every left–right pair $x_r < x_t < x_\tau$, the inequality

$$y_t \ge \lambda_t^{r\tau} y_r + (1 - \lambda_t^{r\tau}) y_\tau$$

can be written, using

$$\lambda_t^{r\tau} = \frac{x_\tau - x_t}{x_\tau - x_r}, \qquad 1 - \lambda_t^{r\tau} = \frac{x_t - x_r}{x_\tau - x_r},$$

as

$$y_t \ge \frac{x_\tau - x_t}{x_\tau - x_r} y_r + \frac{x_t - x_r}{x_\tau - x_r} y_\tau.$$

Since $x_\tau - x_r > 0$, multiplying through and rearranging gives

$$(x_\tau - x_t)(y_t - y_r) \geq (x_t - x_r)(y_\tau - y_t).$$

Finally, since $x_t - x_r > 0$ and $x_\tau - x_t > 0$, this is equivalent to

$$\frac{y_\tau - y_t}{x_\tau - x_t} \leq \frac{y_t - y_r}{x_t - x_r} = \frac{y_r - y_t}{x_r - x_t}.$$

Thus the slope from the intended point to any right point is no larger than the slope from any left point to the intended point.

Choose any $d \in [L_t, U_t]$, and define $\ell(x) = y_t + d(x - x_t)$. By construction, ℓ lies above every point with $x_j > x_t$ and every point with $x_j < x_t$. The same-success inequalities give $\ell(x_j) = y_t \geq y_j$ whenever $x_j = x_t$. Therefore ℓ is an upper-supporting line through (x_t, y_t). □

Proof of Corollary A-1. By Lemma 1, the unrestricted contract-design problem (15) can be written in terms of the allocation–rent profile as

$$R_{\mathsf{SB}} = \max_{\substack{a_t \in \mathcal{A},\ u_t \geq 0 \\ t \in \mathcal{T}}} \sum_t \pi_t \left[v_t(a_t) - c(a_t) - u_t \right] \tag{C-13}$$

$$\text{s.t.} \quad \left(\gamma_t(a_t), v_t(a_t) - u_t\right) \text{ is upper-supported in } \left\{ \left(\gamma_j(a_t), v_j(a_t) - u_j\right) : j \in \mathcal{T} \right\}, \quad \forall t.$$

Applying Lemma A-2 to this option-t point set, with $x_j = \gamma_j(a_t)$ and $y_j = v_j(a_t) - u_j$, gives exactly the same-success and chord inequalities in (A-2). The objective is unchanged, so the two formulations are equivalent. □

Proof of Lemma A-3. Fix an allocation profile $(a_t)_{t \in \mathcal{T}}$ with $s(a_t) > 0$ for all t. Relabel types so that $\alpha_1 < \cdots < \alpha_n$. By Corollary A-1, it is enough to find a rent profile $(u_t)_{t \in \mathcal{T}} \in \mathbb{R}^n_+$ such that, for every $r < t < \tau$,

$$\lambda_t^{r\tau} u_r + (1 - \lambda_t^{r\tau}) u_\tau - u_t \geq \lambda_t^{r\tau} v_r(a_t) + (1 - \lambda_t^{r\tau}) v_\tau(a_t) - v_t(a_t),$$

where

$$\lambda_t^{r\tau} = \frac{\alpha_\tau - \alpha_t}{\alpha_\tau - \alpha_r}.$$

For each triple $r < t < \tau$, define

$$B_t^{r\tau} := \lambda_t^{r\tau} v_r(a_t) + (1 - \lambda_t^{r\tau}) v_\tau(a_t) - v_t(a_t).$$

Also define

$$D_t^{r\tau} := \lambda_t^{r\tau} \alpha_r^2 + (1 - \lambda_t^{r\tau}) \alpha_\tau^2 - \alpha_t^2.$$

Since $\lambda_t^{r\tau}$ is the chord weight satisfying

$$\alpha_t = \lambda_t^{r\tau} \alpha_r + (1 - \lambda_t^{r\tau}) \alpha_\tau,$$

we have

$$D_t^{r\tau} = (\alpha_t - \alpha_r)(\alpha_\tau - \alpha_t) > 0.$$

Because there are finitely many triples, choose $M \geq 0$ such that

$$MD_t^{r\tau} \geq B_t^{r\tau}, \qquad \forall r < t < \tau.$$

Now set

$$u_j := M\alpha_j^2, \qquad j \in \mathcal{T}.$$

Then $u_j \geq 0$ for every $j \in \mathcal{T}$, and for every $r < t < \tau$,

$$\lambda_t^{r\tau} u_r + (1 - \lambda_t^{r\tau}) u_\tau - u_t = MD_t^{r\tau} \geq B_t^{r\tau}.$$

Thus all finite support inequalities hold. By Corollary A-1 and Lemma 1, the allocation-rent profile $((a_t, u_t))_{t\in\mathcal{T}}$ is implementable. That is, for any positive-success allocation profile, we can construct a rent profile $(u_t)_{t\in\mathcal{T}}$ such that $((a_t, u_t))_{t\in\mathcal{T}}$ is implementable. □

C.3 Proofs for Appendix A.3

Proof of Theorem A-1. Consider any feasible menu and keep its allocation profile fixed. For item t, let $C_t(\boldsymbol{\gamma})$ be the buyer's minimum expected cost, including buyer's payment and dispute costs, when the true outcome distribution at item a_t is $\boldsymbol{\gamma} \in \Delta(\Omega)$. For any fixed rule specifying whether the buyer pays or disputes after each report, each true outcome ω_i induces a fixed expected cost, obtained by averaging over the possible reported outcomes. The buyer's overall expected cost is therefore $\sum_{i=1}^K \gamma_i$[expected cost conditional on ω_i], which is linear, and hence affine, in $\boldsymbol{\gamma}$. Therefore C_t, the lower envelope over finitely many fixed dispute rules, is concave in $\boldsymbol{\gamma} \in \Delta(\Omega)$.

Because C_t is the minimum of finitely many affine functions, an affine piece attaining that minimum at $\boldsymbol{\gamma}_t(a_t)$—the expected cost of the pay-or-dispute rule the assigned buyer optimally uses—agrees with C_t there and lies weakly above C_t everywhere on $\Delta(\Omega)$. This yields a supporting affine function ℓ_t at the assigned distribution even when $\boldsymbol{\gamma}_t(a_t)$ sits on the boundary of the simplex:

$$\ell_t(\boldsymbol{\gamma}_t(a_t)) = C_t(\boldsymbol{\gamma}_t(a_t)), \qquad \ell_t(\boldsymbol{\gamma}) \geq C_t(\boldsymbol{\gamma}) \quad \forall \boldsymbol{\gamma} \in \Delta(\Omega).$$

For each menu item t, replace its reporting matrix by I_K and its bills by $\widetilde{p}_{t,\omega_k} := \ell_t(e_k)$. Since $\boldsymbol{\gamma} = \sum_{k=1}^K \gamma_{\omega_k} e_k$, affinity gives $\sum_{k=1}^K \gamma_{\omega_k} \widetilde{p}_{t,\omega_k} = \ell_t(\boldsymbol{\gamma})$. Moreover, under full alignment, disputing verifies the reported outcome and therefore leaves the bill unchanged while adding κ_B. Thus, under seller-favorable tie-breaking, no buyer disputes, and a buyer with outcome distribution $\boldsymbol{\gamma}$ incurs cost $\ell_t(\boldsymbol{\gamma})$.

Because the original menu satisfies IR and IC, the replacement preserves both: it leaves each assigned type's cost unchanged, since $\ell_t(\boldsymbol{\gamma}_t(a_t)) = C_t(\boldsymbol{\gamma}_t(a_t))$, while a type r mimicking item t incurs $\ell_t(\boldsymbol{\gamma}_r(a_t)) \geq C_t(\boldsymbol{\gamma}_r(a_t))$. Thus, assigned utilities are unchanged and every deviation becomes

weakly less attractive. The replacement also eliminates dispute waste: the original seller payoff from type t was $C_t(\boldsymbol{\gamma}_t(a_t))$ minus $\kappa_B + \kappa_S$ times the dispute probability, whereas the new payoff is $C_t(\boldsymbol{\gamma}_t(a_t))$. Since allocations and provision costs are unchanged, profit weakly increases. Applying this replacement to every item proves the result. □

Proof of Theorem A-2. By Theorem A-1, it is without loss to restrict attention to full outcome alignment. A direct menu item t is then an allocation a_t and a vector of unrestricted outcome-contingent bills $\boldsymbol{p}_t = (p_{t,\omega_1}, \ldots, p_{t,\omega_K})$. For each type t, define the truthful rent

$$u_t := v_t(a_t) - \boldsymbol{p}_t \cdot \boldsymbol{\gamma}_t(a_t), \tag{C-14}$$

so the expected payment collected from type t is $v_t(a_t) - u_t$. As in (15), the many-outcome second-best problem reads, in terms of the allocation–rent profile,

$$\begin{aligned} \max_{\substack{a_t \in \mathcal{A},\ u_t \geq 0 \\ t \in \mathcal{T}}} \quad & \sum_{t \in \mathcal{T}} \pi_t \big[v_t(a_t) - c(a_t) - u_t \big] \\ \text{s.t.} \quad & \big((a_t, u_t)\big)_{t \in \mathcal{T}} \text{ is implementable,} \end{aligned} \tag{C-15}$$

where the allocation-rent profile $\big((a_t, u_t)\big)_{t \in \mathcal{T}}$ is *implementable* if there exist bills $(\boldsymbol{p}_t)_{t \in \mathcal{T}}$ under which each u_t is the truthful rent in (C-14) and no type prefers another report.

We first rule out no service for extreme types. Suppose $a_t = 0$ and $\boldsymbol{\alpha}_t \notin \operatorname{co}\{\boldsymbol{\alpha}_r : r \neq t\}$. Let $a^* \in \arg\max_{a \in \mathcal{A}} \{v_t(a) - c(a)\}$. The assumed positive surplus implies $a^* \neq 0$. By the separation argument used in part (i) below, replacing $(0, u_t)$ by (a^*, u_t) preserves implementability while leaving all other items' constraints unchanged. This replacement raises the objective by $\pi_t \max_{a \in \mathcal{A}} \{v_t(a) - c(a)\} > 0$, contradicting optimality. Hence an extreme type cannot receive no service at an attained optimum.

Now consider a non-extreme type t with $a_t = 0$. Since $B(0) = \mathbf{0}$, we have $\boldsymbol{\gamma}_r(0) = \boldsymbol{b}(0)$ for every r; moreover, $v_r(0) = 0$ for every r. Thus, incentive compatibility becomes $u_t \leq u_r$ for every r. Hence u_t is a minimum rent. If $u_t > 0$, subtracting u_t from every rent preserves incentive compatibility because all utilities shift equally and preserves individual rationality because all rents remain nonnegative, while strictly increasing profit. Therefore, at an attained optimum, $a_t = 0$ implies $u_t = 0$.

Implementability reduction.

We next characterize implementability for positive-service items. Fix an allocation–rent profile $\big((a_t, u_t)\big)_{t \in \mathcal{T}}$ and a report t with $a_t \neq 0$. Under full alignment, disputing leaves the bill unchanged and adds κ_B, so no buyer disputes and expected transfers equal expected bills. If true type r chooses report t, her expected payment is

$$\boldsymbol{p}_t^\top \boldsymbol{\gamma}_r(a_t) = \boldsymbol{p}_t^\top \boldsymbol{b}(a_t) + \big(B(a_t)^\top \boldsymbol{p}_t\big)^\top \boldsymbol{\alpha}_r. \tag{C-16}$$

Thus, the bills generate an affine expected-payment function of $\boldsymbol{\alpha}_r$. Conversely, because $B(a_t)$

has full column rank, the map $\boldsymbol{p}_t \mapsto B(a_t)^\top \boldsymbol{p}_t$ is onto $\mathbb{R}^d$. Hence, for any target affine function $\ell_t(\boldsymbol{\alpha}) = q_t + \boldsymbol{\beta}_t \cdot \boldsymbol{\alpha}$, we can choose $\boldsymbol{p}_t^0$ satisfying $B(a_t)^\top \boldsymbol{p}_t^0 = \boldsymbol{\beta}_t$, and then add the common amount $q_t - (\boldsymbol{p}_t^0)^\top \boldsymbol{b}(a_t)$ to every bill. Since $\sum_{k=1}^K \gamma_{r,\omega_k}(a_t) = 1$, this common shift sets the intercept, so the resulting bills charge every type r exactly $\ell_t(\boldsymbol{\alpha}_r)$.

We now translate incentive compatibility into conditions on ℓ_t. To give the assigned type t rent u_t, her expected payment must satisfy $\ell_t(\boldsymbol{\alpha}_t) = v_t(a_t) - u_t$. To prevent type r from choosing report t, we require $v_r(a_t) - \ell_t(\boldsymbol{\alpha}_r) \le u_r$, or equivalently, $\ell_t(\boldsymbol{\alpha}_r) \ge v_r(a_t) - u_r$. For every positive-service report t, full column rank allows its bill vector $\boldsymbol{p}_t$ to implement any affine expected-payment function. Because each report has its own bill vector, this construction can be carried out independently across reports. A no-service report can be implemented with zero bills and zero rent, as established before. Therefore, the allocation–rent profile is implementable if and only if, for every t with $a_t \neq 0$, there exists an affine function ℓ_t such that

$$\ell_t(\boldsymbol{\alpha}_t) = v_t(a_t) - u_t, \qquad \ell_t(\boldsymbol{\alpha}_r) \ge v_r(a_t) - u_r \quad \forall r \in \mathcal{T}. \tag{C-17}$$

Proof of (i). Let $\left((a_t, u_t)\right)_{t\in\mathcal{T}}$ be an attained optimum and fix a type t satisfying $\boldsymbol{\alpha}_t \notin \mathrm{co}\{\boldsymbol{\alpha}_r : r \neq t\}$. By the no-service argument above, $a_t \neq 0$. We first show that a_t can be replaced by any $a' \neq 0$ while holding all rents and all other allocations fixed. Because $\mathcal{T}$ is finite, the convex hull $\mathrm{co}\{\boldsymbol{\alpha}_r : r \neq t\}$ is compact and convex. Since $\boldsymbol{\alpha}_t$ lies outside this set, the strong separating hyperplane theorem provides a vector $\boldsymbol{\eta}$ and a scalar c such that

$$\boldsymbol{\eta} \cdot \boldsymbol{\alpha}_t > c \ge \boldsymbol{\eta} \cdot \boldsymbol{\alpha}_r \qquad \forall r \neq t.$$

Thus, $\boldsymbol{\eta}$ strictly separates type t's index from every other type's index.

Fix any $a' \neq 0$. By (C-17), it suffices to construct an affine expected-payment function that gives type t rent u_t and deters every other type from choosing report t. Define

$$\ell_t'(\boldsymbol{\alpha}) := v_t(a') - u_t - M\,\boldsymbol{\eta} \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}_t).$$

At $\boldsymbol{\alpha}_t$, the last term is zero, so $\ell_t'(\boldsymbol{\alpha}_t) = v_t(a') - u_t$. For every $r \neq t$,

$$\ell_t'(\boldsymbol{\alpha}_r) = v_t(a') - u_t + M(\boldsymbol{\eta} \cdot \boldsymbol{\alpha}_t - \boldsymbol{\eta} \cdot \boldsymbol{\alpha}_r).$$

The expression in parentheses is strictly positive. Because there are finitely many other types, one sufficiently large M ensures $\ell_t'(\boldsymbol{\alpha}_r) \ge v_r(a') - u_r$ for every $r \neq t$. Therefore, report t remains implementable after replacing a_t by a'. The conditions for every report $s \neq t$ depend only on a_s and the rent profile, which are unchanged. Hence the modified allocation–rent profile remains feasible. Choosing $a_t' = a^* \in \arg\max_{a\in\mathcal{A}}\{v_t(a) - c(a)\}$ for extreme type t completes the proof.

Proof of (ii). We first handle no-service reports. By the no-service argument above, $z_t = 0$ implies

$u_t = 0$. Hence, for every $\lambda \in \mathcal{E}_t$,

$$u_t - \sum_{r \in \mathcal{T}} \lambda_r u_r = -\sum_{r \in \mathcal{T}} \lambda_r u_r \leq 0 = z_t \Delta_t^\lambda,$$

because all rents are nonnegative. Thus, the constraints hold automatically for no-service reports. We may therefore focus on a report t with $a_t \neq 0$.

Fix such a report and write $y_r := v_r(a_t) - u_r$. By (C-17), report t is implementable if and only if there exists an affine function through $(\boldsymbol{\alpha}_t, y_t)$ that lies weakly above $(\boldsymbol{\alpha}_r, y_r)$ for every r. Writing this function as

$$\ell_t(\boldsymbol{\alpha}) = y_t + \boldsymbol{\beta} \cdot (\boldsymbol{\alpha} - \boldsymbol{\alpha}_t),$$

this is equivalent to the existence of $\boldsymbol{\beta} \in \mathbb{R}^d$ satisfying

$$\boldsymbol{\beta} \cdot (\boldsymbol{\alpha}_r - \boldsymbol{\alpha}_t) \geq y_r - y_t \qquad \forall r \neq t.$$

We claim that this condition is equivalent to

$$y_t \geq \sum_{r \in \mathcal{T}} \lambda_r y_r \qquad \forall \lambda \in \Lambda_t. \tag{C-18}$$

For necessity, let ℓ_t be such an affine function. For any $\lambda \in \Lambda_t$, affinity and $\sum_r \lambda_r \boldsymbol{\alpha}_r = \boldsymbol{\alpha}_t$ imply

$$y_t = \ell_t(\boldsymbol{\alpha}_t) = \sum_{r \in \mathcal{T}} \lambda_r \ell_t(\boldsymbol{\alpha}_r) \geq \sum_{r \in \mathcal{T}} \lambda_r y_r.$$

For sufficiency, suppose no such $\boldsymbol{\beta}$ exists. Farkas' lemma then provides weights $\mu_r \geq 0$, $r \neq t$, not all zero, such that

$$\sum_{r \neq t} \mu_r (\boldsymbol{\alpha}_r - \boldsymbol{\alpha}_t) = \mathbf{0}, \qquad \sum_{r \neq t} \mu_r (y_r - y_t) > 0.$$

Let $M := \sum_{r \neq t} \mu_r > 0$, and define

$$\lambda_r := \frac{\mu_r}{M} \quad \text{for } r \neq t, \qquad \lambda_t := 0.$$

These weights are nonnegative and sum to one. Moreover, the first Farkas condition implies

$$\sum_{r \in \mathcal{T}} \lambda_r \boldsymbol{\alpha}_r - \boldsymbol{\alpha}_t = \frac{1}{M} \sum_{r \neq t} \mu_r (\boldsymbol{\alpha}_r - \boldsymbol{\alpha}_t) = \mathbf{0},$$

so $\lambda \in \Lambda_t$. The second Farkas condition implies

$$\sum_{r \in \mathcal{T}} \lambda_r y_r - y_t = \frac{1}{M} \sum_{r \neq t} \mu_r (y_r - y_t) > 0,$$

contradicting (C-18). Thus, (C-17) and (C-18) are equivalent. Equivalently, for each t with $a_t \neq 0$, the tuple $(a_t, (u_r)_{r \in \mathcal{T}})$ is implementable if and only if

$$v_t(a_t) - u_t \geq \sum_{r \in \mathcal{T}} \lambda_r \big[v_r(a_t) - u_r \big] \qquad \forall \lambda \in \Lambda_t. \tag{C-19}$$

If $\boldsymbol{\alpha}_t \notin \mathrm{co}\{\boldsymbol{\alpha}_r : r \neq t\}$, then $\Lambda_t = \{e_t\}$, and the condition reduces to the identity

$$v_t(a_t) - u_t \geq v_t(a_t) - u_t.$$

It is therefore automatically satisfied for an extreme type.

Now substitute $v_r(a_t) = \theta_r z_t$, where $z_t = h(a_t)$. For every $\lambda \in \mathcal{E}_t$, (C-19) becomes

$$\theta_t z_t - u_t \geq \sum_{r \in \mathcal{T}} \lambda_r (\theta_r z_t - u_r),$$

or, equivalently,

$$u_t - \sum_{r \in \mathcal{T}} \lambda_r u_r \leq z_t \Delta_t^\lambda. \tag{C-20}$$

These are exactly the constraints (A-7). Together with the domain restriction implied by the no-service condition $z_t = 0 \Longrightarrow u_t = 0$,

$$(z_t, u_t) \in \{(0,0)\} \cup \big((\mathcal{Z} \setminus \{0\}) \times \mathbb{R}_+\big),$$

the objective becomes

$$\sum_{t \in \mathcal{T}} \pi_t \big[\theta_t z_t - \tilde{c}(z_t) - u_t\big].$$

When $\mathcal{Z}$ and $\tilde{c}$ are convex, the objective is concave and the constraints (C-20) are affine. Note that the maximization domain is also convex: because $z \geq 0$, a convex combination can have $z_t = 0$ only if both original points have $z_t = 0$, in which case both also have $u_t = 0$. This proves the convex formulation.

Finally, suppose the program admits a primal–dual saddle point with multipliers $m_t^\lambda \geq 0$ on the constraints (C-20). Dualizing these constraints, the Lagrangian is

$$\mathcal{L} = \sum_{t \in \mathcal{T}} \pi_t \big[\theta_t z_t - \tilde{c}(z_t) - u_t\big] - \sum_{t \in \mathcal{T}} \sum_{\lambda \in \mathcal{E}_t} m_t^\lambda \left(u_t - \sum_{r \in \mathcal{T}} \lambda_r u_r - z_t \Delta_t^\lambda \right).$$

For each type t, collecting the terms involving z_t, the z_t-dependent part of the Lagrangian is

$$\pi_t \big[\widehat{\theta}_t z_t - \tilde{c}(z_t)\big],$$

where

$$\widehat{\theta}_t := \theta_t + \frac{1}{\pi_t} \sum_{\lambda \in \mathcal{E}_t} m_t^\lambda \Delta_t^\lambda.$$

The rest follows similarly to the proof of Theorem 3(ii). □