Pricing Consulting Services in the Age of Agentic AI: A Strategy Map for Executives*

Erik Brynjolfsson† and Georgios Petropoulos‡

August 23, 2026

Abstract

Artificial intelligence (AI) agents are transforming how consulting work is produced, delivered, and valued, yet pricing models remain rooted in assumptions from a human-only era. This article offers consulting leaders a practical framework, grounded in principal–agent theory, for understanding why AI disrupts traditional fee structures and for redesigning pricing to reflect the new information environment. When AI is used internally to augment consultants' productivity, it is likely to initially heighten information asymmetries even as efficiency improves. Over time, however, AI agents can enhance outcome measurability as systems become embedded in client operations. These shifts reshape the economic conditions under which different pricing models (i.e., time-based, project-based, hybrid, subscription, and outcome-based) perform effectively. The article introduces a pricing strategy map that positions each model along two critical dimensions, outcome observability and input cost observability, and distills the analysis into decision rules that executives can apply directly. The framework implies that AI tends to push consulting firms toward hybrid and asset-based structures. Drawing on examples from leading firms, the analysis offers guidance for leaders seeking to align incentives, capture value, and build sustainable pricing strategies in an AI-enabled consulting landscape.

Keywords: principal–agent theory; consulting services pricing; artificial intelligence; information asymmetry; mechanism design.

*We are highly indebted to Tomas Castagnino, Lan Guan, Wang Jin, Ignacio Mamone and Jiaxin Pei for their insights. In addition, the informal expert discussions described in Section 2 were held with three senior executives of Accenture, a firm whose publicly disclosed practices are also among those analyzed in this paper. These discussions were confidential, were not recorded, and served solely as motivating background. No data were collected or analyzed, and the discussions did not involve the collection of data on human subjects. Accenture exercised no control over how the perspectives shared in these discussions were used, interpreted, or presented. Separately, the Stanford Digital Economy Lab has received a monetary gift from Accenture. This gift was general support for the Lab and was not provided for, or in connection with, this study. Accenture did not commission or direct this research, imposed no restrictions on publication, and had no right to review or approve the paper prior to its circulation. All views expressed, and any errors, are solely the responsibility of the authors.

†Brynjolfsson: Stanford Digital Economy Lab and NBER, erikb@stanford.edu.

‡Petropoulos: University of Southern California, georgios@marshall.usc.edu

1 The pricing problem in the age of AI agents

For decades, the consulting industry has priced its work in ways that reflect a deceptively simple production logic. A team of human experts sells expertise, judgment, and effort, and contracts compensating them rest on observable proxies for that effort: hours billed, projects delivered, retainers paid for continued availability. The dominant pricing models, time-and-materials at one end and project-based fixed fees at the other, evolved through decades of trial and error to manage two persistent frictions in professional services. Clients cannot fully observe what consultants do, and clients cannot fully verify what consultants' work has produced. Pricing models are, fundamentally, devices for managing these informational gaps.

The rapid emergence of artificial intelligence (AI) and AI agents has begun to dissolve the assumptions on which these pricing models rest. AI systems now automate analysis, generate insights, standardize workflows, and even interact directly with clients. The boundaries between advisory service, software product, and managed operation are increasingly blurred. Consulting firms that once sold consultant-hours now sell access to internal AI platforms, productized AI agents embedded in client environments, and hybrid arrangements that combine elements of both. The classical price equation, number of consultants times billable hours times rate per hour, describes less and less of what consulting firms actually do and what they actually charge for.

This article is motivated by a consequential question: how should consulting firms price their work when AI fundamentally alters the information environment in which contracts operate? Our central argument is that pricing in the AI era is best understood through the lens of principal–agent contract theory, and that the apparent diversity of pricing models emerges as comparative-statics results from a single underlying contracting problem.[1] The framework we develop allows us to predict, rather than merely catalogue, the directions in which AI pushes consulting pricing.

The framework rests on two informational dimensions. The first is *outcome observability*: how precisely the client can verify the results of the consultant's work. When outcomes are measurable, attributable, and verifiable, pricing can be linked to performance. Because the consultant's effort is itself unobservable, performance-linked pricing is what aligns incentives: this dimension carries a hidden-action (moral-hazard) friction, in that the client uses the outcome signal to motivate effort it cannot see directly. When outcomes are noisy or hard to attribute, performance-linked contracts impose risk on the consultant without an informational return. The second dimension is *input cost observability*: how precisely the

[1]The classical references on the principal–agent problem are Grossman and Hart (1983), Holmström (1979), and Laffont and Tirole (1993).

client can observe the true cost of the consultant's inputs, or equivalently, the complement of how much the consultant privately knows about the engagement's cost structure that the client does not. Low observability on this axis is the footprint of the consultant's private cost knowledge, which is why it carries a hidden-information (adverse selection) friction. When input cost observability is high, contracts can compensate inputs (hours, deliverables, infrastructure spend) without inviting cost padding. When it is low, input-based contracts decouple price from value, and the consultant's private cost knowledge can be exploited rather than compensated.

These two dimensions are not arbitrary axes drawn through the practice. Any pricing model must be evaluated against what it implies about how outcomes will be measured and how costs will be verified, and any change in technology that affects what the client can observe will affect which pricing model is optimal. AI changes both dimensions, but it changes them differently depending on whether AI is deployed internally to augment consultants' capabilities and work, or externally as an autonomous agent operating in the client's environment.

We argue that the two AI transitions act primarily on different informational dimensions, with distinctly different implications for optimal pricing. *AI augmentation*, where generative and analytical models accelerate consultants' work without becoming visible to clients, lowers input cost observability: identical deliverables can now be produced with widely varying combinations of human time, computational spend, and unobservable orchestration effort. *AI productization*, where AI agents are deployed in client environments and generate continuous telemetry of performance, raises outcome observability by compressing outcome noise: accuracy, uptime, conversion lifts, and decision histories become directly observable.

A central result of the article qualifies this second transition. Outcome noise has two economically distinct sources: how precisely a result can be *measured*, and how cleanly it can be *attributed* to the consultant's work rather than to market conditions, concurrent initiatives, or the client's own execution. Telemetry compresses the first and leaves the second untouched. Attribution risk therefore sets a floor on outcome-signal informativeness that no amount of instrumentation can lower, and that floor caps how far productization can push a contract toward outcome-based pricing. The practical upshot is one of the article's central messages for executives: wherever the consultant's contribution is confounded with everything else the client is doing, hybrid contracts are not a transitional stop on the way to outcome-based pricing but the destination, and the highest returns to measurement investment lie in attribution technology, not in more telemetry.

The two dimensions define a *strategy map*: a plane whose axes are outcome observability and input cost observability, on which each engagement occupies a position and each canonical pricing regime in consulting a region. In what follows, we develop a principal–agent framework whose central results explain why the canonical pricing regimes occupy the regions of the strategy map they do. We then use the framework to derive the directions in which AI augmentation and AI productization push pricing. Finally, we distill the framework into concrete decision rules that consulting executives can use to locate an engagement on the strategy map and choose its pricing structure. The article is practice-oriented by design: its contribution is the map and the decision rules, and the body develops the argument in plain language, while the formal model that disciplines it, ensuring that the map's regions and the transitions' directions are derived rather than asserted, is confined to Appendix A.

The article proceeds as follows. Section 2 describes the research approach. Section 3 sets out the contracting problem and the two informational frictions in plain language, stating the central results of the formal analysis. Section 4 derives the pricing strategy map from these results. Section 5 examines the two AI transitions through the framework's lens, drawing on the evolution of AI deployments at leading consulting firms. Section 6 examines two extended cases, Palantir's contractual transition and Salesforce Agentforce's hybrid arrangement, that we use to illustrate how the framework interprets observed pricing practice. Section 7 distills the framework into a set of decision rules for locating an engagement and choosing its pricing structure. Section 8 concludes and identifies a research agenda. Appendix A presents the formal model and its derivations.

2 Research approach

The argument developed in this article rests on three elements: motivating background from expert discussions, desk research on publicly available evidence, and a formal analytical framework. First, during the preparation of this work we held informal, confidential discussions with three senior executives at Accenture. These conversations were not research interviews: no interview protocol was administered, nothing was recorded, and no data were collected or analyzed. Hand-written and typed notes were kept, but they served purely as an aide-mémoire. The executives were approached through existing professional relationships because they had direct, firm-level visibility into how AI is introduced, deployed, and commercialized. At the time of the discussions, all three held senior leadership roles spanning AI strategy formulation, service-offering design, and commercial negotiation. The conversations ranged across the operational, technological, and contractual dimensions of AI adoption, touching on where AI and AI agents are deployed internally versus in client environments, how those deployment choices change what clients can observe about effort and outcomes, how pricing and contract structures are set and renegotiated, and how the firm captures value from AI-enabled work.

Their role in this article is strictly motivational: they helped inspire the two informational dimensions on which the framework rests, outcome observability and input cost observability, and the observation that AI agents operate along two distinct paths, internal augmentation of consultant workflows and external productization in client environments. No statement in the analysis that follows rests on these discussions as evidence. The framework is derived from the formal model of Appendix A and evaluated against the publicly documented practices described below. To preserve confidentiality we do not identify the individual participants, and we make no attributions to them in the analysis.

Second, we conducted desk research on publicly available evidence regarding AI and AI-agent deployment by major consulting firms and AI-enabled service providers. Sources included corporate press releases and product announcements, financial filings and investor communications, trade-press coverage, analyst reports, and firm-issued white papers and methodology documents; where possible we relied on primary disclosures, such as securities filings and official pricing pages, rather than on secondary reporting. Candidate firms were assembled from the largest strategy and technology-consulting firms together with prominent AI-native service providers and from these we selected firms for closer analysis on two criteria fixed in advance: substantive involvement in either the internal-augmentation or the external-productization transition that the framework addresses, and sufficient public disclosure of pricing structure to permit economic interpretation. The firms referenced here, McKinsey, Deloitte, Accenture, Palantir, and Salesforce, meet both criteria and together span the range of AI-enabled service models currently visible in the market.

Third, we developed an analytical economic framework based on a multi-signal principal–agent model that derives optimal pricing schemes under asymmetric information for different information structures, and that allows directional implications to be drawn on how AI and AI agents affect consulting pricing. The framework is presented formally in Appendix A.

The three elements relate iteratively: the expert discussions and desk research generated the observations that motivated the analytical framework, and the framework in turn generated directional predictions that we compared against the publicly observed practices of consulting firms and AI service providers. The consistency between the framework's predictions and observed pricing structures, including transitions in pricing models over time, offers qualitative plausibility for the framework's central claims rather than validation in a confirmatory-empirical sense.

3 The contracting problem and its two informational frictions

Consulting is, in its essential structure, a principal–agent relationship. The client (the principal) hires the consultant (the agent) to do work that the client either cannot do herself or chooses not to do herself. The consultant brings expertise, judgment, and effort that the client values. The client brings the engagement, the data, and the willingness to pay. The pricing contract is the device that converts the consultant's work into the consultant's compensation.

The fundamental challenge of designing this contract is informational. The client cannot, in general, fully observe what the consultant does. She sees deliverables, presentations, recommendations, and reports, but not the effort, judgment, and analytical work that produced them. She cannot directly verify how hard the consultant worked, how much skill was applied, or whether the methods used were appropriate. After the engagement, she may or may not be able to verify what the work produced, whether revenue improved, whether costs fell, whether a strategic decision turned out well. Pricing must accomplish two things simultaneously: it must compensate the consultant fairly for work the client cannot fully see, and it must give the consultant incentives to do work that benefits the client.

Two informational frictions shape what pricing can and cannot achieve. The first concerns outcomes: how precisely can the client measure and attribute the results of the engagement? Some consulting work produces outcomes that are quickly visible, quantitatively measurable, and clearly attributable: a marketing optimization that lifts conversion rates by a documented percentage, a process improvement that reduces operating costs by a documented amount, a predictive system whose accuracy can be benchmarked against ground truth. Other consulting work produces outcomes that are diffuse, slow to manifest, or hard to disentangle from other factors: a strategy recommendation whose effects play out over years, a leadership development engagement whose outputs depend heavily on the client's own execution, an advisory relationship whose value lies in counsel given at moments of uncertainty. We term this dimension *outcome observability*. When outcome observability is high, the client can directly link payment to results; when it is low, she cannot.

The second friction concerns costs and effort: how much does the consultant know about the engagement's true cost structure that the client does not? Some consulting engagements have cost structures that are highly transparent, a fixed team of senior associates at documented billing rates, working a recorded number of hours on clearly delineated tasks, supported by overheads the client can audit or benchmark. Specialized expert advisory engagements, regulated procurement contracts with government clients, and traditional human-intensive audits all fit this pattern: the relationship between consultant inputs and the consultant's true cost of effort is essentially linear, observable, and verifiable. Other consulting engagements have cost structures the client cannot see through. Engagements that rely on proprietary methodologies whose development costs are sunk and unobservable, on reusable intellectual property amortized across many clients, on platform infrastructure whose marginal cost is decoupled from the engagement, or on bundled inputs whose individual contributions cannot be disentangled. In these engagements, identical deliverables can be produced at widely varying combinations of inputs, and the consultant knows what these inputs actually were while the client sees only the deliverable. We term this dimension *input cost observability*, low when the consultant's cost structure is hidden in this way. When input cost observability is high, the client can confidently link payment to inputs. When it is low, input-based pricing decouples price from value-creating effort, and the consultant's private cost knowledge yields an information rent that the client cannot contract away.

These two dimensions are not the only frictions in consulting contracts: reputation, repeated interaction, relational dynamics, and the consultant's outside options all matter. But they are the two that determine which pricing model is optimal in a given engagement, and they correspond to the two canonical informational frictions of contract theory: uncertainty about what the work produced, a hidden-action friction on the outcome side, and the consultant's private knowledge of the engagement's cost structure, a hidden-information friction on the cost side. Other considerations affect the level of compensation, but not its structure. The pricing model is, in our framework, the answer to one question: given what each party can observe about outcomes and costs, what is the most informationally efficient way to convert the consultant's work into compensation?

The contracting problem we have just described is a special case of the multi-signal principal–agent model that has been developed in the contract theory literature since the late 1970s. Two classical results from that literature are important for the remainder of the article.

The first result is the *informativeness principle*. The optimal contract weights observable signals in proportion to how informative they are about the consultant's underlying effort. A signal that is highly informative, one that closely tracks effort with little noise, should receive substantial weight in the pricing formula. A signal that is noisy or only loosely connected to effort should receive little weight. Applied to our two informational dimensions, the principle implies that the relative weight of outcome-linked payment to input-linked payment is inversely proportional to the relative noise of each signal: the cleaner the outcome signal, the more pricing should rely on outcomes. The cleaner the cost signal, the more pricing should rely on inputs. In the limit, when outcomes are vastly more informative than inputs, pricing should be purely outcome-based. When inputs are vastly more informative than outcomes, pricing should be purely input-based. Throughout, we take a signal's *noise* to mean its variance (the square of its standard deviation), not the standard deviation itself; with that convention the relative-weight result has a particularly clean expression:

$$\frac{\text{weight on outcome}}{\text{weight on inputs}} = \frac{\text{noise in cost signal}}{\text{noise in outcome signal}}. \tag{1}$$

The formula captures a fundamental contracting intuition: each signal earns its place in the pricing formula by how informative it is, not by how visible it is. Inputs are usually more visible than outcomes (hours are easy to count, value is hard to measure), but visibility is not the criterion. Informativeness is.[2]

The second result is the *intensity result*. The total power of incentives, the total degree to which the consultant's compensation varies with observable signals rather than being fixed, falls when either signal becomes noisier. Imposing variable compensation on the consultant is costly: the consultant bears risk that the client does not, and must be compensated for bearing that risk through a higher expected payment. This risk premium is only worth paying when the variable component is informationally useful. When both outcomes and costs are noisy, the variable component does little to align incentives but still imposes risk. The principal prefers a low-powered contract (typically a fixed retainer) to a high-powered contract that compensates risk without buying information.[3]

The informativeness principle and the intensity result imply that the optimal contract takes qualitatively different forms in different regions of the (outcome observability, input cost observability) plane. Section 4 walks through each region in turn and explains why the framework predicts the specific pricing structure observed there.

More importantly, these two results tell us what happens to the optimal contract when the parameters themselves change, when a technology, an institutional shift, or an industry trend systematically moves either outcome observability or input cost observability. The two AI transitions are, in our framework, exactly such parameter shifts. Section 5 examines how each transition affects the informational dimensions and what the framework predicts about the direction in which optimal pricing should move in response.

The choice of pricing contract is not a new question for information systems research. It is the central concern of a substantial empirical literature on IT outsourcing and software

[2]The single-signal version of this principle is due to Holmström (1979). The multi-signal extension we rely on here was developed by Banker and Datar (1989).

[3]The intensity result is the standard implication of CARA–normal moral hazard models with risk-averse agents. See Holmström and Milgrom (1987) and the textbook treatment in Bolton and Dewatripont (2005), ch. 4.

development, in which the dominant object of study is the choice between fixed-price (FP) and time-and-materials (T&M) contracts. Gopal et al. (2003) show that requirement uncertainty, team size, and resource availability drive this choice and that the choice in turn shapes vendor profit. Gopal and Koka (2010) argue that the fixed-price structure, by making the vendor the residual claimant on its own effort, induces greater effort and higher delivered quality. This work operationalizes, on project-level data, trade-offs closely related to those our model formalizes. Its procurement-theoretic foundation (Bajari and Tadelis, 2001) emphasizes a distinct but complementary mechanism: fixed-price contracts sharpen cost-reduction incentives while flexible, cost-based contracts ease ex-post adaptation when projects are complex, so that contract choice trades incentives against transaction costs rather than against risk insurance. Our framework's risk-insurance channel and this adaptation channel push the fixed-price-versus-time-and-materials choice in the same direction as complexity and measurement difficulty rise. In the language of our framework, the FP-vs-T&M choice is the bottom edge of the map: movement along the input-cost-observability axis, from input-based pricing toward a fixed fee, while outcome observability stays low.

Our paper is also relevant to contributions on the transformation of knowledge work: evidence that generative AI raises the productivity of professional and analytical labor, unevenly across skill levels, and that whether AI augments or automates human work governs who captures the value it creates (Brynjolfsson, 2022; Brynjolfsson, Li, and Raymond, 2025; Ide and Talamàs, 2025). That literature has concentrated on productivity and skill. We take up the question it leaves open, namely how the *pricing* of knowledge work, and thus the division of the value AI creates, changes with it.

Two threads are especially close to our mechanism. The first is analytical: Dey, Fan, and Zhang (2010) model software-outsourcing contract design and find that fixed-price contracts suit simple, short engagements while time-and-materials suit complex ones, and Wu, Ding, and Hitt (2013) investigate, analytically and experimentally, how IT value, learning, and contract structure interact. The second, and the most direct, is that this literature has already named the information rent at the center of our distribution result: Gopal and Sivaramakrishnan (2008) argue that vendors prefer fixed-price contracts precisely because private knowledge of their own capabilities lets them capture larger information rents. Fitoussi and Gurbaxani (2012) show empirically that the measurability of a contracted objective governs the incentive content of the contract, the outcome-observability axis of our map. Susarla, Subramanyam, and Karhade (2010) and Chang, Gurbaxani, and Ravindran (2017) examine how contractual provisions and asset transfer manage the hold-up and intellectual-property hazards on which a consultant's cost advantage rests, while Huang et al. (2021) show how monitoring and renegotiation substitute for or complement one another in shaping the realized contract. Our contribution is to generalize and to formalize. Where this literature studies a one-dimensional choice between two named contracts, largely in software outsourcing, we embed the choice in a two-dimensional information space, outcome observability against input cost observability, in which the same principal–agent primitives generate five pricing forms as regions rather than two as a binary, and we add the two AI transitions as comparative-statics displacements within that space.

A second and adjacent literature studies not which contract is chosen but how digital offerings are *priced*. Work on the pricing of information goods and software explains why such offerings are sold through subscriptions, usage fees, free tiers, and combinations of these, and how such schedules screen heterogeneous buyers (Sundararajan, 2004; Niculescu and Wu, 2014). A strand within it treats the software-as-a-service model directly, including the agency and task-allocation problems that arise when software is delivered as an ongoing service rather than a licensed product (Susarla, Barua, and Whinston, 2009, 2010). The hybrid schedules we examine in Section 6, in which consulting output is repackaged as subscription, usage, and outcome-linked components, are instances of exactly these forms. Our contribution is to say *when* each becomes optimal for a professional-services engagement, as a function of the two informational coordinates rather than of the properties of software markets as such.

4 The pricing strategy map

The five canonical pricing regimes, time-based, project-based, access-based, outcome-based, and hybrid, each correspond to a region of the two-dimensional space defined by outcome observability and input cost observability. Figure 1 presents this strategy map. The regions of the map reflect the comparative statics of the optimal contract (see Appendix A for a formal exposition).

It is worth being explicit about what *optimal* means here, and why the resulting contract is one both parties can live with. Because the client cannot observe the consultant's effort, she cannot pay for it directly. She can condition payment only on what she does observe, the realized outcome and the reported cost or inputs. The optimal contract is the schedule she commits to in advance, which maximizes the surplus the engagement generates net of the cost of compensating a consultant for whatever risk the schedule imposes.

We walk through each region in turn, explaining why the framework predicts the corresponding pricing structure.

Top-left: Time-based pricing. In this region, outcomes are difficult to measure and attribute, but input cost observability is high. The client cannot link payment to results,

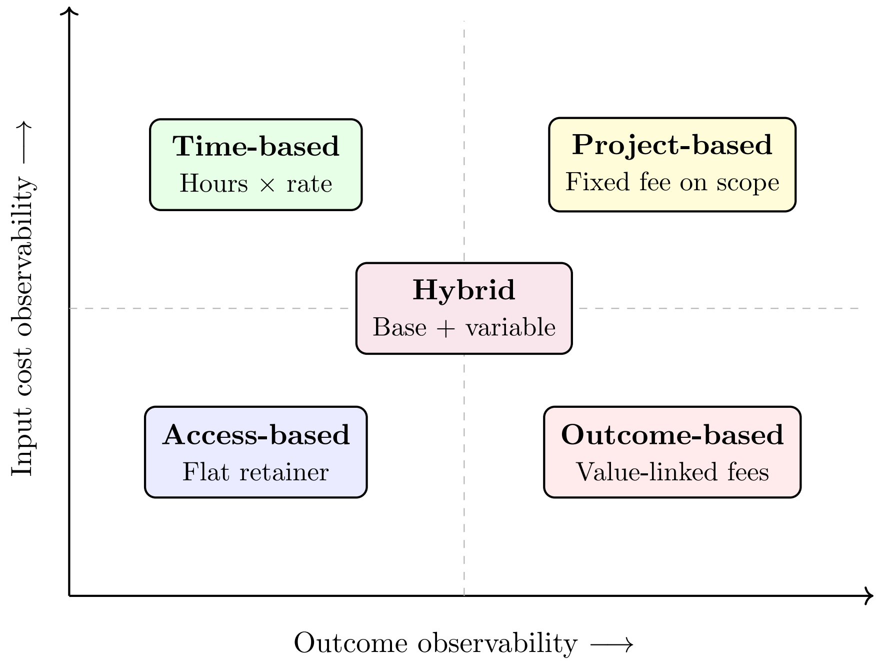

Figure 1: The pricing strategy map. Each region of the parameter plane corresponds to the optimal pricing structure derived from the framework of Appendix A. The horizontal axis indicates increasing outcome observability; the vertical axis indicates increasing input cost observability. Hybrid pricing occupies the interior of the map.

but she can verify hours worked and rates charged. The informativeness principle implies that inputs are the only informative signal of consultant effort, and the optimal contract loads its variable component onto inputs alone. The result is time-and-materials billing: a transparent rate per hour, multiplied by hours worked, with no performance-linked component. This model has historically dominated traditional consulting because it accurately describes the information environment of human-only engagements: outcomes are slow and diffuse, but consultant time is countable and verifiable. The classical pricing formula, number of consultants times hours times rate, is the framework's solution to this informational corner.

Top-right: Project-based pricing. When both outcomes (or proxies for outcomes, such as discrete deliverables) and costs are observable, the optimal contract approaches efficiency from both directions: any reasonable implementation comes close to a first-best contract. In practice, the natural implementation is a fixed fee paid on completion of a defined scope. The fixed fee works because high input cost observability lets the consultant credibly commit to a price (she knows what the engagement will cost her and is willing to accept the price risk), and because the discrete deliverable serves as a sufficient statistic for effort (a satisfactory deliverable signals adequate effort, and an unsatisfactory one signals inadequate effort). The consultant absorbs the cost risk, the client gets a contractually defined scope at a contractually defined price, and the deliverable resolves the informational tension. Project-based pricing is the framework's answer in the top-right not because outcomes are perfectly measured, but because the deliverable is a sufficient proxy and cost commitment is feasible. Appendix A, Proposition 2, states the precise sense in which the fixed fee implements the optimum: exactly when the outcome signal is noiseless, and approximately, with an error that vanishes with the outcome noise, when it is nearly so.

Bottom-left: Access-based pricing. When both outcome observability and input cost observability are low, neither signal is informative enough to justify a high-powered contract. Variable compensation linked to either signal imposes risk on the consultant. The framework predicts that the total power of incentives should fall, and in the limit, the variable component should disappear entirely, leaving only a fixed fee. This is access-based pricing: a retainer paid for continuous availability and advisory access, with no link to specific deliverables or measurable outcomes. The client pays for the right to engage the consultant when needed. The consultant offers responsiveness and expertise on demand. This structure parallels subscription or membership models in other knowledge-intensive sectors: legal counsel on retainer, an investment advisory relationship, a strategic advisor whose presence is the deliverable. The unit of value is availability, because no other unit is reliably measurable.

Bottom-right: Outcome-based pricing. When outcomes are clearly measurable and verifiable but costs are opaque, the optimal contract loads its variable component onto outcomes. The consultant's compensation is tied directly to realized performance: revenue improvements, cost savings, accuracy gains, conversion lifts, or other quantifiable metrics. Cost-side compensation receives little weight because the cost signal is too noisy to inform contracting. Outcome-based contracts allocate risk to the consultant, but they do so in exchange for information: the consultant bears risk on a signal that is highly informative about her effort and contribution. This pricing structure has long been favored in performance-oriented consulting domains such as marketing optimization paid per attributable revenue lift, recovery audits paid per dollar recovered, M&A advisory paid on transaction value, where the outcome is both measurable and reasonably attributable to the consultant's work.

Center: Hybrid pricing. When both signals carry partial but non-redundant information about effort, the optimal contract uses both. The framework predicts a positive base fee (to cover the consultant's cost of taking on the engagement), together with a positive input-linked component (to compensate the informativeness of cost-side signals) and/or a positive outcome-linked component (to compensate the informativeness of outcome signals). The relative weights on the variable components follow the informativeness principle: more weight on the cleaner signal. Hybrid pricing is not a compromise between two pure regimes. It is the strictly optimal contract whenever the information environment is neither purely outcome-informative nor purely cost-informative.

The map's structure carries a managerial implication that the regional descriptions make explicit: pricing decisions are informational decisions. The five regimes are not a menu from which to pick based on convention or comfort. They are conditional optima, each optimal in a specific informational region and suboptimal outside it. A consulting firm using time-based pricing in a region where outcomes are now observable is leaving information unused and revenue along with it. A firm using outcome-based pricing in a region where outcomes remain noisy is imposing unnecessary risk on itself. The strategic question is not which model is best in general but which model matches the information environment of the engagement at hand.

5 How AI reshapes the parameters

The framework's prediction that pricing structures track parameter values means that any technology which systematically alters those parameters will systematically alter the optimal pricing structure. We argue that AI is doing exactly that, in two distinct and analytically separable ways.

5.1 AI augmentation lowers input cost observability

Consulting firms have rapidly deployed AI internally to augment human consultants. Generative AI platforms summarize reports, draft deliverables, and interpret client data. Predictive analytics engines flag emerging trends. Knowledge-management systems compress weeks of research into hours of synthesis. McKinsey's Lilli platform, rolled out firmwide in July 2023 and publicly announced the following month, was built to transform how the firm accesses and synthesizes its institutional knowledge (McKinsey & Company, 2023), and reportedly delivers up to 30 percent time savings in searching and synthesizing knowledge (McKinsey & Company, n.d.). Deloitte's PairD platform, also launched in 2023 for internal deployment, similarly augments consultant productivity across analytical and drafting workflows (Wilkinson, 2024).

These deployments are visible to the consulting firm but largely invisible to the client. The client continues to receive PowerPoint decks, dashboards, implementation roadmaps, and recommendations; the form of deliverables remains unchanged. What changes is the effort required to produce them, and the cost structure underlying that effort. Tasks that previously took a team of analysts several weeks may now take one consultant several hours with AI support. New categories of work, like prompt engineering, model oversight, data preparation for AI consumption, and AI infrastructure maintenance, emerge as hidden inputs whose costs the client cannot observe.

In the framework's language, AI augmentation lowers input cost observability. The consultant knows that the deliverable was produced with two consultant-hours and substantial AI infrastructure spend. The client sees only the deliverable. Reported consultant time becomes a noisy signal of true value-creating effort, because the relationship between hours and effort is no longer linear or visible. The classical pricing formula, number of consultants times hours times rate, describes less and less of what the engagement actually cost the consulting firm and what it actually delivered to the client.

As input cost observability falls, the informativeness principle implies that the optimal weight on the cost signal falls. Time-based pricing becomes a worse instrument than it was, because hours are no longer reliable proxies for effort. If outcome observability has not changed correspondingly, the contract moves downward on the map, into the access-based region if outcomes remain very noisy, or into the hybrid region if outcomes are partially measurable. The intensity result discussed in Section 3 reinforces this directionality: as the cost signal becomes uninformative, the total power of input-linked compensation falls, and the contract relies more heavily on either fixed components (retainer-like elements) or outcome-linked components (where outcomes are observable enough to bear weight).

In practice, this directional shift implies that in the post-augmentation era, optimal pricing increasingly looks like a hybrid scheme that combines a base fixed fee, tied to access or to project-specific arrangements, with outcome-linked bonuses. Pure time-and-materials billing only survives at the margins of the practice, in highly specialized expert advisory, in engagements that genuinely remain human-intensive, but its share of the consulting revenue mix shrinks as AI augmentation diffuses.

This agrees with recent trends in the consulting industry. At a November 2025 media briefing, McKinsey's UK managing partner Michael Birshan said the firm is doing more performance-based arrangements with clients, and McKinsey disclosed that roughly a quarter of its global fees now come from outcome-based pricing (Thompson, 2025).

The implication for clients is also important. Clients negotiating time-based contracts with consulting firms that deploy AI internally pay for hours that no longer correspond to the work being done. Sophisticated clients are pushing back, demanding arrangements that include outcome-linked components (where the consulting firm shares in measurable value creation). PwC's US Tax Leader Krishnan Chandrasekhar has acknowledged publicly that "time's becoming less and less of relevance" in the firm's pricing conversations with clients, and the firm is exploring how to charge based on value and outcomes rather than billable hours (Trentmann, 2026).

Our framework justifies this client pressure: time-based contracts in AI-augmented engagements impose informational costs on clients without compensating benefits.

5.2 AI productization compresses outcome noise

The second AI transition is qualitatively different. When AI is not only an internal augmentation tool but becomes an externalized agent, deployed in the client's environment, integrated with client data, and performing analytical or operational functions continuously, the information environment changes fundamentally. This shift is now visible in the public commercial positioning of major consulting and technology-services firms. PwC's launch of PwC One in early 2026 has been described by the firm as an inversion of its previous AI strategy: where earlier internal tools helped staff work faster while clients experienced the same engagement model, PwC One embeds AI directly into how projects are delivered to clients. As Joe Voyles, Principal of AI and Emerging Technology at PwC, put it, the platform is "an entryway for our clients to start to consume service offerings that are meant to be more kind of autonomous and agent oriented . . . like a front door into the firm" (Taft, 2026). PwC's own announcement is consistent in direction while stopping short of pricing specifics: it presents PwC One as an environment in which AI capabilities perform analytical work under professional oversight, and frames the platform as building toward a new delivery model that moves clients beyond episodic projects toward continuous insight (PwC, 2026).

At the same time, Accenture's AI Refinery, which serves as the architectural backbone for designing and deploying agentic systems across enterprise clients, exemplifies this transition. Through partnerships with major AI hardware and model providers, Accenture offers an agent builder toolset and a multi-agent orchestration framework that enables clients to assemble networks of specialized AI agents performing discrete tasks in collaboration (Accenture, 2025).

When AI agents operate in client environments, they generate continuous outcome telemetry. Their predictive accuracy can be measured against ground truth in real time. Their decisions can be audited. The business effects of their recommendations, conversions lifted, costs reduced, compliance incidents avoided, can be tracked through the same data infrastructure that runs the client's operations. The consultant's product becomes observable in a way that the consultant's effort never was. In the framework's language, AI productization compresses outcome noise sharply.

It is worth being precise about *which* outcome noise it compresses. Low outcome observability has two distinct sources: how precisely the result can be measured, and how cleanly it can be attributed to the consultant's work rather than to market conditions, concurrent initiatives, or the client's own execution. Telemetry attacks the first: it drives measurement noise toward zero. It does not touch the second. An outcome measured perfectly is still uninformative about the consultant's contribution if that outcome is heavily confounded, and no amount of instrumentation resolves the confound. Attribution risk therefore sets a floor on outcome-signal informativeness that productization cannot lower, and outcome-based pricing becomes feasible only where attribution is clean to begin with. This is why the same telemetry that makes a fraud-detection agent's accuracy contractible does little for a strategy engagement whose value is entangled with everything else the client is doing: the measurement problem is solved in both, but the attribution problem is solved only in the first (Appendix A, Proposition 3).

Two features of this measurement gain locate it in a wider information-systems conversation and qualify it. The telemetry that renders the consultant's product observable is *digital trace data*, the operational by-product of a system's use that information-systems research has studied as a route to visibility into work that had previously eluded measurement (Howison, Wiggins, and Crowston, 2011; Newell and Marabelli, 2015). Cast this way, the source of the compression is explicit: AI-embedded delivery datafies the engagement, turning outcomes that were once tacit into contractible signals. But trace data is not a neutral by-product. When the party being measured knows that outcomes are metered and priced, it may begin to act on the traces themselves rather than only on the underlying work, so behavioral visibility can erode the very validity it promises (Aaltonen and Stelmaszak, 2024). The compression of measurement noise is therefore an equilibrium outcome, not a free technical fact: how far outcome noise falls depends on how the measured party responds to being measured, a caveat that compounds the attribution floor just described.

As outcome observability rises, the informativeness principle implies that the optimal weight on outcomes rises. Where attribution is clean, outcome-based pricing becomes informationally feasible in domains where it was previously unworkable: a risk-analytics agent priced on a base subscription plus a gain-sharing component tied to reduced losses. A marketing-optimization agent priced per campaign cycle scaled to performance improvements. A compliance agent priced on a base fee plus a variable component tied to incident reductions. The contract moves rightward on the map.

In combination with the access-style logic that a deployed AI agent is also an ongoing asset rather than a discrete deliverable, the productization transition pushes pricing toward the outcome-augmented hybrid in the bottom-right region. This resembles a contracting problem where the asset itself is valuable, usage is observable, and outcomes are increasingly measurable.

The productization transition also reshapes the consulting firm's economics. Recurring subscription revenue, scalable assets, and lower marginal delivery costs move consulting closer to software economics. The framework explains why this is an equilibrium rather than a fad: outcome observability has crossed a threshold at which value-based pricing becomes informationally efficient, so adopting the new structures maximizes the surplus the engagement can generate. That surplus is shared, not automatically retained by the firm: a firm captures its part through its bargaining share of a now-larger pie and by out-competing rivals still pricing from the old playbook, and plausibly captures a larger share as an embedded, differentiated asset shifts bargaining power its way (Appendix A, Proposition 4). Firms that retain pure time-based billing forgo surplus and, with it, whatever share of it they could have held. The competitive pressure to shift pricing is, in this sense, a pressure to track an informational reality that AI has changed.

Hence, optimal pricing in this case involves an asset-like hybrid scheme: a subscription fee combined with a usage-based component, an outcome-linked component, or both.

The commercial model of PwC is being redesigned along these lines: in an interview with the *Financial Times* at the platform's launch, PwC's US senior partner and CEO Paul Griggs said the firm would begin offering alternatives to traditional hours-based billing, converting some tax and consulting services into AI-powered tools that clients access directly, in some cases priced on a subscription or consumption basis rather than by billed hours (Foley, 2026).

6 Two relevant cases

The framework's predictions can be examined against the contractual practices of two AI-enabled service providers whose pricing structures are publicly observable: Palantir Technologies, which has visibly transitioned its pricing model over the past several years, and Salesforce's Agentforce platform, which has launched with an equilibrium pricing structure that the framework can interpret. Both cases were selected because their pricing engaged the transitions the framework describes (toward asset-based pricing schemes).

6.1 Palantir's transition from project-based to subscription-plus

Palantir Technologies offers an instructive case study of contractual transition that the framework can interpret. Founded in 2003 to provide data integration and analytics services to government and commercial clients, Palantir initially operated a services-intensive, project-based model: bespoke, engagement-scoped deployments built by its own engineers in client environments, under professional services arrangements that resembled traditional consulting engagements augmented by proprietary technology. In the framework's terms, this corresponded to the top-right region: outcomes were measurable in principle (intelligence and operational improvements from data integration) but deliverables like system deployments, integrations, and custom analytical modules served as the contractually anchored sufficient statistic, and the project-based contract form itself.

Over the past several years, Palantir's pricing model has transitioned substantially. The firm now offers software platforms: Foundry for commercial enterprises, Gotham for government, and Apollo as a deployment infrastructure, alongside the professional services. The new pricing structure is hybrid: a subscription/licensing component for platform access plus usage-based compute billing tied to consumption. Palantir's annual report describes revenue from software subscriptions and term licenses with ongoing operations-and-maintenance and professional services (Palantir Technologies, 2025), and the firm's filings explicitly name subscription- and usage-based pricing among its commercial strategies (Palantir Technologies, 2024). Commercial revenue has grown rapidly since the shift to this platform model, with clients accepting software-product terms rather than purely service-engagement terms (Palantir Technologies, 2025).

The framework interprets this transition as a movement on the strategy map from the top-right region toward the bottom-right region, driven by two parameter shifts. First, the productization of Palantir's offerings, turning what were custom integrations into deployable platforms, made the engagement continuously instrumented. The deployed platforms generated telemetry: usage patterns, query volumes, user adoption, decision throughput, and in some implementations, business outcomes from the analytics produced. Most of this telemetry meters consumption rather than value, and it is the outcome component that bears on the horizontal axis. Where the platform's effects are separable from the client's other activity, that component compressed measurement noise and the outcome signal became cleaner than it had been when the product was a custom integration whose effects were diffuse and slow to materialize. Where they are not, the attribution floor binds and the signal remains uninformative however finely it is measured (Appendix A, Proposition 3). Second, the platform model embedded Palantir's proprietary IP, data integration logic, ontology design, query optimization, in software that the client did not need to inspect to use, lowering input cost observability. The client paid for access to and usage of an asset whose underlying cost structure was opaque, not for transparent consultant-hours.

The framework predicts that this combination of rising outcome observability and falling input cost observability should push pricing toward the bottom-right region of the map: outcome-based or outcome-augmented hybrid structures. Palantir's adopted pricing, subscription plus usage plus, in some contracts, outcome-linked components for specific deployments, is such a structure. The subscription component compensates the asset's existence (the platform must be maintained, updated, and supported), the usage component compensates the variable cost of computational resources consumed, the outcome-linked components, where present, compensate the value the platform produces in specific use cases. Each component is the framework's answer to a distinct informational sub-problem: access pricing for the durable asset, usage pricing for the consumable resource, value pricing for the realized outcome.

6.2 Salesforce Agentforce's three-component hybrid

Salesforce's Agentforce represents a different case, illustrating the framework's prediction in equilibrium rather than in transition. Agentforce is Salesforce's AI agent platform, enabling enterprises to deploy autonomous and semi-autonomous AI agents for customer-facing and operational tasks within the Salesforce ecosystem. The pricing that Salesforce adopted for Agentforce combines a set of subscription, usage, and outcome-linked components. Salesforce's own pricing materials offer per-user licensing for platform access alongside consumption-based pricing, metered either per conversation or per action through "Flex Credits" that the firm describes as aligning cost with the business value its agents create (Salesforce, 2025, n.d.). In 2026 Salesforce extended this toward explicit outcome pricing: its Agentforce Help Agent introduced a pay-per-resolution model that charges only when an issue is resolved autonomously from start to finish, with no charge when the customer requests human escalation, so that, in the firm's description, the client's investment is tied directly to successful resolutions rather than to consumption (Salesforce, 2026).

The framework reads this three-component structure as the natural optimal contract for the information environment in which Agentforce operates. The subscription component prices access to the asset, the Agentforce platform, the model infrastructure, the ongoing development of agent capabilities. Without this component, the marginal cost of supporting a client would not be recovered, and the platform's existence would not be financially sustainable. The framework predicts a positive base fee whenever the engagement involves a durable asset whose existence has a positive marginal cost.

The usage component prices a different informational signal: the variable consumption of the platform's computational resources. Agent conversations and transactions consume compute, storage, model inference, and integration costs that scale with usage. The usage component compensates these variable costs and, importantly, aligns incentives: clients pay more when they use the platform more, which prevents low-intensity clients from subsidizing high-intensity clients. The framework's input-side weight, in this context, is the weight on usage rather than on consultant-hours; the underlying logic is identical.

The value-based variable component prices the outcome signal. When Agentforce deployments produce measurable business outcomes, like converted leads, resolved cases, reduced handling times, and increased deal velocity, Salesforce captures a share of the realized value through outcome-linked pricing. Not every deployment supports clean outcome attribution, but those that do generate the strongest revenue capture for Salesforce. The framework's outcome-side weight rises with the cleanliness of the outcome signal, and Salesforce's pricing structure makes this dependence explicit: outcome-linked components are larger in deployments where attribution is cleaner and smaller (or absent) in deployments where it is murky.

The three-component structure is the framework's prediction in the interior of the strategy map. Agentforce engagements are neither purely outcome-observable (some metrics are clean, others diffuse) nor purely cost-transparent (clients see usage but not the full infrastructure cost). The hybrid structure uses each informational signal as a contracting input in proportion to its informativeness: usage gets weight because usage is observable and meters variable consumption; outcomes get weight where outcomes are measurable; the base fee captures the rest as a fixed transfer that ensures Salesforce is willing to take on the engagement. No single-component pricing structure would achieve the same efficiency. The framework's hybrid prediction is consistent with Salesforce's commercial design.

7 Decision rules for consulting executives

The framework's analytical results translate into a small set of decision rules that consulting executives can apply directly. Applying them begins with locating the engagement on the strategy map. Outcome observability is gauged by asking whether the engagement's results can be measured cleanly and quickly and whether those results can be cleanly attributed to the consultant's work rather than confounded by market conditions, concurrent initiatives, or client-side execution. Input cost observability is gauged by asking how far the consultant's true cost of delivery is hidden from the client: how much of the work is AI-augmented in ways the client cannot observe, whether infrastructure costs are opaque, whether the hours–effort link has been broken by automation, and whether proprietary IP or methodology makes inputs unverifiable. The more the true cost of delivery is hidden in these ways, the lower the engagement's input cost observability, and the lower it sits on the map. With the engagement so located, four rules capture the framework's most actionable implications.

Rule 1: Match contract power to total signal informativeness, not to convention or comfort. The total intensity of the variable component in a pricing contract should depend on how informative the available signals are about consultant effort. When both outcome and cost signals are informative (the engagement sits in the top-right of the map), high-powered structures (fixed-fee project pricing) are efficient because variable compensation buys real information at low risk premium. When both signals are noisy (bottom-left of the map), low-powered structures (retainers, access pricing) dominate because variable compensation imposes risk on the consultant without informational return.

Rule 2: Let the AI deployment context dictate the direction of pricing transition. AI affects pricing differently depending on whether it is deployed internally or externally. AI tools used internally to augment consultants, like generative AI platforms, knowledge-management systems, and analytical automation, primarily lower input cost observability without correspondingly raising outcome observability. The framework predicts the contract should move downward on the map, away from pure time-based pricing and toward access-based or hybrid structures. AI used externally as a productized agent (embedded in client environments), generating telemetry, compresses outcome noise. The framework predicts the contract should move rightward, toward outcome-based and outcome-augmented hybrids. A consulting firm running both internal AI augmentation and external AI productization should expect both shifts simultaneously, converging on bottom-right hybrids combining subscription, usage, and outcome components.

Rule 3: Invest in measurement infrastructure before shifting to outcome-based pricing, but check whether measurement is the binding constraint. The framework treats outcome observability as a parameter, not a choice variable, but in practice one of its two components can be improved by investment. Measurement noise is reduced by telemetry, real-time instrumentation, agreed-upon ground-truth metrics, and third-party verification. A firm that shifts to outcome-based pricing before this measurement infrastructure exists imposes risk on itself that the framework would not justify: the outcome signal is too noisy to bear the contractual weight, and the firm absorbs variance without the informational benefit. The correct sequence is to invest in measurement first, then negotiate outcome-linked pricing. The caveat is that measurement is not always the binding constraint. Attribution risk, the confounding of the outcome by factors outside the consultant's work, is not reduced by telemetry, however fine. It falls only under credible causal identification (randomized holdouts, controlled rollouts, clean counterfactuals), which many engagements cannot support. Where attribution is the binding constraint, measurement investment does not unlock outcome-based pricing, and the firm should hold to hybrid or access structures regardless of how precisely it can measure (Appendix A, Proposition 3). Assessing attribution separately from measurability is what identifies which constraint binds.

Rule 4: Treat hybrid pricing as the equilibrium under partial information. Hybrid pricing is the optimal contract whenever both outcome and cost signals carry partial but non-redundant information about effort, which is the typical condition of AI-enabled engagements. The interior of the strategy map is large and is growing as AI proliferates. The share of engagements that sit in pure corner regions is shrinking. Consulting firms should design hybrid pricing capabilities as a strategic competence. The right hybrid for a given engagement is determined by the framework's relative-weighting result: the variable component split between input-linked and outcome-linked compensation should reflect the relative cleanliness of the two signals.

These four rules give consulting executives a small repertoire of decisions: how powerful should the contract be (Rule 1), in which direction is the pricing transition heading (Rule 2), what infrastructure investments must precede pricing changes (Rule 3), and what is the equilibrium contract structure under typical conditions (Rule 4). Applied once the engagement has been located on the map's two informational dimensions, the rules provide an operational toolkit for AI-era consulting pricing.

8 Conclusion

Pricing in the consulting industry has historically evolved through accretion: new models added to the practice as new informational conditions emerged, with little explicit reasoning about why a given model was optimal in a given context. The result is a heterogeneous landscape of time-based, project-based, access-based, outcome-based, and hybrid contracts that practitioners apply by convention and adapt by trial and error.

This article has argued that the heterogeneity is not arbitrary. The five canonical pricing regimes are conditional optima of a single underlying contracting problem, each optimal in a specific region of an information environment defined by two key parameters: outcome observability and input cost observability. The principal–agent framework developed explains why the canonical regimes are optimal choices and identifies the two AI transitions, augmentation and productization, as parameter shifts whose directional effects can be predicted rather than asserted.

With respect to the second transition, we show that outcome noise decomposes into measurement noise, which telemetry compresses, and attribution risk, which it cannot touch.

Outcome-based pricing therefore spreads exactly as far as attribution is clean, and no further. This attribution floor explains an otherwise puzzling pattern in the practice: the same telemetry that makes a fraud-detection agent's accuracy contractible does little for a strategy engagement whose value is entangled with the client's other initiatives. It also converts a vague managerial intuition, that outcome-based pricing "doesn't work everywhere," into a precise diagnostic: ask not whether outcomes are measured, but whether they are attributable, and expect hybrid contracts to be the resting point of the productization transition, not a waystation, wherever they are not.

Our contribution should be read for what it is: a practice-oriented framework. The formal model in Appendix A is deliberately built from standard components of contract theory. Its role is to discipline the strategy map and the decision rules, so that the regions of the map and the directions of the AI transitions follow from explicit assumptions rather than from intuition alone. The value we claim lies in the translation: a map on which an executive can locate an engagement, and rules by which the pricing structure follows from that location.

This framing also defines a research agenda. The framework's comparative statics are directional predictions that invite systematic empirical testing, for example using panel data on contract terms as AI deployment intensifies across engagements, firms, and service lines. The qualitative evidence base would be strengthened by interviews spanning multiple providers and, crucially, clients, whose perceptions of observability ultimately determine which contracts are acceptable. And the model itself invites extensions the practice is already gesturing toward: competition among providers, multi-period reputation effects, and the endogenous investment in measurement and attribution technology that our analysis treats as given.

The broader strategic implication is that consulting firms which understand the informational mechanics of AI's effect on pricing will design contract structures that competitors copying from the old playbook will not. The transition from a human-effort production economy to an AI-augmented intellectual-property economy is, in essential structure, an informational transition. The firms that internalize this insight and act on it by investing in measurement, designing hybrid contract structures, pricing access to assets rather than hours of effort, generate more surplus from a given engagement than rivals who do not. The firms that do not adapt will find that the old pricing models describe less and less of what they actually deliver, leaving unrealized surplus on the table and ceding it to competitors who price to the new informational reality.

A Formal model and derivations

This appendix applies the classical multi-signal principal–agent framework to a two-signal specification calibrated to the consulting context, and shows how the central results of that framework generate the strategy map of Figure 1. The contracting setup follows the CARA–normal moral hazard tradition of Holmström (1979), with the two-signal extension formalized by Banker and Datar (1989). The linearity of the optimal contract is the result of Holmström and Milgrom (1987). The comparative statics are standard implications of the framework as treated in Bolton and Dewatripont (2005). What the appendix contributes is the calibration: the choice of two signals corresponding to outcome observability and input cost observability, the interpretation of the cost signal's variance as the reduced form of the consultant's private cost information, the interpretation of the corner cases as the four canonical consulting pricing regimes, and the identification of AI augmentation and AI productization as parameter displacements in the resulting strategy map.

A.1 Environment and assumptions

A risk-neutral principal P (the client) hires a risk-averse agent A (the consultant) for a single project. The agent chooses an unobservable effort $e \geq 0$, generating gross client value

$$\pi = Ve, \qquad V > 0, \tag{2}$$

at a private disutility

$$\psi(e) = \tfrac{k}{2}e^2, \tag{3}$$

in which the marginal-effort-cost type $k \in \{k_L, k_H\}$ with $0 < k_L < k_H$ is privately known to the consultant, with $\Pr(k_L) = \varphi \in (0, 1)$. The efficient type is the low-k one: reusable intellectual property and methodology amortized across clients let it reach any given deliverable with less effort, an advantage the client cannot verify. Because k scales the marginal cost of effort, private knowledge of it is the hidden-information primitive.

Two signals of effort are contractible ex post:

$$y = e + \eta + \varepsilon_m, \qquad \eta \sim \mathcal{N}\left(0, \sigma_\eta^2\right), \quad \varepsilon_m \sim \mathcal{N}\left(0, \sigma_m^2\right), \tag{4}$$

$$t = e + \varepsilon_c, \qquad \varepsilon_c \sim \mathcal{N}\left(0, \sigma_c^2\right), \tag{5}$$

with η, ε_m, and ε_c mutually independent. The signal y is the client-observable outcome, and its noise has two economically distinct sources. The *attribution* component η is the part of the observed outcome driven by factors outside the consultant's effort (e.g., market conditions, other initiatives, client-side execution) that are confounded with e. It is uncertainty in the outcome-to-effort link itself. The *measurement* component ε_m is instrument error in recording the outcome. Their sum is the aggregate outcome-signal noise,

$$\sigma_y^2 \;\equiv\; \sigma_\eta^2 + \sigma_m^2, \tag{6}$$

and, as shown below, the solved contract depends on the two only through this sum. The distinction nonetheless carries real contractual weight, because the components respond differently to technology: measurement noise σ_m^2 is reducible by telemetry and instrumentation, whereas attribution risk σ_η^2 is not. The signal t is the consultant-side cost or time report. Its noise σ_c^2 is not primitive measurement error: it is the *reduced form* of the consultant's private cost information, in the sense made precise in Section A.2. Because the client cannot verify the report against the consultant's true cost structure, t is only an imperfect signal of value-creating effort, and the degree of that imperfection is what low input cost observability measures.

The principal commits ex ante to a contract payment to the agent

$$w(y,t) = \alpha + \beta_y\, y + \beta_c\, t, \qquad \alpha \in \mathbb{R},\ \beta_y, \beta_c \geq 0. \tag{7}$$

This pricing contract has three components: The fixed component (α) is the part of the payment that does not depend on either signal, a flat amount the consultant receives regardless of how the engagement turns out. It functions as a base fee, a retainer, or the level transfer that ensures the consultant is willing to accept the engagement. In the access-based regime, it is the whole contract. In every other regime, it is one component among several. The loading on the outcome signal (β_y) is the sensitivity of payment to the realized outcome y. If the engagement produces an outcome of y, the consultant receives an additional $\beta_y \cdot y$ on top of the fixed component. In an outcome-based contract, this is the dominant variable component, the gain-sharing fee, the performance bonus, the value-linked payment. In time-based contracts, it is approximately zero. The loading on the cost/input signal (β_c). This is the sensitivity of payment to the realized cost report t. If the consultant reports time/cost of t, she receives an additional $\beta_c \cdot t$ on top of the fixed component. In a time-and-materials contract, this is essentially the hourly rate, payment scales linearly with reported hours. In outcome-based contracts, it is approximately zero.

The agent has constant absolute risk aversion (CARA) with coefficient $r > 0$, and the principal is risk-neutral. The agent's reservation utility is normalized to zero. The two informational axes of Figure 1 correspond one-to-one to the two signal variances: outcome observability is decreasing in σ_y^2, and input cost observability is decreasing in σ_c^2.

Article axis	Model parameter
Outcome observability ↑	$\sigma_y^2 \downarrow$
Input cost observability ↑	$\sigma_c^2 \downarrow$

A.2 Input cost observability as the reduced form of private cost knowledge

The parameter σ_c^2 admits two readings, and the distinction governs what the model may and may not claim. Read narrowly, it is the variance of a second noisy signal of effort, and the input-cost-observability axis is then nothing more than the informativeness of the input signal. Read as we intend it, σ_c^2 is the *reduced form* of the consultant's private information about the engagement's cost structure.

To see this, suppose the consultant privately observes a cost-technology state θ that governs how a given deliverable maps into inputs: how much of the work rests on reusable intellectual property whose development cost is sunk, on methodology amortized across clients, or on infrastructure whose marginal cost is decoupled from the engagement. The client observes neither θ nor the true value-creating effort e; she observes only the reported cost or time t, and she cannot verify that report against θ. When the mapping from inputs to effort is stable and commonly understood (low cost heterogeneity), t tracks e closely and is a precise signal, so σ_c^2 is small. When the consultant privately commands a wide range of input-to-effort mappings (e.g., when identical deliverables can be produced at widely varying, unverifiable combinations of inputs), the client's best read of e from any given t is correspondingly diffuse, so σ_c^2 is large. Private cost knowledge is thus what generates the noise in the cost signal, and the dispersion of the privately known state is what low input cost observability measures. In the limit, when the report is wholly unverifiable and the state distribution is wide, t carries no information about e from the client's side and $\sigma_c^2 \to \infty$.

This reading is what makes the two axes genuinely distinct informational objects rather than one friction measured twice. The outcome axis σ_y^2 is a hidden-action friction: effort is unobservable and y reads it with exogenous noise. The cost axis σ_c^2 is, in reduced form, a hidden-information friction: the consultant privately knows the cost structure, and it is that private knowledge that degrades t as a signal of effort. The two canonical frictions of contract theory therefore enter on separate axes.

A.3 The contracting problem and its solution

Conditional on the contract $(\alpha, \beta_y, \beta_c)$ and effort e, the agent's certainty equivalent is

$$\mathrm{CE}_A(e) = \alpha + (\beta_y + \beta_c)\, e - \tfrac{k}{2}e^2 - \tfrac{r}{2}\left(\beta_y^2\sigma_y^2 + \beta_c^2\sigma_c^2\right). \tag{8}$$

Having accepted the contract, the agent chooses effort by solving

$$\max_{e \geq 0}\ \mathrm{CE}_A(e), \tag{9}$$

subject to the single constraint that effort be non-negative. This is because effort is a physical quantity, hours worked, intensity of analysis, care applied that cannot be negative. The agent can always choose $e = 0$ if she finds the contract unrewarding at the margin. No upper bound on effort is imposed because the quadratic disutility $\psi(e) = \frac{k}{2}e^2$ rises sharply enough that the agent never chooses effort near any plausible physical maximum. The risk premium term $\frac{r}{2}(\beta_y^2\sigma_y^2 + \beta_c^2\sigma_c^2)$ does not depend on e and therefore is taken as given by the agent when she chooses her optimal effort level.

Note that the non-negativity constraint does not bind at the optimum, and the first-order condition characterizes the agent's effort choice. The incentive-compatible effort is

$$e^*(\beta_y, \beta_c) = \frac{\beta_y + \beta_c}{k}. \tag{10}$$

Total effort responds only to the sum of pay sensitivities.

Two further constraints govern the contract design problem. The agent will accept the contract only if her certainty equivalent at the optimal effort weakly exceeds her outside option, which we have normalized to zero. This is the *individual rationality* constraint:

$$\mathrm{CE}_A(e^*(\beta_y, \beta_c))\ \geq\ 0. \tag{11}$$

Substituting the optimal effort (10) into the certainty equivalent (8) gives the explicit form

$$\alpha + \frac{(\beta_y + \beta_c)^2}{2k} - \frac{r}{2}\left(\beta_y^2\sigma_y^2 + \beta_c^2\sigma_c^2\right)\ \geq\ 0. \tag{12}$$

Because the principal extracts surplus only through the fixed transfer α, and because lowering α reduces the agent's payment without affecting her effort choice, the principal sets α at the lowest level consistent with the agent's participation. The individual rationality constraint therefore binds at the optimum, and solving (12) at equality for α yields

$$\alpha^* = -\frac{(\beta_y + \beta_c)^2}{2k} + \frac{r}{2}\left(\beta_y^2\sigma_y^2 + \beta_c^2\sigma_c^2\right). \tag{13}$$

Binding the participation constraint at the reservation level assigns all bargaining power to the client. We relax it in Section A.7, where the consultant retains a share of the surplus and the paper's claim that firms capture the value of the AI transition is made precise.

Substituting into the principal's objective $Ve^* - \mathbb{E}[w]$ and simplifying gives the reduced-form problem

$$\max_{\beta_y,\beta_c \geq 0} \ \frac{V(\beta_y + \beta_c)}{k} \ - \ \frac{(\beta_y + \beta_c)^2}{2k} \ - \ \frac{r}{2}\left(\beta_y^2\sigma_y^2 + \beta_c^2\sigma_c^2\right). \tag{14}$$

Because participation binds, this objective is exactly total certainty-equivalent surplus—value Ve^* net of the effort cost $\psi(e^*)$ and the risk premium, so the contract the map describes is *constrained-efficient*: it maximizes the joint surplus given unobservable effort and a risk-averse agent, and a planner facing the same information would choose it. The map is thus optimal in the efficiency sense, not from one party's side. By Proposition 4 the loadings that solve (14) are invariant to how that surplus is divided. Client-optimal, planner-optimal, and any bargaining outcome coincide on the map, differing only in the fixed transfer α.

Let $s \equiv \beta_y + \beta_c$ denote total pay-for-performance power. For any fixed s, the principal minimizes the risk premium subject to $\beta_y + \beta_c = s$; first-order conditions yield $\beta_y\sigma_y^2 = \beta_c\sigma_c^2$ and the *informativeness shares*:

$$\boxed{\frac{\beta_y^*}{s} = \frac{\sigma_c^2}{\sigma_y^2 + \sigma_c^2}, \qquad \frac{\beta_c^*}{s} = \frac{\sigma_y^2}{\sigma_y^2 + \sigma_c^2}.} \tag{15}$$

Each signal loads in inverse proportion to its own variance, which is the multi-signal generalization of the Holmström (1979) informativeness principle.

Substituting the optimal split back into (14) and maximizing over s yields

$$\boxed{s^* = \frac{V}{1 + rk\,\dfrac{\sigma_y^2\sigma_c^2}{\sigma_y^2 + \sigma_c^2}}.} \tag{16}$$

The quantity $\sigma_y^2\sigma_c^2/(\sigma_y^2 + \sigma_c^2)$ is the residual aggregate noise after optimally combining the two signals. As either variance falls toward zero, this aggregate falls and $s^* \to V$ (first best). As both grow without bound, the aggregate grows and $s^* \to 0$.

Proposition 1 (Comparative statics)**.** *The optimal contract* $(\alpha^*, \beta_y^*, \beta_c^*)$ *satisfies:*

(i) $\partial\beta_y^*/\partial\sigma_y^2 < 0$ *and* $\partial\beta_c^*/\partial\sigma_c^2 < 0$.

(ii) *The relative weight* β_y^*/β_c^* *rises with* σ_c^2 *and falls with* σ_y^2.

(iii) $\partial s^*/\partial\sigma_y^2 < 0$ *and* $\partial s^*/\partial\sigma_c^2 < 0$.

(iv) *As* $\sigma_y^2, \sigma_c^2 \to 0$, $s^* \to V$ *(first best).*

(v) *As* $\sigma_y^2, \sigma_c^2 \to \infty$, $s^* \to 0$ *and the contract collapses to the fixed component* α^*.

Proof. Combining the informativeness split (15) with the power (16) and writing $D \equiv \sigma_y^2 + \sigma_c^2 + rk\,\sigma_y^2\sigma_c^2 > 0$ gives the closed forms

$$\beta_y^* = \frac{V\sigma_c^2}{D}, \qquad \beta_c^* = \frac{V\sigma_y^2}{D}, \qquad s^* = \beta_y^* + \beta_c^* = \frac{V\left(\sigma_y^2 + \sigma_c^2\right)}{D}.$$

(i) Since $\partial D/\partial\sigma_y^2 = 1 + rk\sigma_c^2$ and $\partial D/\partial\sigma_c^2 = 1 + rk\sigma_y^2$,

$$\frac{\partial\beta_y^*}{\partial\sigma_y^2} = -\frac{V\sigma_c^2\left(1 + rk\sigma_c^2\right)}{D^2} < 0, \qquad \frac{\partial\beta_c^*}{\partial\sigma_c^2} = -\frac{V\sigma_y^2\left(1 + rk\sigma_y^2\right)}{D^2} < 0,$$

every factor being positive.

(ii) The split (15) makes the relative weight independent of s^*, namely $\beta_y^*/\beta_c^* = \sigma_c^2/\sigma_y^2$, so

$$\frac{\partial}{\partial\sigma_y^2}\left(\frac{\beta_y^*}{\beta_c^*}\right) = -\frac{\sigma_c^2}{\sigma_y^4} < 0, \qquad \frac{\partial}{\partial\sigma_c^2}\left(\frac{\beta_y^*}{\beta_c^*}\right) = \frac{1}{\sigma_y^2} > 0.$$

(iii) Differentiating $s^* = V(\sigma_y^2 + \sigma_c^2)/D$, the numerator terms cancel except the $-rk$ contribution, leaving

$$\frac{\partial s^*}{\partial\sigma_y^2} = -\frac{Vrk\,\sigma_c^4}{D^2} < 0, \qquad \frac{\partial s^*}{\partial\sigma_c^2} = -\frac{Vrk\,\sigma_y^4}{D^2} < 0.$$

Equivalently, (16) makes s^* a decreasing function of the aggregate noise $\sigma_y^2\sigma_c^2/(\sigma_y^2 + \sigma_c^2)$, which is increasing in each variance (its partial in σ_y^2 is $\sigma_c^4/(\sigma_y^2 + \sigma_c^2)^2 > 0$).

(iv)–(v) From (16), as $\sigma_y^2, \sigma_c^2 \to 0$ the aggregate noise vanishes and $s^* \to V$; as $\sigma_y^2, \sigma_c^2 \to \infty$ it grows without bound, so $s^* \to 0$ and $\beta_y^*, \beta_c^* \to 0$, leaving only the fixed transfer α^*. □

The private type now enters the effort choice $e^*(k) = s/k$ with $s = \beta_y + \beta_c$. Write the aggregate signal noise as $\Sigma \equiv \sigma_y^2\sigma_c^2/(\sigma_y^2 + \sigma_c^2)$. The split $\beta_y{:}\beta_c$ that determines which signal the contract loads on, and hence which regime the engagement occupies, is fixed by the informativeness condition (15), $\beta_y\sigma_y^2 = \beta_c\sigma_c^2$, which depends only on the signal variances, not on k. Every type therefore contracts on the same ray from the origin. The type affects only how far out along that ray the contract sits, its power s, not its direction. The map as drawn is the efficient (symmetric-information) benchmark, at which a type-k contract has power $s^*(k) = V/(1 + rk\Sigma)$ from (16). Private cost information displaces the *inefficient* type's contract radially inward from this benchmark.

A.4 Mapping the solution onto the strategy map

The five regimes in the main text each correspond to a region of the (σ_y^2, σ_c^2) plane, and the map is a statement about the optimal linear contract's structure there—its power $s^* = \beta_y^* + \beta_c^*$ and its split $\beta_y^*{:}\beta_c^*$. That structure is derived by the model in all five regions alike. In four regions the reading is literal: loading on the input signal is time-and-materials, loading on the outcome is value-based pricing, loading on neither ($s^* \to 0$) is a flat retainer, and loading on both is a hybrid. The top-right is the exception. Its structure is derived like the rest: $s^* \to V$, a high-powered contract near first best, but its recognizable instrument, a fixed fee paid on completion, is a lump sum rather than a linear contract in the signals. Identifying that instrument with the derived structure is therefore not a reading but a limiting argument, given in Proposition 2.

Top-left: Time-based pricing (σ_y^2 large, σ_c^2 small). By (15), $\beta_y^*/s^* \to 0$ and $\beta_c^*/s^* \to 1$, the contract reduces to $w \approx \alpha^* + \beta_c^* t$. Pay is anchored to the input/time signal. This recovers time-and-materials billing.

Bottom-right: Outcome-based pricing (σ_y^2 small, σ_c^2 large). Symmetrically, $\beta_y^*/s^* \to 1$ and $\beta_c^*/s^* \to 0$ imply $w \approx \alpha^* + \beta_y^* y$. Pay is anchored to the realized outcome. This is value-based or performance-linked pricing.

Bottom-left: Access-based pricing (σ_y^2 and σ_c^2 both large). From (16), $s^* \to 0$, both pay-loadings collapse and only the fixed transfer α^* survives. The principal pays a flat retainer for availability.

Top-right: Project-based pricing (σ_y^2 and σ_c^2 both small). From (16), $s^* \to V$, the contract approaches first best with negligible risk premium. It is tempting to read the familiar practical implementation, a single fixed fee F paid on completion of a defined scope, as simply this optimal contract in disguise, but the two are not equivalent at positive noise: an unconditional lump sum provides no incentives and implements zero effort, while a completion-contingent fee on a noisy deliverable still loads risk on the risk-averse agent.

Only in the noiseless limit is there nothing to insure away. The fixed-fee implementation is therefore a *limiting* claim, exact only as the outcome signal becomes noiseless. Two points are worth flagging. First, the mechanism runs through the outcome signal: it is $\sigma_y^2 = 0$ that makes the deliverable a sufficient statistic for effort, so the commonly cited rationale that high input cost observability (σ_c^2 small) lets the consultant commit to a price is a complementary practical consideration rather than what underwrites first-best implementation in the model. Second, for σ_y^2 small but positive the optimal linear contract does not *become* a lump sum. Its realized payment merely converges to a constant, and the gap is governed by the residual risk premium.

Proposition 2 (Fixed-fee implementation of the top-right corner)**.** *Let $e^{\mathrm{FB}} = V/k$ denote first-best effort and $F^* = \psi(e^{\mathrm{FB}}) = V^2/(2k)$.*

(i) *If $\sigma_y^2 = 0$, then $y = e$ is a sufficient statistic for effort and the cost signal t is redundant. The forcing contract that pays F^* when the deliverable meets the completion standard $y \geq e^{\mathrm{FB}}$ and pays $\underline{w} < 0$ otherwise implements e^{FB} at zero risk premium, binds participation, and attains first best. This holds for any σ_c^2: outcome-side noiselessness alone underwrites the fixed fee.*

(ii) *As $(\sigma_y^2, \sigma_c^2) \to (0, 0)$, the optimal linear contract attains $e^* \to e^{\mathrm{FB}}$ with vanishing risk premium, and its realized payment converges in probability to the constant F^*. The optimal contract is then* ex post *indistinguishable from a fixed fee F^* paid on delivery.*

(iii) *For $\sigma_y^2 > 0$ the equivalence fails: an unconditional lump sum implements $e = 0$; the optimal (linear) contract carries strictly positive incentive power $s^* > 0$ and is not a fixed fee; and any completion-contingent fee on a noisy deliverable imposes a strictly positive risk premium on the agent. The accuracy of the fixed-fee reading is governed by the residual risk premium $\frac{r}{2}(\beta_y^{*2}\sigma_y^2 + \beta_c^{*2}\sigma_c^2)$, which vanishes as $(\sigma_y^2, \sigma_c^2) \to (0, 0)$.*

Proof. (i) With $\sigma_y^2 = 0$, $y = e$ almost surely, so the conditional law of y given e is degenerate at e and y determines e; by the sufficiency criterion for signals, t carries no information about e beyond y. Under the stated contract, any $e \geq e^{\mathrm{FB}}$ yields F^*. Because ψ is strictly increasing, the agent's net payoff $F^* - \psi(e)$ is strictly decreasing in e on $[e^{\mathrm{FB}}, \infty)$ and hence maximized on that interval at e^{FB}, where it equals $F^* - \psi(e^{\mathrm{FB}}) = 0$. Any $e < e^{\mathrm{FB}}$ yields at most $\underline{w} < 0$. Thus e^{FB} is optimal for the agent, participation binds at zero, and payment is nonrandom on the equilibrium path. Principal surplus is $Ve^{\mathrm{FB}} - F^* = V^2/(2k)$, the first-best level.

(ii) By (16), $s^* \to V$, so $e^* = s^*/k \to e^{\mathrm{FB}}$ and the risk premium $\frac{r}{2}(\beta_y^{*2}\sigma_y^2 + \beta_c^{*2}\sigma_c^2) \to 0$. Realized pay is $w = \alpha^* + s^*e^* + (\beta_y^*\varepsilon_y + \beta_c^*\varepsilon_c)$; the stochastic term has variance $\beta_y^{*2}\sigma_y^2 + \beta_c^{*2}\sigma_c^2 \to 0$

and so vanishes in probability, while $\alpha^* \to -V^2/(2k)$ by (13) and $s^*e^* \to V^2/k$. Hence $w \to V^2/(2k) = F^*$ in probability.

(iii) Setting $\beta_y = \beta_c = 0$ in (10) gives $e^* = 0$. For $\sigma_y^2 > 0$ with σ_c^2 finite, (16) gives $s^* > 0$, so the optimal contract, linear by Holmström and Milgrom (1987), has positive total power and is not a lump sum. A completion-contingent fee pays the Bernoulli amount $F^*\mathbf{1}\{y \geq \bar{y}\}$, whose conditional variance is strictly positive whenever $0 < \Pr(y \geq \bar{y} \mid e) < 1$. The CARA agent therefore bears a strictly positive risk premium. The stated order bound is the risk-premium term of (8). □

Center: Hybrid pricing (σ_y^2 and σ_c^2 both interior). Neither informativeness share in (15) is close to 0 or 1, and s^* in (16) is strictly interior. All three contractual components, α, β_y, and β_c, are active. Hybrid pricing is the strictly optimal contract whenever both signals carry partial, non-redundant information about effort.

The four corners are limiting reference cases: no engagement has literally zero or infinite noise, so none is realized exactly. Real engagements are interior points, which the model solves exactly, and the corners serve to name the poles they lean toward.

A.5 Comparative statics of the AI transitions

The article's two AI transitions are displacements in (σ_y^2, σ_c^2) space. Each transition has a dominant effect on one parameter and a secondary effect on the other.

AI augmentation (internal). AI augmentation widens the range of unverifiable input-to-effort mappings available to the consultant: identical deliverables can now be produced at widely different, privately known combinations of consultant-hours and infrastructure spend, so reported hours and infrastructure spend decouple from true value-creating effort. In the reduced form of Section A.2, this is a widening of the privately known cost-technology state, and its effect falls on σ_c^2: input cost observability falls. By Proposition 1(i) and (15), β_c^* falls and the relative weight on the outcome signal rises. Hence, the optimum moves *up* on the map.

Corollary 1 (The augmentation hybrid)**.** *Let σ_c^2 rise with σ_y^2 held fixed. The optimal contract remains hybrid and re-weights toward the outcome signal: β_c^* falls, β_y^* rises, the relative weight $\beta_y^*/\beta_c^* = \sigma_c^2/\sigma_y^2$ rises, and total power s^** falls. *As $\sigma_c^2 \to \infty$ the cost loading vanishes and only the outcome signal is priced, at the reduced power $s^* \to V/(1 + rk\sigma_y^2) < V$. The post-augmentation scheme is thus a* lower-powered *hybrid, a larger fixed retainer alongside a* *shrinking input-linked component and a residual outcome-linked one, because degrading the cost signal removes information without supplying any.*

AI productization (external). An AI agent embedded in the client's environment produces continuous outcome telemetry, uptime, predictive accuracy, conversion lifts, error rates. Its dominant effect is on the *measurement* component of outcome noise: telemetry compresses σ_m^2, and through (6) the aggregate σ_y^2 falls only to the extent that measurement, rather than attribution, was the binding source of noise. By Proposition 1(i), the fall in σ_y^2 raises β_y^*. Crucially, telemetry leaves the attribution component σ_η^2 untouched, so the rightward move is bounded by attribution risk in the sense made precise in Proposition 3 below. In practice productization typically runs on top of an augmentation capability the firm has already built, so there is also a secondary effect from the augmentation transition, which reduces the relative weight on the cost signal through (15). Under the typical condition that both shifts are present, and where attribution is clean enough for telemetry to bite, the optimum moves *right and up* on the map.

A.6 Attribution risk and the reach of productization

Even where productization sharply improves outcome observability, the optimal contract is unlikely to become a *pure* outcome-based scheme. Productization improves only the measurement component of outcome observability, not attribution: telemetry drives σ_m^2 toward zero but leaves σ_η^2 untouched, so the outcome signal retains an irreducible floor of noise, the cost signal keeps positive weight, and the optimum settles in the hybrid interior rather than the outcome-based corner, unless attribution, too, is clean. This places a ceiling on how far productization can move an engagement toward outcome-based pricing, which the following makes precise.

Proposition 3 (Attribution risk bounds the productization transition)**.** *Decompose outcome-signal noise as $\sigma_y^2 = \sigma_\eta^2 + \sigma_m^2$ per (6), with σ_η^2 attribution risk and σ_m^2 measurement noise. As telemetry perfects ($\sigma_m^2 \to 0$), the outcome loading rises only to the ceiling*

$$\frac{\beta_y^*}{s^*} \longrightarrow \frac{\sigma_c^2}{\sigma_\eta^2 + \sigma_c^2}, \tag{17}$$

which is strictly below one whenever $\sigma_\eta^2 > 0$, and total power is capped at $s^* \to \frac{V}{1+rk\,\sigma_\eta^2\sigma_c^2/(\sigma_\eta^2+\sigma_c^2)} < V$. Pure outcome-based pricing is reachable through productization only if attribution is also clean, $\sigma_\eta^2 \to 0$.

Proof. Setting $\sigma_m^2 = 0$ in (15) gives (17), which is below one iff $\sigma_\eta^2 > 0$ and tends to one iff $\sigma_\eta^2 \to 0$. The bound on s^* is (16) evaluated at $\sigma_y^2 = \sigma_\eta^2$. $\square$

This result sharpens the fixed-fee corner of Proposition 2: the exact first-best implementation there requires $\sigma_y^2 = 0$, which now means *both* clean attribution and perfect measurement. Telemetry alone does not deliver it.

Corollary 2 (The productization hybrid)**.** *Let σ_y^2 fall through its measurement component σ_m^2, with σ_c^2 held fixed. The optimal contract remains hybrid and re-weights toward the outcome signal: β_y^* rises, β_c^* falls, and total power s^** rises. *The re-weighting is bounded by attribution: as $\sigma_m^2 \to 0$ the outcome share is capped at $\beta_y^*/s^* \to \sigma_c^2/(\sigma_\eta^2 + \sigma_c^2) < 1$ whenever $\sigma_\eta^2 > 0$, so the cost component never vanishes. The post-productization scheme is thus a* higher-powered *hybrid, a base access (subscription) component plus a larger outcome-linked (gain-sharing) component that approaches but does not reach pure outcome-based pricing.*

The two schemes share a direction but differ in power. Both re-weight toward the outcome signal, yet for opposite reasons: augmentation degrades the cost signal and so lowers total incentive power, while productization improves the outcome signal and so raises it. This is the formal counterpart of the article's two hybrid schemes: augmentation yielding a lower-powered, retainer-heavy hybrid and productization a higher-powered, gain-sharing hybrid. The two transitions traverse contract space differently even though both push pay toward outcomes.

A.7 Who captures the value: bargaining power and the information rent

The baseline binds participation at the consultant's reservation utility, so, under symmetric cost information, the client captures the entire surplus and the consultant earns zero rent in every regime. Read literally, this says better pricing enriches only the client, which sits uneasily with the paper's claim that consulting firms capture the value of the AI transition. The tension dissolves once pie-size is separated from pie-split. The strategy map governs the *size* of the surplus. The baseline merely normalizes its *division.*

Let total certainty-equivalent surplus be

$$S(\beta_y, \beta_c) = \frac{V(\beta_y + \beta_c)}{k} - \frac{(\beta_y + \beta_c)^2}{2k} - \frac{r}{2}\left(\beta_y^2\sigma_y^2 + \beta_c^2\sigma_c^2\right), \tag{18}$$

which is exactly the reduced-form objective (14). We write $S^* = \max_\beta S(\beta_y, \beta_c)$. Generalize participation by letting the consultant receive a share $\lambda \in [0, 1]$ of surplus above her reservation utility, the generalized-Nash split with consultant weight λ, of which the baseline is the corner $\lambda = 0$.

Proposition 4 (Efficiency, distribution, and value capture). (i) *The surplus-maximizing loadings* (β_y^*, β_c^*) *and effort* e^* *solve* (14) *and are independent of* λ. *Distribution is implemented through the fixed transfer alone,* $\alpha_\lambda = \alpha^* + \lambda S^*$, *with* α^* *as in* (13). *Every result on the strategy map is therefore a statement about efficiency and is invariant to bargaining power.*

(ii) *The consultant's rent is* λS^* *and the client's is* $(1-\lambda)S^*$. *The baseline is* $\lambda = 0$, *in which the client captures all surplus.*

(iii) *For any* $\lambda > 0$, *the consultant's captured value strictly rises as a transition raises surplus:*

$$\frac{\partial(\lambda S^*)}{\partial \sigma_y^2} = -\tfrac{\lambda r}{2}\,\beta_y^{*2} < 0, \qquad \frac{\partial(\lambda S^*)}{\partial \sigma_c^2} = -\tfrac{\lambda r}{2}\,\beta_c^{*2} < 0. \tag{19}$$

Because $0 < \lambda < 1$ *leaves both shares increasing in* S^*, *a transition that expands surplus is captured in part by* both *parties. A firm that retains a structure mismatched to the information environment forgoes* $S^* - S(\beta) > 0$ *of surplus, of which it loses the share* λ.

Proof. (i) The transfer α does not enter (10), so effort and the surplus-maximizing loadings are independent of the level transfer; maximizing (14) gives (β_y^*, β_c^*). Setting the agent's certainty equivalent in (12) equal to λS^* and solving for α gives $\alpha_\lambda = \alpha^* + \lambda S^*$, with the baseline (13) recovered at $\lambda = 0$. (ii) is then immediate. (iii) By the envelope theorem applied to $S^* = \max_\beta$ of (18), $\partial S^*/\partial\sigma_y^2 = -\frac{r}{2}\beta_y^{*2}$ and $\partial S^*/\partial\sigma_c^2 = -\frac{r}{2}\beta_c^{*2}$; multiply by λ. $\square$

The incentive invariance in (i)[4] is what licenses the rest of the paper: the strategy map and every comparative static describe how AI reshapes the *structure* of efficient pricing, a claim that holds whoever captures the resulting surplus.

The value-capture claim of Sections 5 and 8 is the statement that adopting the efficient structure maximizes S^*, the surplus available to be shared; a firm captures value both through its bargaining share λ of a larger pie and, in a market with several firms, by out-competing rivals still pricing from the old playbook, since a firm that can offer the client more surplus wins the mandate while clearing its own participation constraint.

[4]This separation, in which the incentive loadings are chosen independently of the fixed transfer that divides the surplus, is the contracting-problem counterpart of a familiar result in the pricing of information goods: Sundararajan (2004) shows that the optimal usage-based pricing schedule is independent of the fixed fee, so the two can be designed sequentially. Here the loadings play the role of the usage schedule and α_λ that of the fixed fee.

The productization transition plausibly raises λ itself: an embedded, differentiated asset with switching costs shifts bargaining power toward the consultant, so the firm gains on both margins, a larger S^* and a larger share of it. Because λ is implemented through the negotiated transfer, the analysis is agnostic about which party nominally designs the contract, matching the practice in which pricing is negotiated rather than dictated.

A second source of consultant surplus operates alongside the bargaining share and, unlike it, requires no bargaining power at all: the information rent from the privately known marginal-cost type k. The reduced-form bridge of Section A.2 treated private cost knowledge only as the source of the cost signal's noise, delivering the decoupling of input-based pay from value but not a rent. The same private information, admitted directly into the model, yields a rent. Under a contract of power s written at the informativeness split, the agent chooses $e^* = s/k$, is paid $\alpha + s^2/k$ in expectation, bears effort cost $s^2/(2k)$, and carries a risk premium $\frac{r}{2}\Sigma s^2$, so its certainty equivalent is $\alpha + s^2/(2k) - \frac{r}{2}\Sigma s^2$. Only the term $s^2/(2k)$ depends on the type, and it is larger the lower is k: a low-k consultant draws more from any given contract than a high-k one. This is the single-crossing property that drives the screening solution. The client cannot separate the types without conceding the efficient one a rent. Each type's contract is written at the informativeness split, so it is summarized by its power s_i and its risk premium $\frac{r}{2}\Sigma s_i^2$; the surplus a type-k contract of power s generates is $S(k, s) = Vs/k - s^2/(2k) - \frac{r}{2}\Sigma s^2$.

The efficient type can always take the inefficient type's contract and deliver any given effort at lower cost, so truthful revelation requires leaving it the surplus it would gain by doing so. The standard screening solution binds the inefficient type's participation and the efficient type's incentive constraint. The efficient type is served at its undistorted benchmark power, $s_L^* = V/(1 + rk_L\Sigma)$ (no distortion at the top), while the inefficient type's power is distorted downward,

$$s_H^* = \frac{(1-\varphi)V/k_H}{(1-\varphi)\left(\frac{1}{k_H} + r\Sigma\right) + \varphi\left(\frac{1}{k_L} - \frac{1}{k_H}\right)} < \frac{V}{1 + rk_H\Sigma}, \tag{20}$$

and the efficient type is left an information rent

$$U_L^* = \frac{s_H^{*2}}{2}\left(\frac{1}{k_L} - \frac{1}{k_H}\right) > 0, \tag{21}$$

while the inefficient type earns zero. The rent and the distortion both vanish as $k_L \to k_H$ and both increase in the type spread $1/k_L - 1/k_H$ and in φ. As the invariance argument of Section A.3 guarantees, both contracts share the informativeness split, so the distortion is purely radial: the inefficient type sits at lower power on the *same* ray, and the engagement does not change regime.

The rent is the formal content of the claim in Section 1 that the client's information disadvantage "can be exploited rather than compensated": the efficient consultant, whose low marginal cost derives from reusable IP, commands a rent the client cannot contract away, because it can always understate its efficiency, take the easier contract, and pad the effort it claims the work required. The client's only instrument for shrinking that rent is to under-power the inefficient type's contract, pushing it toward cost-plus and distorting its effort downward. This is the equilibrium counterpart of the main model's prescription to reduce the loading on a poorly attributed cost signal. The rent is additive to the bargaining share: the consultant captures $\lambda S^* + \varphi U_L^*$ in expectation, the rent term accruing even at $\lambda = 0$, where the client holds all bargaining power. This is the precise sense in which firms capture the value they create in Sections 5 and 8.

References

Aaltonen, A. and M. Stelmaszak (2024). "The Performative Production of Trace Data in Knowledge Work," *Information Systems Research* 35(3): 1448–1462.

Accenture (2025). "Accenture Launches AI Refinery for Industry to Reinvent Processes and Accelerate Agentic AI Journeys," Accenture Newsroom, January 6, 2025. `https://newsroom.accenture.com/news/2025/accenture-launches-ai-refinery-for-industry-to-reinvent-processes-and-accelerate-agentic-ai-journeys`.

Bajari, P. and S. Tadelis (2001). "Incentives versus Transaction Costs: A Theory of Procurement Contracts," *RAND Journal of Economics* 32(3): 387–407.

Banker, R. D. and S. M. Datar (1989). "Sensitivity, Precision, and Linear Aggregation of Signals for Performance Evaluation," *Journal of Accounting Research* 27(1): 21–39.

Bolton, P. and M. Dewatripont (2005). *Contract Theory.* Cambridge, MA: MIT Press.

Brynjolfsson, E. (2022). "The Turing Trap: The Promise and Peril of Human-Like Artificial Intelligence," *Daedalus* 151(2): 272–287.

Brynjolfsson, E., D. Li, and L. Raymond (2025). "Generative AI at Work," *Quarterly Journal of Economics* 140(2): 889–942.

Chang, Y. B., V. Gurbaxani, and K. Ravindran (2017). "Information Technology Outsourcing: Asset Transfer and the Role of Contract," *MIS Quarterly* 41(3): 959–973.

Dey, D., M. Fan, and C. Zhang (2010). "Design and Analysis of Contracts for Software Outsourcing," *Information Systems Research* 21(1): 93–114.

Fitoussi, D. and V. Gurbaxani (2012). "IT Outsourcing Contracts and Performance Measurement," *Information Systems Research* 23(1): 129–143.

Foley, S. (2026). "PwC US Boss Says Partners Who Resist AI Have No Place at the Firm," *Financial Times*, March 19, 2026. `https://www.ft.com/content/cd365ae8-0f9c-4c33-8ee0-7fad89abd125`.

Gopal, A. and B. R. Koka (2010). "The Role of Contracts on Quality and Returns to Quality in Offshore Software Development Outsourcing," *Decision Sciences* 41(3): 491–516.

Gopal, A. and K. Sivaramakrishnan (2008). "Research Note—On Vendor Preferences for Contract Types in Offshore Software Projects: The Case of Fixed Price vs. Time and Materials Contracts," *Information Systems Research* 19(2): 202–220.

Gopal, A., K. Sivaramakrishnan, M. S. Krishnan, and T. Mukhopadhyay (2003). "Contracts in Offshore Software Development: An Empirical Analysis," *Management Science* 49(12): 1671–1683.

Grossman, S. J. and O. D. Hart (1983). "An Analysis of the Principal–Agent Problem," *Econometrica* 51(1): 7–45.

Holmström, B. (1979). "Moral Hazard and Observability," *Bell Journal of Economics* 10(1): 74–91.

Holmström, B. and P. Milgrom (1987). "Aggregation and Linearity in the Provision of Intertemporal Incentives," *Econometrica* 55(2): 303–328.

Howison, J., A. Wiggins, and K. Crowston (2011). "Validity Issues in the Use of Social Network Analysis with Digital Trace Data," *Journal of the Association for Information Systems* 12(12): 767–797.

Huang, H., M. Hu, R. J. Kauffman, and H. Xu (2021). "The Power of Renegotiation and Monitoring in Software Outsourcing: Substitutes or Complements?," *Information Systems Research* 32(4): 1236–1261.

Ide, E. and E. Talamàs (2025). "Artificial Intelligence in the Knowledge Economy," *Journal of Political Economy* 133(12): 3762–3800.

Laffont, J.-J. and J. Tirole (1993). *A Theory of Incentives in Procurement and Regulation*. Cambridge, MA: MIT Press.

McKinsey & Company (2023). "Meet Lilli, Our Generative AI Tool That's a Researcher, a Time Saver, and an Inspiration," New at McKinsey Blog, August 2023. https://www.mckinsey.com/about-us/new-at-mckinsey-blog/meet-lilli-our-generative-ai-tool.

McKinsey & Company (n.d.). "Rewiring the Way McKinsey Works with Lilli, Our Generative AI Platform." Accessed August 2026. https://www.mckinsey.com/capabilities/tech-and-ai/how-we-help-clients/rewiring-the-way-mckinsey-works-with-lilli.

Newell, S. and M. Marabelli (2015). "Strategic Opportunities (and Challenges) of Algorithmic Decision-Making: A Call for Action on the Long-Term Societal Effects of 'Datification'," *Journal of Strategic Information Systems* 24(1): 3–14.

Niculescu, M. F. and D. J. Wu (2014). "Economics of Free Under Perpetual Licensing: Implications for the Software Industry," *Information Systems Research* 25(1): 173–199.

Palantir Technologies (2024). *Annual Report (Form 10-K) for Fiscal Year 2023*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/0001321655/000132165524000022/pltr-20231231.htm.

Palantir Technologies (2025). *Annual Report (Form 10-K) for Fiscal Year 2024*. U.S. Securities and Exchange Commission. https://www.sec.gov/Archives/edgar/data/0001321655/000132165525000022/pltr-20241231.htm.

PwC (2026). "PwC Introduces PwC One as Part of Its Vision for AI-Enabled Professional Services," PwC Newsroom, March 19, 2026. https://www.pwc.com/us/en/about-us/newsroom/press-releases/pwc-one.html.

Salesforce (2025). "Salesforce Introduces New Flexible Agentforce Pricing to Accelerate the Digital Labor Revolution," Salesforce Newsroom, May 15, 2025. https://www.salesforce.com/news/press-releases/2025/05/15/agentforce-flexible-pricing-news/.

Salesforce (2026). "Salesforce Launches Agentforce Help Agent That Deploys in Minutes and Only Charges for Resolutions," Salesforce Newsroom, June 25, 2026. https://www.salesforce.com/news/stories/agentforce-help-agent-announcement/.

Salesforce (n.d.). "Agentforce Pricing." https://www.salesforce.com/agentforce/pricing/.

Sundararajan, A. (2004). "Nonlinear Pricing of Information Goods," *Management Science* 50(12): 1660–1673.

Susarla, A., A. Barua, and A. B. Whinston (2009). "A Transaction Cost Perspective of the 'Software-as-a-Service' Business Model," *Journal of Management Information Systems* 26(2): 205–240.

Susarla, A., A. Barua, and A. B. Whinston (2010). "Multitask Agency, Modular Architecture, and Task Disaggregation in SaaS," *Journal of Management Information Systems* 26(4): 87–117.

Susarla, A., R. Subramanyam, and P. Karhade (2010). "Contractual Provisions to Mitigate Holdup: Evidence from Information Technology Outsourcing," *Information Systems Research* 21(1): 37–55.

Taft, D. K. (2026). "PwC's AI Agents Are Now Your Consultants—Whether You're Ready or Not," *The New Stack*, March 24, 2026. `https://thenewstack.io/pwcs-ai-agents-are-now-your-consultants-whether-youre-ready-or-not/`.

Thompson, P. (2025). "AI Is Reshaping How McKinsey Makes Money," *Business Insider*, November 17, 2025. `https://www.businessinsider.com/ai-reshapes-how-mckinsey-makes-money-consultancy-025-11`.

Trentmann, N. (2026). "AI Efficiency Gains Push Accounting Firms to Reimagine Pricing," *Bloomberg Tax*, March 5, 2026. `https://news.bloombergtax.com/financial-accounting/ai-efficiency-gains-push-accounting-firms-to-reimagine-pricing`.

Wilkinson, L. (2024). "Deloitte Rolls Out In-House Generative AI Platform for Coding, Researching," *CIO Dive*, January 8, 2024. `https://www.ciodive.com/news/deloitte-generative-ai-use-platform-employees-PairD/703962/`.

Wu, D. J., M. Ding, and L. M. Hitt (2013). "IT Implementation Contract Design: Analytical and Experimental Investigation of IT Value, Learning, and Contract Structure," *Information Systems Research* 24(3): 787–801.