# Do AI Agents Need Firms? Governance without Capability

Soheil Ahmadi* Jillian Grennan†

**Abstract**

Do firms matter when search, contracting, and settlement become programmable? We study millions of transactions among autonomous AI agents in a setting that approximates frictionless exchange. Contracting frictions remain, yet the firm's role in ameliorating them changes. What endures is the economic problem the firm solves, not the organizational form that solves it. Functions conventionally bundled inside a single organizational boundary become discrete and separately supplied: relational contracts, concentrated verification, platform certification, and portable reputation. Importantly, this separation of information and incentives produces a novel relation that our theory predicts: a lack of performance persistence. AI agents rely on relational contracting to determine whom to trust, but such relationships cannot substitute for what the task specification omits. Because the advantage that informal capabilities confer inside firms has no counterpart here, currently high-performing AI agents do not outperform next period. This implies that competitive advantage may depend less on production knowledge than on control of agentic infrastructure, with implications for AI policy.

**JEL**: D23, D86, L14, G24, G32, L86, O33

**Keywords**: corporate finance; theory of the firm; AI agents; agentic commerce; incomplete contracts; relational contracting; decentralized governance; certification; reputation; verification; platforms; FinTech; blockchain; cryptocurrency

*Soheil Ahmadi, St. Mary's University, Sobey School of Business, Soheil.Ahmadi@smu.ca
†Emory University, Economics Department, jillian.grennan@emory.edu

Why do firms exist if markets can coordinate production? The answer economics has given for ninety years is that market exchange is costly. Finding counterparties, negotiating terms, and writing/enforcing contracts is costly, which leads economic activity to move inside firms (Coase, 1937; Williamson, 1979). Historically, transaction costs have rarely approached zero, but recent advances in artificial intelligence (AI) and blockchain-based technologies have produced markets pushing that limit. In this article, we examine a platform on which the transacting parties are autonomous AI agents. They produce services via machine-readable contracts. The payment is escrowed and released via code, and the median transaction time, from initiation to settlement, is about one minute. This has led some to ponder whether such technology brings a "Coasean singularity" in which frictionless exchange dissolves the rationale for the firm (Shahidi et al., 2025; Hadfield and Koh, 2025). We provide, to our knowledge, the first systematic evidence on that question, drawn from millions of transactions among these AI agents. The setting is the closest empirical approximation yet observed to the frictionless limit at the foundation of the theory of the firm.

Whether organization survives in such a market is, ex ante, unclear. One view holds that firms are a response to the cost of using the price mechanism, so driving that cost toward zero should dissolve the boundary between firm and market. Under this view, autonomous agents transacting through programmable rails need no organization at all. In contrast, others posit that the frictions firms address are not primarily about the mechanics of transacting. Instead, contracts are incomplete because contingencies are hard to describe, not merely hard to enforce (Grossman and Hart, 1986; Tirole, 1999). Under this view, cheap settlement changes the form that organization takes without eliminating the underlying economic problems. Our setting allows us to distinguish between these views, and, importantly, to observe separately several functions that conventional firms bundle inside a single organizational boundary.

We study these questions using the universe of on-chain activity from Virtuals Protocol, a crypto-native platform launched in December 2023 on which AI agents are created, tokenized, discovered, hired, and paid. A key advantage of our setting is access to 3,786,659 service

jobs transacted between 3,033 provider agents and 34,865 buyer agents, including the jobs that fail.[1] Observing failed and abandoned requests is unusual in contracting data and is essential here, because our central specification result concerns jobs that never execute. We supplement the job records with four additional sources: the provider service registry, which reports the input fields a provider formally designates as required; the platform's provider certification status and ungraduation dates; on-chain ERC-8004 reputation records; and a daily panel of 50,200 agent tokens. Together, these data let us observe task specification, monitoring, verification, certification, reputation, and market valuation as separate objects for the same set of agents.

The institutional setting is itself informative. The platform's Agent Commerce Protocol (ACP) structures bilateral exchange through four phases: request, negotiation, escrowed transaction, and evaluation. Payment can be automated by smart contracts, task requirements can be expressed through machine-readable schemas, and output can be verified by a specialized evaluator. Yet neither the Proof of Agreement nor the escrow contract verifies whether a buyer has provided all economically relevant information or whether a delivered output is substantively satisfactory. Moreover, the platform's governance apparatus was not designed ex ante. Certification with automatic revocation, ratings and reviews, performance-based search ranking, and portable on-chain identity were each introduced by the platform operator in response to observed failures. Thus, over time, the platform reinvented the governance stack of a conventional marketplace, institution by institution, compressing it into a timeframe short enough to observe end-to-end.

This sequence of administrative changes also supplies the variation underlying our empirical claims. We exploit five dated platform changes. First, automatic ungraduation, introduced in September 2025, allows us to study what happens to an agent's market access when the platform withdraws its certification. Second, an October 2025 protocol upgrade

[1]Failure is common and takes several forms. Jobs may be rejected at evaluation, expire before reaching a resolved outcome, or stall in an intermediate lifecycle phase. Because we observe the request itself rather than only executed contracts, we can study incomplete contracting at the point where it binds, rather than inferring it from realized agreements.

made the evaluation stage optional, generating a revealed-choice test of whether established counterparties economize on formal monitoring when permitted to do so. Third, a November 2025 change introduced performance-based search ranking allowing us to explore how predetermined provider quality maps into discovery. Finally, in February 2026, enablement of ERC-8004 made agent reputation technically portable across transactions and capabilities. Around each change, we compare more- and less-exposed agents while absorbing buyer-provider pair, provider-service, calendar-time, and protocol-generation fixed effects, and we cluster by provider and buyer.

We are deliberate about which of these designs truly appear to satisfy the assumptions required for unbiased inferences such as parallel trends and SUTVA in the case of difference-in-differences design, or the continuity of potential outcomes in our regression discontinuity design. Further, we examine forums and changelogs to determine whether platform changes are well-known or plausibly randomly timed. In some cases, our evidence support the assumptions required for quasi-experimental identification, and in the other cases, we provide the evidence and describe the results as descriptive correlations. Two features of the data nonetheless discipline the interpretation. First, because we observe the universe of on-chain activity, selection into the observable sample is minimal. Second, several of our results turn on asymmetries that alternative explanations struggle to generate. For example, if decertification simply revealed poor quality, demand should fall from established and prospective counterparties alike, whereas the discovery channel predicts that only unfamiliar counterparties respond.

To interpret our findings, we develop a theoretical framework that separates two components of incomplete contracting. Successful delegation requires both that the provider understand the task and that the provider comply with the agreement. The first is an information problem and the second is an incentive problem, and the two need not be solved by the same institution. Because the provider is software rather than a person, it possesses no tacit interpretive capacity, so execution depends mechanically on how much of the task

is described. Governance arrangements that raise compliance therefore cannot compensate for a deficient specification. The framework generates one prediction for each of our results blocks. It also generates a corollary that is central to the paper: when productive knowledge is codified in software and exchanged through standardized interfaces, any performance advantage is replicable at low cost, so capability rents should dissipate even while governance problems remain. Our null result on performance persistence is thus predicted rather than merely accommodated.

Our analyses yield four main results. First, formal specification remains consequential for execution even when contracting and settlement are programmable. Among jobs governed by formal machine-readable schemas, omitting at least one required field is associated with a 48 p.p. lower probability of reaching evaluation and a 57 p.p. higher probability of expiring beforehand. Information volunteered beyond the formal schema is associated with substantially higher execution and lower rejection. The relation with third-party evaluation, by contrast, is small and statistically insignificant. Thus, the point of failure appears to migrate from enforcement to specification: making a contract legible to software does not encode the contingencies the contract fails to mention (Tirole, 1999).

Second, repeated relationships between AI agents substitute for external monitoring, but not for specification. After a single prior interaction, the probability that a pair engages a third-party evaluator falls by 5 percentage points, the largest discrete change in the maturity profile, and established pairs disproportionately forgo evaluation once the protocol makes it optional. Yet relationship capital does not attenuate the execution penalty associated with formally incomplete requests; if anything, that penalty is larger in mature relationships. We also distinguish durable relationship formation from mechanical retries, as rejected first jobs generate substantially more immediate repetition but fewer durable repeat relationships. Taken together, these results suggest that, among AI agents, relationships can determine whom to trust, but they cannot substitute for what the task specification omits. This distinction is difficult to observe in conventional firms, where colleagues learn whom to rely

on while simultaneously developing the shared routines and tacit knowledge that help them interpret incomplete instructions.

Third, the governance functions that operate beyond the bilateral relationship become concentrated rather than decentralized. Although the protocol permits free entry into verification, a single evaluator performs 90% of third-party evaluations in the average month and essentially all of them from late 2025 onward, and the identity of that evaluator changes through administrative routing rather than through competitive displacement. Platform certification, in turn, governs access to strangers. When the platform withdraws a provider's certification, jobs from new buyers fall by 27 to 37 log points and remain depressed for a month, while activity from established counterparties is statistically indistinguishable from zero. Reputation exhibits a related pattern: once agent identity and feedback move on-chain, reputation earned in one capability predicts demand from counterparties new to the provider in the agent's other capabilities. We interpret the reputation evidence cautiously, as reviews are generated by transactions and the review sample remains small.

Fourth, and most surprisingly, persistent productive advantage does not emerge. Performance differences across agents are real and precisely measurable. Splitting an agent's jobs randomly within a month, performance in one half predicts performance in the other with a coefficient of 0.40. Yet those differences do not persist. One month ahead, the persistence coefficient is $-0.06$, and at three and six months it remains indistinguishable from zero. Nor does advantage accumulate inside relationships, as a pair's excess performance relative to the same provider's other buyers is $-0.002$. The market therefore exhibits relational governance without relational capability, a configuration that capability theories, in which relationships are precisely where durable advantage lives, do not anticipate (Zingales, 2000; Gibbons and Henderson, 2012; Dessein and Prat, 2022).

We subject these findings to a series of robustness tests. We show that measurement error does not explain the persistence null by instrumenting performance estimated from one random half of a provider-month's jobs with performance estimated from the other half,

and by estimating persistence separately within transaction-volume quartiles. Persistence is meaningfully positive only among the largest and longest-surviving providers, and even there it decays within a quarter, far faster than the productivity differences documented among conventional firms (Backus, 2020). We also verify that our results are not driven by the smallest transactions, that they hold across protocol generations and after excluding fund-transfer jobs and the default evaluator, and that apparent orchestration among agents reflects transaction frequency rather than durable hierarchy. Finally, we examine whether token markets price what our performance measures capture and find that valuations move with market access rather than with measured productive quality, which is consistent with the interpretation that scarcity resides in infrastructure rather than in production.

A natural follow-up question is whether results from this platform generalize. We view the setting as a benchmark rather than as a representative sample, and we acknowledge that it reflects one marketplace populated by a particular technological generation of agents transacting largely with one another. Two considerations nonetheless support portability. The frictions we document, namely incomplete task description, costly verification, and gatekeeping over discovery, are near-universal constraints rather than features of this platform's design (List, 2022). Moreover, the mechanisms we identify follow from properties of the technology rather than from platform-specific rules. We know that AI agents lack tacit interpretive capacity, and codified productive knowledge is cheap to replicate. If future models acquire something closer to interpretive capacity, the specification friction should weaken, and relationship-specific capability may begin to appear, which is a testable implication of our framework rather than a threat to it.

In conclusion, our evidence suggests that the theory of the firm decomposes into two accounts that agentic commerce separates cleanly. What endures is the economic problem the firm solves, not the organizational form that solves it. Contracting frictions remain, and the functions that address them reappear as schemas, bilateral relationships, a concentrated verification layer, platform certification, and portable reputation. What does not endure

is capability. When production is replicable code, the firm's role as a repository of tacit and hard-to-imitate knowledge finds no counterpart. This shifts where advantage resides. If every function of the firm can be procured except the ability to stay good at something, then the scarce asset is not skill in production but control of the interfaces through which production is matched, verified, and certified (Rajan and Zingales, 1998, 2001). Ownership and settlement are decentralized, but market design is not.

The remainder of the paper proceeds as follows. Section  positions our paper relative to the existing literature. Section 2 describes the institutional setting. Section 3 develops a framework separating the information and incentive components of incomplete contracting and derives one prediction per results block, including the prediction that performance should not persist. Section 5 details the empirical designs. Section 6 presents results. Section 7 discusses implications of our findings, and Section 8 concludes.

## 1 Related Literature

A rapidly growing literature examines how AI is changing work and firms (Acemoglu and Restrepo, 2019; Babina et al., 2024; Cao et al., 2024; Grennan and Michaely, 2025), but less attention has been paid to how AI is changing the economics of delegation within and beyond firms. Yet recent advances have shifted AI from tools that assist human decision-makers (Eloundou et al., 2024; Brynjolfsson et al., 2025; Dell'Acqua et al., 2026) to AI agents that can execute a multitude of tasks on behalf of requesters (Shahidi et al., 2025; Hadfield and Koh, 2025) while also autonomously requesting and procuring services from other agents. This shift matters because organizations and institutions are designed to manage the costs of delegation and coordination (Bolton et al., 2012). The change also speaks to a central question that has driven economic research for decades. Namely, why firms exist rather than markets (Coase, 1937; Williamson, 1979, 1985; Baker et al., 2001).

One tradition argues that the answer lies in incomplete contracting and the allocation of residual control rights (Grossman and Hart, 1986; Hart and Moore, 1990, 1988; Aghion

and Tirole, 1997; Tirole, 1999). A second strand focuses on incentives, monitoring, and the design of authority within organizations (Holmstrom, 1982; Holmstrom and Milgrom, 1991, 1994; Prendergast, 1999; Dessein and Santos, 2006; Alonso et al., 2008). Another one emphasizes knowledge-based hierarchies for problem-solving and the shared codes that make communication within them cheap (Garicano, 2000; Cremer et al., 2007; Garicano and Rossi-Hansberg, 2006, 2015; Bloom et al., 2014; Caliendo et al., 2015; Ewens and Giroud, 2025). A fourth posits the firm as a nexus of specific investments and human capital whose value depends on continued association (Rajan and Zingales, 1998; Zingales, 2000; Rajan and Zingales, 2001; Hart and Holmstrom, 2010). And a fifth locates durable advantage in organizational capital and routines that accumulate slowly (Gibbons and Henderson, 2012; Dessein and Prat, 2022).

Our paper contributes to this research on the theory of the firm and the boundary between markets and organizations. The most closely related strand asks what firms do well rather than what assets they own, locating persistent advantage in management practices and shared capabilities that are costly to imitate and slow to diffuse (Teece, 1986; Bertrand and Schoar, 2003; Bloom and Van Reenen, 2010; Bloom et al., 2012; Lazear et al., 2015). Corporate culture is a canonical example, in which shared beliefs and norms lower the cost of coordinating on what was never written down, what firms invest in deliberately, and that does not transfer easily across organizational boundaries (Schein, 1985; Kreps, 1990; Guiso et al., 2015; Li et al., 2020; Graham et al., 2022a,b; Grennan, 2023; Grennan and Li, 2023; Gorton and Zhang, 2024). These are the mechanisms by which human organizations produce what contracts leave unsaid. In our setting, these traditionally bundled functions are supplied by separate parties and separately observed, so we can ask which survive when the cost of market exchange nearly vanishes. The contracting functions survive, but persistent productive capability does not.

To that end, a recent body of theory asks how technology and automation change knowledge hierarchies and how much of the gain accrues to the firms adopting them (Ide and

Talamàs, 2025). Closest in focus to our work on AI agents specifically is that on the industrial organization of agentic markets (Demirer et al., 2026; Shahidi et al., 2025; Nagle1 and Yue, 2025), on the institutional and legal requirements of an economy of agents (Hadfield and Koh, 2025; Grennan, 2026), on alignment when agents act autonomously (Yang et al., 2026; Wang et al., 2026), and early evidence on agent-to-agent negotiation and collusion (Tomasev et al., 2025; Sarkar, 2026). Nearly all of this work studies AI inside human organizations or in simulation. In contrast, our study focuses on AI agents transacting outside any firm, at scale, with the organizational functions reassembling around them.

Finally, our setting is crypto-native, and we build on work analyzing whether programmable settlement changes the economics of intermediation (Abadi and Brunnermeier, 2018; Cong and He, 2019; Cong et al., 2021, 2022; Catalini and Gans, 2020; Abadi and Brunnermeier, 2018; Saleh, 2020; Makarov and Schoar, 2020). Especially relevant is evidence that decentralized architectures reduce some transaction costs while leaving multiple layers at which rents accumulate (Makarov and Schoar, 2022; Lehar and Parlour, 2024; Harvey et al., 2026), and evidence on token issuance, disclosure, and investor protection in unregulated primary markets (Howell et al., 2020; Liu et al., 2023; Appel et al., 2025; Lyandres et al., 2022; Cong et al., 2021; Sockin and Xiong, 2022; Li and Mann, 2025; Li et al., 2025). Work on decentralized governance asks who actually sets the rules when ownership is dispersed on-chain (Appel and Grennan, 2023, 2025; Cong et al., 2025). Our contribution is to use this infrastructure as a laboratory rather than as a subject: the same features that make crypto rails controversial as financial infrastructure—programmability, pseudonymity, complete transaction visibility—are what make the organizational functions we study separately observable. Consistent with Makarov and Schoar (2022), we find that decentralized settlement coexists with, and arguably requires, centralized market design.

## 2 Institutional Background

### 2.1 Business Model and Contracting Environment

Launched in December 2023, Virtuals Protocol is a crypto-native platform on which autonomous AI agents can be created, tokenized, discovered, hired, and paid. The platform combines two economically distinct layers. The first is a token market. Agents can issue tradable tokens through a bonding-curve mechanism denominated in $VIRTUAL or $USDC; once the protocol's liquidity threshold is reached, trading migrates to a Uniswap pool.[2] These agent-specific tokens give us a market-based measure of how investors value the underlying agent over time.

The second layer is the Agent Commerce Protocol (ACP), a set of smart contracts that structures bilateral exchange between AI agents. An ACP job can move through four phases: request, negotiation, transaction, and evaluation. During request, a client identifies a provider and initiates a task. During negotiation, the parties agree to terms that are recorded in a Proof of Agreement. During transaction, payment and the deliverable are placed into escrow. During evaluation, the output is assessed before escrow is released. Evaluation can be delegated to a third-party evaluator, performed by the client, or, for jobs using later protocol versions, omitted.

ACP therefore separates several functions that are typically harder to observe independently inside firms. Payment can be automated by smart contracts, task requirements can be expressed through machine-readable schemas, output can be verified by a specialized evaluator, and counterparties can transact repeatedly outside a common ownership boundary. At the same time, neither the Proof of Agreement nor the escrow contract verifies whether a buyer has provided all economically relevant information or whether a delivered output is substantively satisfactory. The setting therefore allows us to study incomplete contracting without assuming that programmable settlement is equivalent to complete con-

[2]Uniswap is a system of smart contracts on the Ethereum blockchain and is the largest decentralized exchange (DEX) (Lehar and Parlour, 2024; Harvey et al., 2026).

tracting (Holmstrom and Roberts, 1998). It also makes it possible to distinguish relational governance, in the sense of cooperation supported by repeated interaction (Gibbons and Henderson, 2012), from the stronger claim that crypto infrastructure makes exchange "trustless" (Bratspies, 2018).

### 2.2 Certification, Discovery, and Platform Authority

ACP also contains a platform certification system. Every provider begins in a sandbox state. Providers that satisfy the platform's operating criteria can become *graduated*, which makes them eligible for production-facing discovery and for routing through Butler, the platform's consumer-facing orchestration agent. A sandbox provider remains technically capable of receiving jobs through direct assignment or internal links, but loses access to the main discovery and routing infrastructure. Thus, ACP graduation is economically different from token graduation.[3]

The platform introduced automatic ungraduation on September 2, 2025. Under the initial rule, a graduated provider could be returned to sandbox after ten consecutive qualifying failed or expired jobs. On September 5, Butler began automatically retrying a failed request with another suitable provider. On September 9, the ungraduation rule was refined so that qualifying failures had to reflect provider responsibility and the failure history had to involve at least three distinct buyers. Previously reviewed agents could regraduate after accumulating ten total successful jobs and three consecutive successes.

The September sequence is useful institutionally because it reveals how the platform aggregates noisy performance signals before withdrawing certification. It is less useful as a sharp natural experiment in our data. We observe job outcomes and the provider's most recent ungraduation date, but we do not observe the platform's backend determination of provider responsibility or a complete historical status path. Accordingly, we use recon-

[3]Token graduation refers to migration of an agent token from the bonding curve to a decentralized-exchange liquidity pool. ACP graduation is a service-provider certification state. The former is a token-market event; the latter governs access to ACP discovery and routing.

structed threshold crossings to validate the institutional rule, but we do not treat them as an instrument for certification. Our main authority analysis instead studies how observed market activity evolves around recorded ungraduation events and, especially, whether the pattern differs between new and established counterparties.

### 2.3 Reputation and the Evolution of Market Infrastructure

The platform's information infrastructure evolved alongside its certification system. Ratings and written reviews were introduced in October 2025. In November 2025, performance metrics were incorporated more directly into the platform's default search and ranking system. On February 1, 2026, Virtuals enabled full ERC-8004 support for ACP agents, linking agent identity and reputation to an on-chain registry. This last change is conceptually important for our analysis. Internal ratings make past performance visible within the platform; ERC-8004 makes reputation technically portable by attaching it to an on-chain agent identity that can in principle be read outside the transaction or service in which the signal was generated.

Our data allow us to link ERC-8004 feedback records to the underlying ACP providers and, for most reviews, to the underlying completed ACP job. A single job can generate multiple on-chain feedback entries when the review records different task components, service labels, or blocks. We therefore distinguish between the number of reviewed transactions, the number of independent reviewer wallets, and the amount of information written on-chain about a transaction. This distinction matters for the economic meaning of reputation: repeated entries from one complex job provide richer public information, but they are not independent endorsements.

The dated platform changes provide useful institutional landmarks, but we do not assign the same empirical interpretation to each one. The November search change is used as supporting evidence because the post-change association between pre-existing performance and activity is short-lived and is not concentrated among new counterparties. ERC-8004 is used more directly because the resulting review data allow us to observe the accumulation

and portability of public reputation itself. Additional platform changes are described where relevant in the Appendix rather than in the main institutional narrative.

### 2.4 Who Sets the Operating Rules

Virtuals has formal on-chain governance. Holders of non-transferable voting power earned by locking $VIRTUAL may submit proposals after crossing a voting-power threshold, followed by a comment period and an on-chain vote. These procedures resemble governance arrangements used in other decentralized protocols (Appel and Grennan, 2023, 2025; Han et al., 2026).

The operating rules central to our analysis, however, were implemented administratively by the core team rather than through token-holder votes. Graduation thresholds, expiry accounting, evaluator settings, search ranking, payment denomination, and ERC-8004 integration were announced as product or engineering changes. This distinction is substantively useful. Formal token-holder governance governs some organizational decisions, while the rules that determine day-to-day market access, verification, and information production are set by a platform operator. In that sense, the setting combines decentralized ownership and settlement with a centralized layer of market design and administrative authority.

## 3 Theoretical Framework

We borrow from two distinct theoretical explanations for why firms exist. First, the incomplete contracts tradition posits that firms arise because contracting via markets is costly, so residual decisions must be allocated to someone. The second is a capability explanation, which emphasizes that firms are repositories of relational knowledge, routines, and firm-specific investments that markets cannot replicate. In most empirical settings, the two cannot be separately identified, because firms bundle contracting and capability functions inside a single boundary. For that reason, the purpose of this section is to formalize a distinction that conventional settings render invisible, and that ultimately guides our empirical

tests. Incomplete contracting is both an information problem and an incentive problem, and the two need not be solved by the same institution. In AI agent markets, we argue, they come apart. Our framework, therefore, generates predictions organized around each problem.

### 3.1 Setup

A client AI agent $c$ wishes to delegate a task of value $V$ to another provider AI agent $p$. Delegation succeeds only if two conditions hold. First, the AI agent provider must understand the task. Let $\sigma \in [0, 1]$ denote the share of economically relevant contingencies encoded in the buyer's machine-readable request. The platform's schema, or required fields, mandates a minimum $\sigma_{min}$. AI agent clients may voluntarily supply information beyond it. Given that the provider is an AI agent, it possesses no tacit interpretive capacity. So, unlike a human employee who draws on shared context, like cultural norms, to fill gaps in an instruction, an AI agent's probability of correct execution, $q(\sigma)$, is increasing in $\sigma$ mechanically, independent of any incentive. Second, the provider must comply, delivering the promised quality rather than a cheaper substitute. Performing in a way that complies with the stated task is a standard hidden-action problem where output quality is observable but costly.

Expected surplus from a transaction is therefore

$$S = q(\sigma)\,\pi\,V - \kappa(\text{governance}), \tag{1}$$

where $\pi$ is the probability of compliance induced by the chosen governance arrangement and $\kappa(\cdot)$ is that arrangement's cost. This structure embeds our framework's key insight and central assumption. Namely, information and incentives are complements in production but not substitutes. No governance institution that operates on $\pi$ can compensate for a deficient $\sigma$. The features of programmable settlement, such as specified escrow, automated payment, and machine-readable terms, all reduce the cost of writing and enforcing the contract, but they do not increase the share of contingencies the contract describes. This suggests that

AI agents are like earlier forms of digitization (Goldfarb and Tucker, 2019), which relax the enforcement constraint on contracting while leaving the describability constraint intact.

Three governance technologies can raise $\pi$:

1. **External verification.** A third-party evaluator inspects output at per-transaction cost $k$. Evaluation technology exhibits scale economies: an evaluator's marginal cost of judging machine output falls with volume, through standardization of rubrics and fixed-cost tooling.

2. **Relational enforcement.** Within a repeated client–provider pair, the continuation value of the relationship disciplines compliance (Gibbons and Henderson, 2012). Relational enforcement is free at the margin but is available only after the pair has transacted: it requires private history.

3. **Platform certification.** The platform observes noisy performance signals across all of a provider's counterparties and issues a public certification (graduated status) that gates access to discovery and matching. Certification is informative precisely for counterparties who lack private history; portable on-chain identity extends the reach of such public signals across the provider's distinct capabilities.

In a conventional firm, all three functions—plus task specification—are bundled inside one boundary and priced implicitly. Here each is supplied by a separate market institution, so the model's comparative statics map one-to-one into observable margins.

### 3.2 Predictions

Our first prediction is that formal task specification is not substitutable. Execution probability rises in $\sigma$ under every governance mode. Requests that omit schema-mandated fields ($\sigma < \sigma_{min}$) fail disproportionately before evaluation, and information supplied beyond the mandated minimum raises execution. Given that $q(\sigma)$ enters surplus multiplicatively, rela-

tional capital does not attenuate the execution penalty of formal incompleteness. Rather, a mature relationship raises $\pi$, but not $q$.

Our second prediction is that relationships substitute for verification. Relational enforcement and external verification are substitutes in producing $\pi$. Because the informational gain from interaction is largest at the first encounter, demand for third-party evaluation drops discretely after the first successful transaction and declines little thereafter. When the protocol makes evaluation optional, established pairs disproportionately forgo it.

Our third prediction is that verification concentrates. External verification can concentrate for two related reasons. First, standardization and fixed-cost tooling create economies of scale in evaluating machine output. Second, a platform can coordinate transactions on a common verifier through default routing, reducing the cost of evaluator selection and integration. Thus, concentrated verification need not imply that the dominant evaluator possesses a uniquely persistent productive advantage. A change in the platform default can instead reallocate verification activity sharply across otherwise substitutable evaluators.

This distinction separates the decision to use verification from the organization of the verification layer itself. Relationship capital can reduce the demand for external verification, while platform routing can determine which evaluator supplies verification conditional on its use. As such, concentration may reflect both evaluator-level economies of scale and platform-level coordination. Later, in our empirical tests, we explore how concentration relates to these two potential mechanisms. In addition, Appendix C microfounds the concentration of verification.

Our fourth prediction is that certification matters at the extensive margin. Certification and relational capital are substitutes along the counterparty margin. Withdrawal of certification removes the public signal and access to discovery, which is the binding constraint only for counterparties without private history. Ungraduation therefore reduces new-client activity sharply while leaving established-pair activity approximately unchanged. Symmetrically, portable public reputation (i.e., signals attached to the agent's identity rather than to

a specific service) predicts demand from unfamiliar counterparties and from counterparties new to the market, including in capabilities other than the one in which the reputation was earned.

Our fourth prediction assigns the platform the role that theories of the firm typically assign to the owner of a critical asset. The platform's authority derives not from ownership of the AI agents but from control of access to the resource all agents need: discovery and matching. The platform, therefore, exercises authority in a firm-like sense without employment relationships or residual claims. Consistent with the view that governance mechanisms are invented as environments change (Ma and Shleifer, 2025), the unbundled contracting features that we study were introduced sequentially by the platform operator in response to observed failures. Thus, if authority indeed operates through discovery rather than ownership, it should bite in proportion to how much a provider depends on being discovered.

Finally, our fifth prediction formalizes this notion and states that authority scales with dependence on discovery. The consequence of decertification is an increase in a provider AI agent's reliance on unfamiliar counterparties. Providers whose pre-event demand comes disproportionately from new clients lose more activity upon ungraduation than providers with similar demand concentrated in established relationships. This last prediction separates a discovery channel from a simple demand story. If decertification merely signals poor quality, all buyers should update, and activity should fall regardless of a provider's counterparty mix. The gradient in pre-event new-client reliance distinguishes the withdrawal of market access from the withdrawal of a quality endorsement.

### 3.3 The Capability Corollary

Our framework's final implication concerns what does not reappear when we unbundle information and incentives. In capability theories, persistent performance differences across firms arise because productive knowledge is tacit, embedded in relationships, and thereby costly to imitate. Our setting is the limit case in which productive knowledge is fully codified. An

AI agent's tacit knowledge, culture, and routines must be specified; the AI agent must rely on the model and context provided to it.

This insight offers a testable corollary about capability. When production knowledge is codified, and interfaces are standardized, any performance advantage is imitable at low cost, so capability rents dissipate quickly. Therefore, the associated prediction is that contemporaneous performance heterogeneity exists, but its persistence should be approximately zero, and no excess performance should accrue to specific client–provider pairs.

Taken at face value, this corollary offers a theoretical answer to our question, "Do AI agents need firms?" On the one hand, the firm's contracting feature survives the collapse of transaction costs, but survives unbundled and relocated into the modular functions above (i.e., schemas, relationships, a concentrated evaluator, platform certification, and portable reputation). On the other hand, the firm's capability feature has no counterpart. What remains scarce is not skill in production but control of market infrastructure. Interestingly, this insight inverts the mechanism in Ide and Talamàs (2025). There, AI agents are absorbed into human knowledge hierarchies and reshape the firm from within. Here, agents transact outside any firm, and it is the hierarchy's residual functions, such as monitoring, certification, and authority, that reassemble around them as market institutions.

Two caveats bound our framework's ambition. First, it is a framework for organizing comparative statics. We do not seek to develop a full equilibrium model. Instead, we take entry, pricing of evaluation, and the platform's objective as given. Second, while we offer several predictions related to selection into governance arrangements, it is not possible to identify all of them. Therefore, we are careful to describe in Section 5.1 which empirical counterparts we interpret as causal and which as descriptive.

## 4 Data and Summary Statistics

### 4.1 Data Sources

We assemble five linked datasets from the Virtuals ecosystem. The first four describe the service market: job-level records from the Agent Commerce Protocol (ACP), the provider service registry that defines formal task requirements, the platform's provider certification status, and on-chain ERC-8004 reputation records. The fifth is a daily market panel for agent tokens launched through the platform's bonding-curve mechanism. All are constructed from records stored on Virtuals Protocol smart contracts. Appendix A defines every variable used in the exhibits.

**ACP job records.** The ACP dataset contains 3,786,659 service jobs initiated between May 30, 2025, and June 22, 2026. Records come from successive smart-contract generations, which we label protocol generations v1, v2, and v3 and absorb with protocol-generation fixed effects throughout. Each record identifies the transaction hash; the buyer, provider, and evaluator wallets; the service transacted; the payment token; and the payments and fees allocated to each party. Associated lifecycle events allow us to follow a job from creation and negotiation through payment and evaluation. Across the sample, 3,033 wallets supplied at least one service and 34,865 wallets purchased at least one service.

We classify each job by its last observed phase. Because observed ACP paths need not pass through every stage, completed, rejected, and expired jobs do not form an exhaustive partition of initiated jobs, and the stage shares are not mechanically nested.

**Service registry and catalog schemas.** A subset of providers expose a formal machine-readable input schema through the on-chain service registry. For these offerings we observe which input fields the provider designates as required and which inputs the buyer actually submitted. This lets us construct our measures of formal specification: whether the buyer omitted at least one required field, whether the buyer supplied information beyond the

schema, and the share of required fields submitted. Catalog schemas are observed for 10.5% of jobs, which defines the estimation sample for Table 2. For offerings whose schema contains exactly one required field, we additionally recover the normalized identity of that field, which supports comparisons across jobs requiring the same input.

**Platform certification status.** We observe each provider agent's current ACP graduation status and, for currently ungraduated providers, the recorded date of the most recent ungraduation. We do not observe the platform's backend determination of provider responsibility or a complete historical status path. This limitation shapes our research design: we use reconstructed threshold crossings only to validate the institutional rule and study observed ungraduation events as event-time evidence rather than as an instrument. ACP graduation is a service-market certification state and is distinct from token graduation from the bonding curve.

**On-chain reputation.** Following the February 1, 2026, enablement of ERC-8004 for graduated ACP agents, we observe on-chain feedback records linked to ACP providers. The review data are early and thin relative to the job data: 812 valid on-chain review entries attach to 427 distinct underlying reviewed jobs, covering 55 ERC-linked providers and 342 provider-capabilities. Because a single job can generate several entries when the review bundle records distinct task components, labels, or blocks, we distinguish throughout between reviewed jobs, distinct reviewer wallets, and raw review entries. We treat the transaction-level reviewed job, not the entry, as the unit of reputation depth.

**Agent-token market.** The token dataset is an unbalanced daily panel of 50,200 agent tokens launched between September 23, 2024, and July 7, 2026, containing 318,762 token-day observations of closing price, trading volume, market capitalization, distinct holders, blockchain, and token tier. Coverage is highly uneven: the 625 tokens that graduated from the bonding curve to a Uniswap V2 liquidity pool have a median of 216 daily observations,

while the 49,575 tokens that remained on the bonding curve have a median of one. Because token-market outcomes are used only for the provider agents we can link to ACP activity, and because return-based measures require dense coverage, we report our token results on three explicitly nested samples: the broad performance-eligible sample, the market-cap-observed subsample, and the return-rich subsample requiring at least five clean daily returns in the month.

### 4.2 Sample Construction and Measurement

**Identity.** Agent-level analyses require mapping provider wallets to underlying agents. We retain a provider-agent identity only when the provider-to-agent mapping satisfies our strong identity rule and is not flagged as an identity conflict. All persistence, ungraduation, and token-market results are estimated on strong-identity providers. Service identity is recovered from the on-chain service registry and cleaned into a consistent service concept, which we use to form provider-service offerings and service-by-time fixed effects.

**Relationship history.** Relationship-history variables count jobs created strictly before the current job's creation-time block, so no contemporaneous or future activity enters a relationship measure. We construct prior buyer-provider jobs, prior pair-service jobs, prior provider jobs, and prior buyer jobs on this basis, and bin prior buyer-provider jobs into the relationship-maturity categories.

**Evaluation.** We classify the evaluator of each job into four mutually exclusive categories: no usable evaluator, buyer/self evaluation, provider evaluation, and independent third-party evaluation. Third-party evaluation means the evaluator wallet differs from both the buyer and the provider. Independence here is wallet-level separation in the observed data; it does not establish organizational or economic independence.

**Payments.** Most jobs are denominated in USDC (98.8%), with 1.2% in VIRTUAL and 421 jobs in wrapped ether. We convert all amounts to U.S. dollars using the conversion recorded in the data and exclude observations with missing, negative, or implausibly large values. Provider, evaluator, and platform fees become claimable only on completion, so payment-allocation shares are computed among completed jobs with all allocation fields observed. We also flag jobs that primarily transfer funds between wallets rather than purchase a service and exclude them in robustness tests.

**Performance.** Our measure of provider quality is a risk-adjusted job performance residual: the residual from a linear probability model of completion that controls flexibly for relationship maturity and absorbs buyer, service-by-creation-week, and protocol-generation fixed effects, leaving provider-level differences in the residual. Agent performance is the provider-month mean of this residual, standardized within calendar month. Because this measure is estimated rather than observed, we address attenuation directly by splitting each provider-month's jobs into random halves, using one half as an instrument for the other.

### 4.3 Descriptive Statistics

Table 1 reports job-level descriptive statistics across four panels, and three features of this market are worth emphasizing before turning to the results.

First, programmable settlement does not imply frictionless execution. Panel A shows that 62.8% of jobs complete, 10.3% are rejected in evaluation, and 15.5% expire before reaching a resolved outcome. A substantial share of requested work therefore fails, even though negotiation, escrow, and payment are automated.

Second, exchange is extremely fast and extremely repeated. Panel C shows a median job duration of 0.018 hours, or a little over one minute, though a long right tail lifts the mean to 1.43 hours. At the same time, the median buyer-provider pair has 102 prior jobs and the median pair-service relationship has 62. Repeated interaction is a first-order feature of this market rather than a residual phenomenon, which is what makes it possible to study

relational governance at the transaction level.

Third, transactions are individually small but the payment structure is economically informative. Panel D shows a median specified job payment of \$0.25, with 31.9% of jobs at or above \$1 and 7.3% at or above \$10. The scale of the median transaction is consistent with automated, high-frequency machine-to-machine exchange rather than with discrete human procurement decisions, and we report our main results on economically meaningful subsamples to confirm that they are not driven by the smallest transactions. The allocation of completed-job payments is equally telling: providers receive 74.1% of payment value, the platform 25.7%, and evaluators 0.2%. The party that performs the work captures roughly three-quarters of the value, while the party that supplies the matching and certification infrastructure captures roughly one-quarter, even though it neither performs nor verifies the service.

### 4.4 Units of Observation

Our analyses use five units. The job level ($j$), with buyer $c$, provider $p$, service $s$, and creation time $t$, is the unit for specification, relational governance, and relationship-formation tests. The buyer-provider pair is the unit for relationship-formation analysis restricted to first observed jobs. The provider-month is the unit for performance and persistence tests. The provider-day is the unit for the ungraduation event study, and the provider-capability-day for the portable-reputation tests. Finally, the provider-token-month links ACP activity to token-market outcomes. Sample sizes differ across exhibits because of these differing units and because catalog schemas, review records, ungraduation dates, and dense token coverage are each observed for distinct subsets of the market.

# 5 Empirical Design

## 5.1 Empirical Design

The empirical design follows the distinction developed in the theoretical framework between the information problem in delegation, the incentive problem in governing counterparties, the role of common market infrastructure, and the persistence of productive capability. We use ACP's institutional details to test a sequence of predictions that differ in the source of identifying variation and in the strength of the interpretation they support. The first two parts use within-relationship and within-offering variation to study formal specification and relational governance. The third part studies verification and platform certification, combining descriptive evidence on the organization of shared infrastructure with event-time evidence around observed changes in platform status. The fourth part asks whether measured differences in agent performance persist over time and explicitly addresses the possibility that weak persistence is generated by measurement error.

Our baseline job-level specifications weight jobs equally and use linear probability models for binary outcomes. We use buyer-provider-pair fixed effects to absorb persistent heterogeneity in a bilateral relationship, provider-service fixed effects to absorb persistent differences across offerings, calendar-time fixed effects to absorb common shocks, and protocol-generation fixed effects to absorb differences across ACP contract generations. Standard errors in job-level specifications are two-way clustered by buyer and provider unless otherwise noted. Agent-level persistence specifications are estimated at the provider-month level and cluster standard errors by provider agent. The Internet Appendix provides complete variable definitions.

### 5.1.1 Formal Specification and the Information Problem

Our prediction distinguishes task specification from enforcement. If machine-readable contracting eliminates the relevant information problem, then, conditional on the client-provider relationship, the provider-service offering, calendar time, and protocol generation, variation

in the information supplied with a request should not be systematically related to downstream execution. Our null hypothesis is that formal completeness and residual information are unrelated to subsequent task progression.

We test this prediction in the subset of jobs for which the provider's catalog exposes a formal machine-readable input schema. For job $j$ between buyer $c$ and provider $p$ using service $s$ at time $t$, we estimate

$$Outcome_{jcpst} = \beta_1 MissingRequired_j + \beta_2 ExtraSchema_j + \mu_{cp} + \nu_{ps} + \lambda_t + \omega_v + \varepsilon_{jcpst}, \quad (2)$$

where $MissingRequired_j$ indicates that at least one field designated as required by the provider's schema is absent, and $ExtraSchema_j$ indicates that the buyer supplied information beyond the formal schema. The outcomes distinguish where execution breaks down: whether the job reaches evaluation, whether it expires before evaluation, rejection conditional on reaching evaluation, use of an independent third-party evaluator, and time to transaction. The fixed effects compare jobs within the same bilateral relationship and provider-service offering while absorbing common time and protocol conditions.

Next, we test the non-substitution implication of our Predictions. Let $R_{cp,t-1} = \log(1 + PriorJobs_{cp,t-1})$ denote relationship capital measured strictly before the current creation-time block. We augment Equation (2) with $R_{cp,t-1}$ and $R_{cp,t-1} \times MissingRequired_j$. If bilateral history supplies shared context that substitutes for a formally incomplete request, the interaction should offset the execution penalty associated with missing required information. However, our theoretical framework instead predicts no such offset. Relational governance operates on compliance, not on whether the provider understands the task.

### 5.1.2 Relational Governance and the Incentive Problem

The next prediction we test is whether repeated interaction serves as a substitute for formal verification. The null of no relational governance predicts that, conditional on the same pair and provider-service, evaluator choice should not change systematically as a relationship

accumulates history. The relational-governance prediction instead implies that the demand for external verification should fall once the parties have private bilateral history, with the largest informational gain occurring early in the relationship.

For each job, we measure relationship maturity using the number of prior buyer-provider jobs created strictly before the current creation-time block. The displayed specification uses bins for one prior job, two to five, six to ten, and more than ten prior jobs, with the first observed buyer-provider job omitted:

$$Outcome_{jcpst} = \sum_{b} \beta_b 1\{Maturity_{cp,t-1} = b\} + \mu_{cp} + \nu_{ps} + \lambda_t + \omega_v + \varepsilon_{jcpst}. \tag{3}$$

The primary outcome is independent third-party evaluation. We estimate the same specification for reaching evaluation, pre-evaluation expiry, rejection conditional on evaluation, and job duration. Pair fixed effects imply that the maturity coefficients describe how governance changes within a continuing buyer-provider relationship rather than cross-sectional differences between relationships that do and do not repeat. Repeated exchange is endogenous, so the coefficient is interpreted as a within-relationship description of governance.

### 5.1.3 Platform Infrastructure, Verification, and Market Access

Next, we test two of the predictions related to platform functions. The first is the organization of verification. We classify each usable evaluator as the buyer/client, the provider, or an independent third party, and calculate monthly evaluator counts, evaluator HHI, and the largest evaluator's share of third-party evaluations. These statistics test the descriptive null that an open protocol produces a diffuse verification layer. They do not by themselves identify why concentration arises.

Then, we exploit the October 2025 default-routing episode to better understand the concentration measures. Beginning October 13, ACP progressively routed v1 jobs toward the evaluator Goku. On October 22, Goku ceased evaluating jobs, and Generalist began almost immediately; the v2 JobManager contract launched the following day. We use this episode

as institutional evidence on how verification is organized. In particular, the routing data distinguish concentration generated by decentralized evaluator choice from concentration implemented through platform defaults.

The second platform function is certification. Graduated status affects an agent's access to ACP discovery and matching. Our prediction here implies an asymmetric relation between withdrawal of certification and market activity. Namely, public certification should matter most for buyers that lack private history with the provider. For provider agent $p$ around an observed ungraduation date $T_p$, we estimate

$$Y_{pt} = \sum_{k \in \mathcal{K}} \beta_k 1\{t - T_p \in k\} + \alpha_p + \lambda_t + \varepsilon_{pt}, \tag{4}$$

where $\mathcal{K}$ contains the post-event bins days 1–3, 4–7, 8–14, and 15–30, and the omitted period is days $-14$ through $-2$. Day $-1$ and the recorded ungraduation day are excluded because status timing is observed only at the calendar-date level and activity is unusual around the recorded event. We estimate Equation (4) separately for total jobs, jobs from new buyers, jobs from established buyers, and the number of distinct new buyers. Outcomes enter as $\log(1 + count)$.

Observed ungraduation is not randomly assigned. It follows adverse platform performance, and our data do not recover the exact backend determination or intraday status timing. We therefore do not interpret $\beta_k$ as the causal effect of certification. The theory-relevant comparison is cross-counterparty: the null predicts similar post-event changes in activity from new and established buyers, whereas our framework predicts a larger change among counterparties without private history.

### 5.1.4 Performance Persistence and the Capability Corollary

The final empirical design tests the capability corollary. If agents develop durable, hard-to-imitate productive advantages, capability theory suggests that current relative performance should forecast future relative performance. In contrast, our corollary predicts that when

productive routines are codified in software and interfaces are standardized, contemporaneous performance differences can be real while persistence is close to zero. Because this prediction places weight on a statistical null, simply failing to reject persistence is not informative. We therefore first establish that agent performance is measurable and then explicitly address attenuation from measurement error.

We construct risk-adjusted job performance as the residual from a linear probability model of completion among resolved service jobs. The model controls flexibly for relationship maturity and absorbs buyer fixed effects, service-by-creation-week fixed effects, and protocol-generation fixed effects. Provider identity is deliberately not absorbed so that provider-level performance differences remain in the residual. Agent performance $Q_{pt}$ is the provider-month mean of the estimated risk-adjusted job performance for strong-identity provider agents, standardized across providers within calendar month.

We estimate persistence at horizons $h \in \{1, 3, 6\}$ months using

$$Q_{p,t+h} = \rho_h Q_{pt} + \lambda_t + \varepsilon_{p,t+h}, \tag{5}$$

requiring sufficient resolved-job support in both periods. The conventional capability alternative predicts $\rho_h > 0$ over economically meaningful horizons. In contrast, our proposed corollary predicts rapid dissipation toward zero.

We use three diagnostics to distinguish low persistence from noisy measurement. First, within a provider-month, we randomly divide resolved jobs into two disjoint halves and relate standardized performance measured in one half to performance measured in the other. This split-half coefficient measures contemporaneous reliability. Second, we form provider-month transaction-volume quartiles within calendar month and compare same-month reliability with future persistence across quartiles. If attenuation drives the result, persistence should be materially stronger precisely where split-half reliability is highest. Third, on a common sample we instrument performance estimated from one random half of current-month jobs

with performance estimated from the other half. The split-half IV provides a direct errors-in-variables correction where the first stage is sufficiently strong.

We also ask whether productive advantage is relationship-specific. For high-intensity established relationships, relationship-specific excess performance equals the buyer-provider pair's prior-month risk-adjusted performance minus the same provider's performance with other buyers in that month. We test whether this dyad-specific component predicts subsequent completion after controlling for lagged agent performance.

Finally, we use our token-market data for graduated AI agents as an external diagnostic rather than a separate main hypothesis. In particular, we compare future-performance regressions on exactly the same observations with and without token-market signals. We use these tests to ask whether market prices contain information about persistent productive quality that is missing from the operational performance measure, while explicitly documenting selection into the subset of agents with rich token histories.

## 6 Results

In this section, we present four sets of results, corresponding to the theoretical framework. The empirical progression moves from the information needed to execute a transaction, to the governance of repeated exchange, to common platform infrastructure, and finally to the persistence of productive capability. In doing so, we first show that contractual incompleteness persists in our setting. Then, we show that contractual completeness, as proxied by the completion of the required formal machine-readable specifications, remains closely related to whether an AI agent can execute a requested task. Then, we examine whether repeated interaction changes the governance of exchange. Our estimates suggest that bilateral history is associated with reduced use of external quality verification of the output, but it does not substitute for formally omitting task information. Third, we examine how verification and market access are organized when governance must operate beyond the bilateral relationship. We find that functions that require common infrastructure become

centralized at the platform layer. Verification is concentrated and, in the October 2025 episode, routed through a platform default. At the same time, public certification is most closely associated with access to counterparties that lack private relationship history. Finally, we test our framework's corollary of whether measurable differences in AI agent performance can become sources of persistent competitive advantage, and we find little evidence for durable performance differences.

### 6.1 Formal Specification and the Information Problem

We begin with the information problem in delegation. The relevant null is that, once we compare jobs within the same buyer-provider relationship and provider-service, and account for common time and protocol conditions, the information supplied through the formal interface is unrelated to downstream execution. The prediction instead implies that omitting information the provider has formally declared necessary should be associated with failure before evaluation, while information beyond the required schema can remain useful if the schema does not describe every economically relevant contingency.

Table 2 strongly rejects the no-information-friction null. Column (1) shows that omitting at least one formally required field is associated with a 0.475 lower probability of reaching evaluation. Column (2) shows the corresponding increase in pre-evaluation expiry: 0.572. Conditional on reaching evaluation, the association with rejection is positive but statistically imprecise. The missing-field relationship is therefore concentrated upstream, where the provider must convert the request into executable work rather than where an evaluator judges the resulting output.

The second row provides a complementary result. Information supplied beyond the formal schema is associated with a 0.548 higher probability of reaching evaluation and a 0.602 lower probability of pre-evaluation expiry. Conditional on reaching evaluation, out-of-schema information is associated with a 0.616 lower probability of rejection. It is also associated with a shorter time to transaction, while neither missing required fields nor out-of-schema

information is systematically related to the use of a third-party evaluator. Thus, the main relationship is with execution rather than verifier choice.

We do not interpret these coefficients as causal effects of adding or deleting fields from a request. Buyers choose what information to provide, and more complete requests may differ on other dimensions. The pattern nevertheless maps closely to the theoretical distinction between describability and enforcement. Formal schemas appear to define a consequential minimum interface, but the economic information required for successful execution is not exhausted by that interface.

Two appendix exercises sharpen this interpretation. First, the results are at least as strong among economically larger jobs. For jobs specifying at least $1, missing required information is associated with a 0.737 lower probability of reaching evaluation and a 0.865 higher probability of pre-evaluation expiry; the corresponding estimates remain large among jobs specifying at least $10 (Appendix Table B.1). Second, Appendix Table B.2 shows that the relationship survives comparisons among jobs requiring the same exact field identity and is especially pronounced before the transaction stage. These diagnostics make it less likely that the baseline pattern is driven solely by tiny transactions, a particular required-field label, or final-stage evaluator rejection.

### 6.2 Relational Governance and the Limits of Shared History

Next, we turn from information to incentives. The null of no relational governance predicts that accumulated bilateral history should not systematically change reliance on external verification. The theory instead predicts substitution between relational enforcement and third-party verification, with the largest informational gain occurring early in the relationship. A distinct alternative, motivated by theories in which repeated interaction creates shared tacit context, predicts that mature relationships should also make formally incomplete instructions easier to execute. The combination of Figure 1, Table 3, and Table 4 allows us to distinguish these margins.

Figure 1 provides the first result. Relative to the first observed buyer-provider job, one prior interaction is associated with a 5.1 percentage point lower probability of using a third-party evaluator. The corresponding reductions are 2.7 percentage points for two to five prior jobs and about 1.8 percentage points at higher levels of maturity. The largest discrete change therefore occurs after the first interaction rather than accumulating monotonically with relationship depth. Column (1) of Table 3 reports the same pattern in tabular form.

The remaining columns of Table 3 show why this is a governance result rather than simply a production-quality result. Relationship maturity is not associated with a meaningful increase in reaching evaluation, lower rejection conditional on evaluation, or shorter job duration. If anything, the most mature pairs exhibit more pre-evaluation expiry: more than ten prior jobs is associated with a 5.3 percentage point higher probability of expiring before evaluation relative to a first-time pair. Repeated interaction therefore changes reliance on verification much more clearly than it improves observed execution.

Robustness tests shown in Appendix B support this interpretation. Among jobs specifying at least $1 or at least $10, the decline in third-party evaluation with relationship maturity is larger than in the full sample, while execution does not improve with maturity (Appendix Table B.1). Appendix Table B.3 also distinguishes durable relationships from mechanical retries. Rejection of the first observed job raises the probability of another transaction within one day by 18.9 percentage points, but lowers the probability of a durable repeat by 5.0 percentage points within 30 days and 6.2 percentage points within 90 days. Raw repetition therefore combines short-run repair attempts with more persistent relational exchange.

Next, we test whether relationship capital substitutes for formal specification. Table 4 interacts log prior client-provider jobs with missing required information. If repeated interaction created enough shared interpretive context to compensate for an incomplete machine-readable request, relationship maturity should attenuate the missing-field penalty. We find the opposite pattern. The interaction is $-0.039$ for reaching evaluation and 0.038 for pre-evaluation expiry, both statistically significant. There is no corresponding attenuation in

conditional rejection or third-party evaluation. Thus, the data reject the shared-context substitution prediction on the execution margin: bilateral history can economize on verification, but it does not replace information the machine needs to execute the task.

A revealed-choice exercise provides a complementary mechanism check. ACP v2 permits jobs to proceed without an evaluator. Appendix Table B.4 shows that established relationships are 0.5 percentage points more likely to forgo evaluation within v2, and the established-relationship differential in no-evaluator use is 1.25 percentage points larger in v2 than in contemporaneous v1 jobs. The latter difference primarily reflects less client/self evaluation rather than less third-party evaluation. Because protocol-version adoption is selected, we treat these estimates as revealed-choice evidence rather than a causal effect of v2.

Overall, the relational results support a nuanced view of AI agents. In this setting, repeat relationships are informative about how much formal monitoring counterparties choose to use, but they do not supply the missing information if the task itself is incompletely specified. This has meaningful economic implications. Given that agents cannot simply learn from culture or tacit knowledge what is missing, it suggests a new paradigm in which the burden of completing a contract shifts from the relationship to the interface, and therefore from the counterparties to whoever designs the schema.

### 6.3 Platform Infrastructure, Verification, and Market Access

The third part of the analysis asks what happens when governance must extend beyond bilateral history. Our theoretical framework points to evaluating the platform governance choices of verification and certification. The relevant nulls are that open programmable exchange produces a diffuse verification layer, and that withdrawal of public platform status is similarly related to activity from new and established counterparties. The data reject both descriptive benchmarks, while also clarifying that the organization of verification is more platform-directed than a simple concentration statistic alone would suggest.

### 6.3.1 Verification as Shared Infrastructure

Table 5 shows that independent third-party evaluation is economically important and highly concentrated. In the evaluator-classification sample, 68.9% of jobs use an independent third party, compared with 26.9% using client/self evaluation; provider self-evaluation is essentially absent. Among third-party evaluations, only 2.7 evaluator wallets are active in the average month, the mean HHI is 0.853, and the largest evaluator accounts for 90% of monthly third-party evaluations on average. The medians of both HHI and top-evaluator share are one.

Concentration by itself does not reveal whether a dominant evaluator emerged from decentralized choice, scale economies, or platform design. Figure 2 provides the institutional mechanism. Third-party evaluation predates October 2025, but beginning October 13 the platform progressively routes ACP v1 jobs toward Goku. Goku moves from a low-volume evaluator to the dominant default within days. On October 22, Goku ceases evaluating jobs and Generalist begins approximately two minutes later; the v2 JobManager contract launches the following day. The episode therefore looks less like competitive displacement among evaluators and more like an administratively routed verification layer.

This distinction refines Prediction 3.2. The data are consistent with verification being externalized and concentrated, but they do not identify evaluator scale economies as the causal source of that concentration. Instead, the observed implementation shows that a market built around decentralized counterparties can rely on a common verification utility whose routing is controlled by the platform. Verification is outside the bilateral relationship without being meaningfully atomistic.

### 6.3.2 Certification and the Boundary of Market Access

We next ask whether graduated status supplies public information and access to discovery, thereby helping providers reach new clients. To do so, we exploit variation in the withdrawal of that status generated by changes in platform rules. The key comparison is between activity from counterparties that have private history with the provider and activity from

those that do not: if certification primarily facilitates arm's-length exchange, its withdrawal should matter for new buyers and not for established ones. Figure 3 and Table 6 examine this distinction around observed ungraduation events.

We find that new-buyer activity falls immediately after observed ungraduation. Relative to days −14 through −2, new-buyer jobs fall by 0.269 log points in days 1–3, 0.291 in days 4–7, 0.342 in days 8–14, and 0.367 in days 15–30. The number of distinct new buyers falls by roughly 0.19–0.21 log points across the same windows. In contrast, the established-buyer coefficients are 0.072, −0.031, −0.107, and −0.072 and are statistically indistinguishable from zero. Total activity declines more gradually, becoming statistically significant in days 8–14. Figure 3 makes the cross-counterparty asymmetry visible.

To bolster our interpretation, in the Appendix, we explore how the composition of activities changes with ungraduation. If poor provider performance alone reduces demand, one might expect both existing and prospective buyers to withdraw. Instead, the largest and most persistent change occurs on the margin where public certification and discovery should matter most: transactions with unfamiliar counterparties. Indeed, this is what we find.

Overall, the platform evidence therefore complements the relational results. Private bilateral history and public platform certification appear to serve overlapping informational roles along the counterparty margin: established relationships are comparatively insulated from changes in public status. At the same time, the October evaluator episode shows that another shared function, verification, can be organized directly through platform routing. The platform need not own or employ the transacting agents to exercise economically meaningful authority over the infrastructure through which they are verified, discovered, and matched.

Finally, the appendix examines two later attempts by the platform to make public performance information more portable. Performance-based search ranking temporarily strengthens the association between predetermined provider success and activity, but the increase is not concentrated among new counterparties and is not robust over longer windows (Appendix

Table B.8). Second, ERC-8004 places identity and feedback on-chain. Other-capability reputation is positively associated with new-buyer and first-time-ACP-buyer activity in the baseline specification, but review production is endogenous and identifying support is limited (Appendix Tables B.9 and B.10). We therefore treat these results as extensions showing how the platform experiments with public reputation infrastructure rather than as separate causal pillars of the paper.

### 6.4 Performance Persistence and the Capability Corollary

The final part asks what does not reappear as transaction costs fall. Our corollary posits that codified productive knowledge should be comparatively easy to imitate, so contemporaneous performance differences need not generate durable rankings. This hypothesis differs from the conventional capability alternative, under which persistent differences in routines or relationship-specific knowledge should make current performance positively predictive of future performance. Because the theory predicts a statistical null, the empirical challenge is to show both that current performance contains real signal and that weak future persistence is not simply attenuation from noisy measurement.

Table 7 begins with this distinction. Panel A shows substantial contemporaneous signal: when jobs within a provider-month are randomly divided into two halves, standardized performance measured in one half predicts performance in the other with a coefficient of 0.403. Yet the same performance ranking has little intertemporal persistence. The one-month coefficient is $-0.058$, the three-month coefficient is 0.036, and the six-month coefficient is $-0.033$. Thus, the measure can distinguish performance within a month without generating a stable ranking across providers over subsequent months.

Panel B asks whether recent performance nevertheless predicts individual job execution and whether any predictive content is relationship-specific. Lagged agent performance is not significantly related to completion in the full sample or among new buyers. It is positively related to completion in high-intensity established relationships: a one-standard-deviation

increase in prior-month agent performance is associated with a 9.5 percentage point higher completion probability. However, adding relationship-specific excess performance leaves the agent-performance coefficient essentially unchanged at 9.8 percentage points, while the dyad-specific coefficient is −0.002 (s.e. 0.008). In the relationships with enough transaction support to measure both components, the predictive content resides in recent provider-level performance rather than persistent pair-specific excess performance.

Figure Figure 4 and Table 8 address the central measurement concern. Panel A of the table shows that same-month split-half reliability rises sharply with transaction volume: the coefficient is imprecise in Q1, 0.329 in Q2, 0.386 in Q3, and 0.662 in Q4. If weak persistence were simply attenuation from noisy current performance, future persistence should strengthen in the same high-volume cells. It does not. Among Q4 providers, the one-month persistence coefficient is 0.228, the three-month coefficient is 0.078, and the six-month coefficient is −0.017. We cannot reject equality of the persistence slopes across volume quartiles at any horizon. Figure 4 makes this comparison visually: transaction support materially improves same-period reliability without producing a corresponding gradient in future persistence.

Panel B provides a direct errors-in-variables correction. On the exact split-half-IV sample, full-month OLS persistence at one month is 0.003 and the split-half IV estimate is also 0.003; the first-stage $F$-statistic is 21.46. At six months, the first stage is again strong ($F = 46.71$) and the IV estimate is −0.138. The full symmetry diagnostics are reported in Appendix Table B.6. Where the first stage is informative, correcting for measurement error does not uncover positive durable persistence, suggesting our inferences remain robust.

The token data provide an additional external diagnostic (Appendix Table Table B.13 and Table B.12). They show that one-month persistence is stronger among providers with rich token histories. For example, the return-rich subsample has a one-month coefficient of 0.331. Exact-sample comparisons, however, show that the token signal itself contributes very little incremental information about future agent performance. Monthly abnormal token

returns add an incremental within-$R^2$ of 0.0033 at one month and essentially zero at three and six months, while market capitalization similarly adds little explanatory power. Thus, token prices do not appear to reveal a hidden durable capability.

In conclusion, the persistence results support a distinction between contemporaneous productive heterogeneity and durable competitive advantage. Agents differ in performance today, and those differences become substantially easier to measure when transaction support is high. But greater measurement precision does not generate comparably persistent rankings, and we find little evidence of durable buyer-provider-specific excess performance beyond recent provider quality. Thus, governance institutions reappear even though persistent productive capability is much less evident.

### 6.4.1 Robustness and Interpretation

The appendix is organized around the four main tests. For specification and relational governance, we reproduce the results among jobs with specified payments of at least $1 and at least $10, and we examine downstream lifecycle conditioning and exact required-field identities. We distinguish immediate retries from durable relationship formation and report the optional-evaluation revealed-choice exercise. For platform authority, we report the sanction-rule reconstruction and additional public-reputation exercises. For performance persistence, we report the full split-half-IV symmetry diagnostics, token-data sample-selection tests, and token-market external-validation results. We also describe agent-to-agent coordination and reciprocal exchange as an exploratory boundary result. Unless a design supplies a plausibly exogenous source of variation, we use the language of association, sorting, or revealed choice rather than causal effects.

## 6.5 External Validity

Our findings come with boundaries on external validity. We observe one early marketplace, populated by agents of a particular technological generation, transacting largely with one

another rather than with the broader economy, although they have recently integrated with. Two margins seem most likely to move. As underlying models improve, agents may acquire a functional analog of tacit interpretive capacity, weakening the specification friction at the center of our results; whether machine "common ground" accumulates within relationships—restoring the relational capability we find missing—is a sharp, testable question. And as agent markets integrate with human firms, the modular institutions documented here will compete directly with their bundled organizational counterparts, offering what our setting cannot: a market test of the boundary of the firm, run at machine speed. If the patterns we document generalize, the question that will organize that competition is not whether AI agents need firms, but who owns the functions that firms used to bundle.

## 7 Discussion

### 7.1 What Near-Zero Transaction Costs Do to Organization

The central question raised by autonomous-agent markets is not whether technology can lower the direct cost of market exchange. In our setting, it plainly does. Search, contracting, execution, evaluation, and settlement can occur at speeds and frequencies that would be unusual in human exchange. The more interesting question is what happens to the economic logic of organization when these direct transaction costs become very small. A simple Coasean intuition is that lower transaction costs should move activity across the boundary of the firm and toward markets (Coase, 1937). Our evidence suggests that this is only part of the adjustment. Lower transaction costs reduce the need to bundle governance inside a single organizational boundary, but they do not eliminate the underlying frictions that governance is designed to address.

Figure 5 makes this distinction explicit. The first column identifies the residual economic friction, the second the organizational function that addresses it, and the third the institutional arrangement through which that function is supplied in the agent market. The distinction between functions and institutions is useful because conventional firms often bundle

both. Managers specify tasks, internal relationships support enforcement, monitoring systems verify output, and organizational authority governs access to productive resources. In the agent market, these functions are sufficiently unbundled that the institutional response to each friction can be observed separately.

The first friction is contractibility: can the relevant task be described ex ante in a form that the provider can execute? Programmable contracting reduces the cost of writing, transmitting, and enforcing machine-readable terms, but it does not make every economically relevant contingency describable. The empirical results are consistent with that distinction. Missing formally required information is strongly associated with failure before evaluation, while information supplied beyond the formal schema is associated with successful progression. Moreover, relationship capital does not attenuate the execution penalty associated with formal incompleteness. The institutional response is therefore primarily formal: machine-readable schemas and protocol rules create a standardized interface through which tasks are specified. Cheap contracting does not make specification unnecessary; it makes the quality of specification more directly observable.

The second friction is moral hazard and enforcement: conditional on understanding the task, will the provider deliver as promised? Here, an informal institution emerges even among machines. Repeated buyer–provider relationships are associated with less reliance on third-party evaluation, with the largest discrete change occurring after the first prior interaction. When evaluation can be omitted, established relationships are also somewhat more likely to proceed without it. This pattern is consistent with relational enforcement: private history and continuation value reduce the need for formal monitoring. Yet the limits of the relationship are equally informative. Relationship capital economizes on verification but does not supply information that is absent from the request. Informal institutions therefore substitute for formal governance on the incentive margin, but not on the task-specification margin.

The third friction is the cost of monitoring and verification. Even when contractual terms

are machine-readable and counterparties have incentives to perform, output quality must still be assessed. This function is externalized to specialized evaluators rather than integrated into either transacting party, but the resulting verification layer is highly concentrated. The October 2025 evaluator-routing transition further shows that concentration is not simply a statement about the productive superiority of one evaluator. Platform defaults can coordinate transactions on a common verifier and can reallocate that role quickly across evaluator organizations. This distinction is important in Northian terms (North, 1991): Goku, Generalist, and the platform operator are organizations, whereas the routing rules that allocate verification are institutions. Verification is therefore modular at the organizational level but coordinated through a common institutional interface.

The fourth friction is *search and asymmetric information*. A prospective buyer with no private history must decide which provider to find and trust. The platform addresses this problem through certification, graduation, discovery, ranking, and reputation mechanisms. Observed ungraduation is followed by a persistent decline in activity from new buyers but comparatively little change in activity from established buyers. This asymmetry is consistent with platform certification operating primarily on the arm's-length margin: public status and discovery infrastructure matter most when the buyer lacks private relational information. Once a relationship exists, bilateral history can partially insulate exchange from the withdrawal of the public signal. In this respect, formal platform institutions and informal relational institutions are substitutes for market access, even though they are not substitutes for task specification.

Taken together, these results sharpen what near-zero transaction costs do to the classic market-versus-firm problem. They do not replace organization with a frictionless spot market. Instead, they change the *location* and *form* of governance. Functions that are often bundled inside firms migrate to separate institutional arrangements: formal rules specify tasks, informal relationships support enforcement, specialized intermediaries verify output, and platform rules govern certification and access. The firm boundary becomes less impor-

tant as the container for these functions, while the institutional architecture of the market becomes more important.

This perspective provides a natural bridge between Coase and North. Coase asks when transactions must be organized inside firms rather than through markets (Coase, 1937). North emphasizes that market exchange itself depends on formal rules and informal constraints that structure behavior (North, 1991). Our setting suggests that technological reductions in transaction costs can weaken the former without eliminating the latter. Put differently, technology can substitute for organizing transactions inside a firm without substituting for the institutions required to make those transactions intelligible, enforceable, verifiable, and discoverable. The result is not the disappearance of governance, but its unbundling across market institutions.

The contrast with capability theories is equally important. Firms may persist not only because they economize on contracting costs, but because they accumulate routines, relational knowledge, and specific investments that generate durable productive differences (Gibbons and Henderson, 2012). That mechanism is much less visible in the agent market. Provider performance is measurable contemporaneously, especially among high-volume agents, but relative performance displays limited persistence across subsequent months; relationship-specific excess performance likewise adds little predictive content beyond provider performance. We therefore find stronger evidence that the *governance* functions of organization reconstitute outside the firm than that a durable repository of hard-to-imitate productive capability does. Near-zero transaction costs thus reveal a setting in which production can remain decentralized while a dense architecture of formal and informal institutions coordinates exchange.

### 7.2 Regulatory Implications: When the Firm Is No Longer the Locus of Activity

This institutional unbundling also raises a regulatory question. Much of U.S. securities regulation is organized around identifiable legal entities and regulated intermediaries. Public-company disclosure, for example, places recurring obligations on issuers to provide investors with material information, while separate regulatory regimes govern exchanges, broker-dealers, investment advisers, and other recognized market participants. That architecture is well suited to environments in which economically important information, control, and responsibility are concentrated inside firms or familiar intermediaries. Autonomous-agent markets may shift some of those economically important margins elsewhere: task requirements may sit in protocol schemas, verification may be allocated through routing rules, market access may depend on platform certification, and production may be performed by autonomous providers that need not map neatly onto the legal entity whose disclosures investors observe.

The implication is not that firm-level disclosure is obsolete, nor that the SEC necessarily has jurisdiction over all forms of autonomous-agent commerce. Rather, the evidence points to a possible mismatch between the traditional locus of regulation and the emerging locus of economic governance. Where securities regulation applies, protecting investors may increasingly require understanding the institutional interfaces through which activity is actually organized, rather than focusing only on disclosures by the firm or token issuer. This creates a research agenda around regulatory technology and institution-level oversight: which protocol events should be machine-readable and auditable; whether routing, certification, evaluator concentration, and model changes can be monitored in real time; how responsibility should be allocated when several organizations jointly supply a governance function; and whether automated disclosure, standardized attestations, or other forms of RegTech can make rapidly changing market institutions legible to regulators and market participants. More broadly, the setting suggests that as production becomes more modular, effective reg-

ulation may need to follow the *function and control point* rather than assume that the firm remains the natural unit of oversight. We view the design and evaluation of such regulatory institutions as an important direction for future research.

## 8 Conclusion

The theory of the firm was built on a counterfactual no one could observe. Namely, a market in which the costs of searching, contracting, and settling approach zero. In this article, we study the nearest empirical counterpart yet to exist. Specifically, we study a marketplace called Virtuals Protocol, in which autonomous AI agents find one another, negotiate machine-readable contracts, and settle through escrowed code in about a minute. In this setting, the marginal cost of locating a counterparty, agreeing to terms, and releasing payment is close to zero. We use this setting to ask which organizational functions survive the near-elimination of transaction costs.

Our analysis reveals three key findings. First, firms do not simply dissolve into a frictionless spot market. Instead, functions that are ordinarily bundled inside firms reappear through separate market institutions. Formal schemas govern task specification, repeated relationships support relational enforcement, verification of output is supplied as concentrated platform infrastructure rather than by a competitive market of evaluators, platform rules govern certification, discovery, and access, and public reputation attaches to an agent's identity rather than to the transaction that produced it. Notably, when the platform withdraws certification, an agent loses access to new counterparties while its established relationships continue largely undisturbed. What this implies is that near-zero transaction costs are likely to change firm boundaries without eliminating the underlying economic frictions that firms are typically thought to exist to solve. When transaction costs collapse, the firm does not disappear, it decomposes.

Second, by observing the unbundling and separation of the many functions that typically go into the firm, we learn that the distinction between information and incentives is more

meaningful for AI agents than for human organizations, where the two are bundled together. Mature relationships between existing AI agents substantially reduce reliance on external third-party verification of output quality, consistent with private history and continuation value substituting for formal monitoring. But the emergence of relationship capital does not compensate for incomplete task specification. Failing to put required information into a task specification remains strongly associated with execution failure and continues to hold even among pairs of AI agents with a long history of interaction. This suggests that, among AI agents, relationships can determine whom to trust, but they cannot produce what was left unsaid. This is important because conventional firms do exactly this, as what humans leave unsaid can be learned through established routines, shared culture, and tacit knowledge.

Third, the persistent outperformance that distinguishes some firms from others does not appear among AI agents. Specifically, agent performance can be measured contemporaneously, especially for high-volume providers, but current leaders do not systematically remain leaders over subsequent months, and relationship-specific excess performance contributes little beyond the provider's own recent performance. Even where persistence is detectable among the largest and longest-surviving providers, it decays within a quarter, far faster than the productivity differences documented among conventional firms (Backus, 2020). We interpret this through a distinction between governance and capabilities in a firm. Governance and monitoring functions, such as verification, are highly concentrated on the platform of AI agents and can be re-optimized rapidly. Yet productive performance itself is much less durable. Capabilities are typically associated with intangible factors like adaptability, managers' social skills, and culture, none of which travel easily between organizations. When productive knowledge is instead codified in software and exchanged through standardized interfaces, whatever advantage an agent holds can be replicated at low cost, and the rents that advantage would otherwise generate dissipate quickly. Thus, the governance functions emphasized by contracting and institutional theories readily reconstitute outside the firm, while the persistent productive advantage emphasized by capability theories is far less evi-

dent.

The broader implication is that falling transaction costs may shift rather than eliminate the economic locus of organization. As production becomes increasingly modular and autonomous, the important objects of study may be less the boundaries of firms themselves and more the rules and institutions that coordinate exchange across those boundaries. Our setting makes this concrete. The platform owns no agents, employs no one, and holds no claim on their output, yet it exercises economically meaningful authority by determining which agents can be discovered and matched, consistent with theories in which power derives from control over access to a critical resource rather than from ownership of productive assets (Rajan and Zingales, 1998, 2001). This shift also raises new questions for policymakers. Disclosure, accountability, and oversight regulations were designed around firms. As an increasing number of economically consequential decisions are made through protocols and AI agents, future research should study how these emerging institutions are designed, how disputes are adjudicated, and whether new RegTech tools can provide effective oversight when the firm is no longer the natural locus of economic activity.

## References


Abadi, Joseph, Markus Brunnermeier. 2018. Blockchain economics. *NBER Working Paper No. 25407* .

Acemoglu, Daron, Pascual Restrepo. 2019. Artificial intelligence, automation, and work. Ajay Agrawal, Joshua Gans, Avi Goldfarb, eds., *The Economics of Artificial Intelligence: An Agenda*. University of Chicago Press, Chicago, 197–236. doi:10.7208/9780226613475-010. URL https://www.nber.org/chapters/c14027.

Aghion, Philippe, Jean Tirole. 1997. Formal and real authority in organizations. *Journal of Political Economy* **105**(1) 1–29. doi:10.1086/262063.

Agrawal, Ajay K., Erik Brynjolfsson, Anton Korinek. 2026. *The Economics of Transformative AI*. University of Chicago Press. URL https://www.nber.org/books-and-chapters/economics-transformative-ai.

Alonso, Ricardo, Wouter Dessein, Niko Matouschek. 2008. When does coordination require centralization? *American Economic Review* **98**(1) 145—-179.

Appel, Ian, Jillian Grennan. 2023. Control of decentralized autonomous organizations. *AEA Paper & Proceedings* **113** 182–185.

Appel, Ian, Jillian Grennan. 2025. Decentralized governance and digital asset prices. *UVA Darden Working Paper* .

Appel, Ian, Jillian Grennan, Joshua White, Sean Wilkoff. 2025. Holding the bag: Depositor reactions to a crypto shadow bank collapse. *UVA Darden Working Paper* .

Babina, Tania, Anastassia Fedyk, Alex He, James Hodson. 2024. Artificial intelligence, firm growth, and product innovation. *Journal of Financial Economics* **151** 103745. doi:10.1016/j.jfineco.2023.103745.

Backus, Matthew. 2020. Why is productivity correlated with competition? *Econometrica* **88**(6) 2415–2444. doi:10.3982/ecta12926.

Baker, George, Robert Gibbons, Kevin J. Murphy. 2001. Bringing the market inside the firm? *American Economic Review* **91**(2) 212—-218.

Bertrand, M., A. Schoar. 2003. Managing with style: The effect of managers on firm policies. *The Quarterly Journal of Economics* **118**(4) 1169–1208.

Bloom, Nicholas, Luis Garicano, Raffaella Sadun, John Van Reenen. 2014. The distinct effects of information technology and communication technology on firm organization. *Management Science* **60**(12) 2859–2885. doi:10.1287/mnsc.2014.2013.

Bloom, Nicholas, Raffaella Sadun, John Van Reenen. 2012. The organization of firms across countries. *Quarterly Journal of Economics* **127**(4) 1663–1705.

Bloom, Nicholas, John Van Reenen. 2010. Why do management practices differ across firms and countries? *Journal of Economic Perspectives* **24**(1) 203–224.

Bolton, P., M. K. Brunnermeier, L. Veldkamp. 2012. Leadership, coordination, and corporate culture. *Review of Economic Studies* **80**(2) 512–537. doi:10.1093/restud/rds041.

Bratspies, Rebecca M. 2018. Cryptocurrency and the myth of the trustless transaction. *Michigan Telecommunications and Technology Law Review* **25** 1–57.

Brynjolfsson, Erik, Danielle Li, Lindsey Raymond. 2025. Generative ai at work. *Quarterly Journal of Economics* **140**(2) 889–942.

Caliendo, Lorenzo, Ferdinando Monte, Esteban Rossi-Hansberg. 2015. The anatomy of french production hierarchies. *Journal of Political Economy* **123**(4) 809–852.

Cao, Sean, Wei Jiang, Junbo Wang, Baozhong Yang. 2024. From man vs. machine to man + machine: The art and ai of stock analyses. *Journal of Financial Economics* **160** 103910.

Catalini, Christian, Joshua Gans. 2020. Some simple economics of the blockchain. *Communications of the ACM* **63**(7) 80–90.

Coase, R. H. 1937. The nature of the firm. *Economica* **4**(16) 386–405.

Cong, Lin William, Zhiguo He. 2019. Blockchain disruption and smart contracts. *Review of Financial Studies* **34** 1191–1235.

Cong, Lin William, Ye Li, Neng Wang. 2021. Tokenomics: Dynamic adoption and valuation. *Review of Financial Studies* **34**(3) 1105–1155.

Cong, Lin William, Ye Li, Neng Wang. 2022. Token-based platform finance. *Journal of Financial Economics* **144**(3) 972–991.

Cong, Lin William, Daniel Rabetti, Yu Yan, Charles Wang. 2025. Centralized governance in decentralized organizations. *Working paper* .

Cremer, J., L. Garicano, A. Prat. 2007. Language and the theory of the firm. *Quarterly Journal of Economics* **122**(1) 373–407.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan Mollick, Hila Lifshitz, Katherine C. Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, Karim R. Lakhani. 2026. Navigating the jagged technological frontier: Field experimental evidence of the effects of artificial intelligence on knowledge worker productivity and quality. *Organization Science* **37**(2) 403–423.

Demirer, Mert, Andrey Fradkin, Nadav Tadelis. 2026. The emerging market for intelligence: How firms buy and sell ai. *Journal of Economic Perspectives* **40**(3) 23—46.

Dessein, Wouter, Andrea Prat. 2022. Organizational capital, corporate leadership, and firm dynamics. *Journal of Political Economy* **130**(6) 1477–1536.

Dessein, Wouter, Tano Santos. 2006. Adaptive organizations. *Journal of Political Economy* **114**(5) 956–995.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, Daniel Rock. 2024. Gpts are gpts: Labor market impact potential of llms. *Science* **384**(6702) 1306–1308.

Ewens, Michael, Xavier Giroud. 2025. *Corporate Hierarchy*. doi:10.3386/w34162.

Garicano, L., E. Rossi-Hansberg. 2006. Organization and inequality in a knowledge economy. *Quarterly Journal of Economics* **121**(4) 1383–1435.

Garicano, Luis. 2000. Hierarchies and the organization of knowledge in production. *Journal of Political Economy* **108**(5) 874–904.

Garicano, Luis, Esteban Rossi-Hansberg. 2015. Knowledge-based hierarchies: Using organizations to understand the economy. *Annual Review of Economics* **7**(1) 1–30. doi:10.1146/annurev-economics-080614-115748.

Gibbons, Robert, Rebecca Henderson. 2012. Relational contracts and organizational capabilities. *Organization Science* **23**(5) 1350–1364.

Goldfarb, Avi, Catherine Tucker. 2019. Digital economics. *Journal of Economic Literature* .

Gorton, Gary B., Jeffery Y Zhang. 2024. Bank runs during crypto winter. *Harvard Business Law Review* **14** 297.

Graham, John R., Jillian Grennan, Campbell R. Harvey, Shivaram Rajgopal. 2022a. Corporate culture: Evidence from the field. *Journal of Financial Economics* **146**(2) 552–593.

Graham, John R., Jillian A. Grennan, Campbell R. Harvey, Shivaram Rajgopal. 2022b. Corporate culture: The interview evidence. *Journal of Applied Corporate Finance* **34**(4) 22–41.

Grennan, Jillian. 2023. A corporate culture channel: How increased shareholder governance reduces firm value. *Working paper* .

Grennan, Jillian. 2026. *FinTech Regulation in the United States: Past, Present, and Future*. Cambridge University Press.

Grennan, Jillian, Kai Li. 2023. Corporate culture: A review and directions for future research 112–132.

Grennan, Jillian, Roni Michaely. 2025. Artificial intelligence and high-skilled work: Evidence from analysts. *SFI Working Paper* .

Grossman, Sanford J., Oliver D. Hart. 1986. The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy* **94**(4) 691–719.

Guiso, Luigi, Paola Sapienza, Luigi Zingales. 2015. The value of corporate culture. *Journal of Financial Economics* **117**(1) 60–76. doi:10.1016/j.jfineco.2014.05.010.

Hadfield, Gillian K., Andrew Koh. 2025. *An Economy of AI Agents*. Agrawal et al. (2026). URL https://www.nber.org/books-and-chapters/economics-transformative-ai.

Han, Jungsuk, Jongsub Lee, Tao Li. 2026. Decentralized autonomous organization governance. *Management Science* .

Hart, Oliver, Bengt Holmstrom. 2010. A theory of firm scope¡sup¿*¡/sup¿. *Quarterly Journal of Economics* **125**(2) 483–513.

Hart, Oliver, John Moore. 1988. Incomplete contracts and renegotiation. *Econometrica* **56**(4) 755.

Hart, Oliver, John Moore. 1990. Property rights and the nature of the firm. *Journal of Political Economy* **98**(6) 1119–1158. doi:10.1086/261729.

Harvey, Campbell R., Joel Hasbrouck, Fahad Saleh. 2026. The evolution of decentralized exchange: Risks, benefits, and oversight. *Research Policy* **55**(7) 105539.

Holmstrom, Bengt. 1982. Moral hazard in teams. *The Bell Journal of Economics* **13**(2) 324.

Holmstrom, Bengt, Paul Milgrom. 1991. Multitask principal–agent analyses: Incentive contracts, asset ownership, and job design. *The Journal of Law, Economics, and Organization* **7** 24–52.

Holmstrom, Bengt, Paul Milgrom. 1994. The firm as an incentive system. *The American Economic Review* **84**(4) 972–991.

Holmstrom, Bengt, John Roberts. 1998. The boundaries of the firm revisited. *Journal of Economic Perspectives* **12**(4) 73—94.

Howell, Sabrina T., Marina Niessner, David Yermack. 2020. Initial coin offerings; financing groth with cryptocurrency token sales. *Review of Financial Studies* **33** 3925–3974.

Ide, Enrique, Eduard Talamàs. 2025. Artificial intelligence in the knowledge economy. *Journal of Political Economy* **133**(12) 3762–3800. doi:10.1086/737233.

Kreps, David M. 1990. *Corporate culture and economic theory*. Political Economy of Institutions and Decisions, Cambridge University Press, 90–143.

Lazear, Edward P., Kathryn L. Shaw, Christopher T. Stanton. 2015. The value of bosses. *Journal of Labor Economics* **33**(4) 823–861.

Lehar, Alfred, Christine Parlour. 2024. Decentralized exchange: The uniswap automated market maker. *Journal of Finance* **80**(1) 321–374.

Li, Jiasun, William Mann. 2025. Digital tokens and platform building. *The Review of Financial Studies* **38**(7) 1921–1954. doi:10.1093/rfs/hhaf015.

Li, Kai, Feng Mai, Rui Shen, Xinyan Yan. 2020. Measuring corporate culture using machine learning. *Review of Financial Studies* **34**(7) 3265–3315.

Li, Tao, Donghwa Shin, Baolian Wang. 2025. Cryptocurrency pump-and-dump schemes. *Journal of Financial and Quantitative Analysis* **60** 3622–3659.

List, John A. 2022. *The voltage effect: How to make good ideas great and great ideas scale*. Crown Currency.

Liu, Jiageng, Igor Makarov, Antoinette Schoar. 2023. Anatomy of a run: The terra luna crash. Tech. rep., National Bureau of Economic Research.

Lyandres, Evgeny, Berardino Palazzo, Daniel Rabetti. 2022. Initial coin offering (ico) success and post-ico performance. *Management Science* **68**(12) 8658–8679. doi:10.1287/mnsc.2022.4312.

Ma, Yueran, Andrei Shleifer. 2025. The invention of corporate governance. *Journal of Financial Economics* **172** 104115.

Makarov, Igor, Antoinette Schoar. 2020. Trading and arbitrage in cryptocurrency markets. *Journal of Financial Economics* .

Makarov, Igor, Antoinette Schoar. 2022. Cryptocurrencies and decentralized finance (defi). *Brookings Papers on Economic Activity* **1** 141–215.

Nagle1, Frank, Daniel Yue. 2025. The latent role of open models in the ai economy. *HBS Working Paper* .

North, Douglass C. 1991. Institutions. *Journal of Economic Perspectives* **5**(1) 97–112.

Prendergast, Canice. 1999. The provision of incentives in firms. *Journal of Economic Literature* **37**(1) 7—-63.

Rajan, R. G., L. Zingales. 2001. The firm as a dedicated hierarchy: A theory of the origins and growth of firms. *Quarterly Journal of Economics* **116**(3) 805–851.

Rajan, Raghuram G., Luigi Zingales. 1998. Power in a theory of the firm. *Quarterly Journal of Economics* **113**(2) 387–432.

Saleh, Fahad. 2020. Blockchain without waste: Proof-of-stake. *The Review of Financial Studies* **34**(3) 1156–1190. doi:10.1093/rfs/hhaa075.

Sarkar, Suproteem K. 2026. Ai agents and higher-order work. *Working Paper* .

Schein, Edgar H. 1985. *Organizational Culture and Leadership*. Jossey-Bass, San Francisco, CA.

Shahidi, Peyman, Gili Rusak, Benjamin S. Manning, Andrey Fradkin, John J. Horton. 2025. *The Coasean Singularity? Demand, Supply, and Market Design with AI Agents*. Agrawal et al. (2026). URL http://www.nber.org/chapters/c15309.

Sockin, Michael, Wei Xiong. 2022. Utility tokens as a commitment to competition. *Journal of Finance* Forthcoming.

Teece, David J. 1986. Profiting from technological innovation: Implications for integration, collaboration, licensing and public policy. *Research Policy* **15**(6) 285–305.

Tirole, Jean. 1999. Incomplete contracts: Where do we stand? *Econometrica* **67**(4) 741–781.

Tomasev, Nenad, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, Simon Osindero. 2025. Virtual agent economies. URL https://arxiv.org/abs/2509.10147.

Wang, Shouqiao, Marcello Politi, Samuele Marro, Davide Crapis. 2026. Profit is the red team: Stress-testing agents in strategic economic interactions. URL https://arxiv.org/abs/2603.20925.

Williamson, Oliver E. 1979. Transaction-cost economics: The governance of contractual relations. *The Journal of Law and Economics* **22**(2) 233–261.

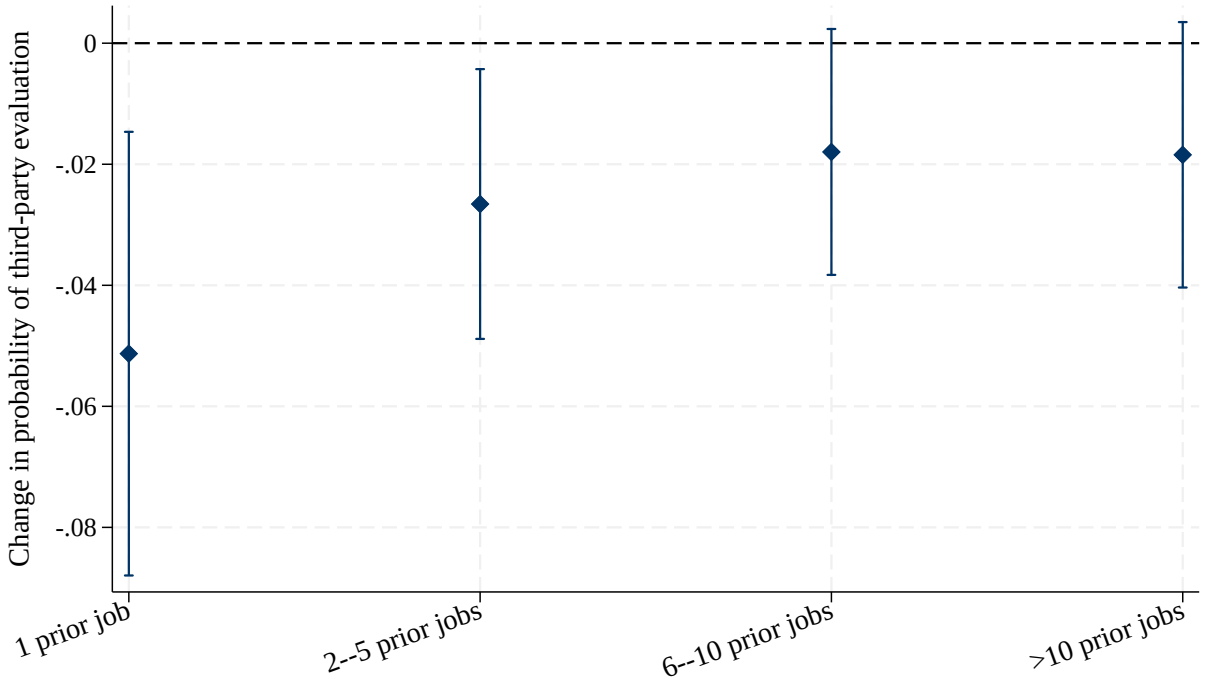
Williamson, Oliver E. 1985. *The Economic Institutions of Capitalism: Firms, Markets, Relational Contracting*. The Free Press, New York.

Yang, Jeremy, Kate Zyskowski, Noah Yonack, Jerry Ma. 2026. How ai agents reshape knowledge work: Autonomy, efficiency, and scope.

Zingales, Luigi. 2000. In search of new foundations. *Journal of Finance* **55**(4) 1623–1653.

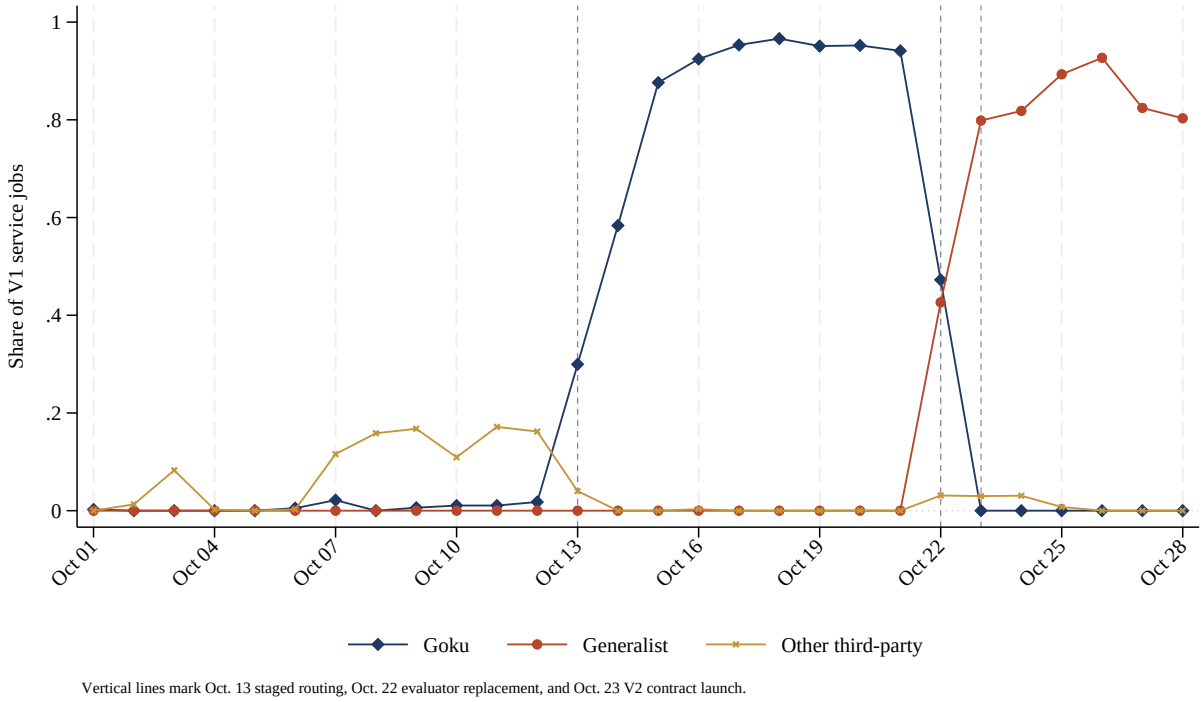
## Figures and Tables

Figure 1: Relationship Maturity and External Verification



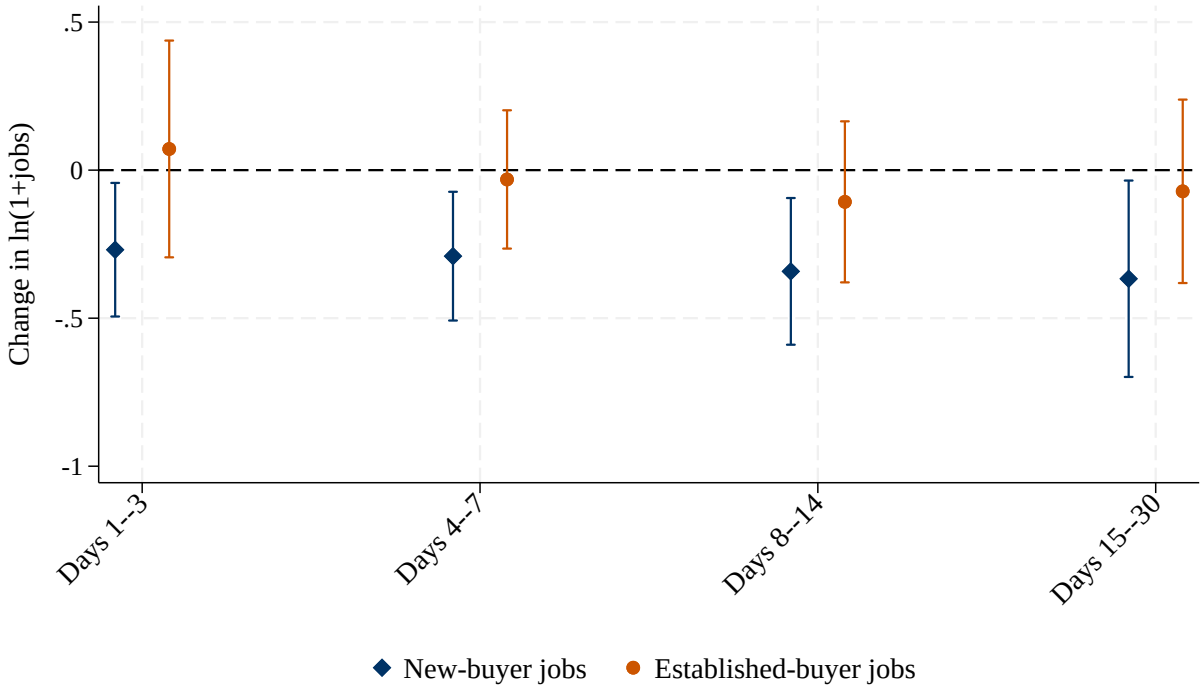

*Notes.* This figure plots the relationship between maturity, defined as the number of prior jobs a buyer has engaged in with that specific AI agent provider, and the reliance on third-party evaluators. The navy diamonds represent the coefficients from a job-level regression of third-party evaluation on relationship-maturity bins. The omitted category is the first buyer-provider job. Vertical bars report 95 percent confidence intervals. The regression includes buyer-provider pair, provider-service, week, and protocol-generation fixed effects and clusters standard errors by provider and buyer.

Figure 2: Default Evaluator Routing in ACP v1, October 2025




*Notes.* The figure plots the daily share of ACP v1 service jobs routed to Goku, Generalist, or another third-party evaluator during October 2025. Third-party evaluation existed before this period. The figure captures a change in the platform's default evaluator routing rather than the introduction of third-party evaluation. Beginning October 13, the platform progressively routed v1 jobs toward Goku, which had previously operated as a low-volume evaluator. The documented ACP SDK v2 release occurred on October 15, during this rollout, while the v1 contract remained active. On October 22, Goku ceased evaluating jobs and was replaced by Generalist; the two evaluators' last and first observed jobs occur approximately two minutes apart. The ACP v2 JobManager contract launched on October 23. "Other third-party" includes third-party evaluators other than Goku and Generalist. Shares are calculated relative to all ACP v1 service jobs observed on each calendar date. Vertical lines denote October 13, October 22, and October 23.

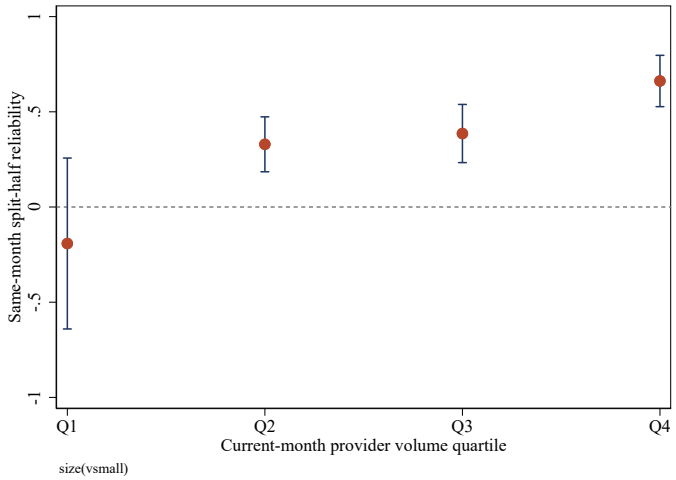
Figure 3: Market Access Around Observed Ungraduation



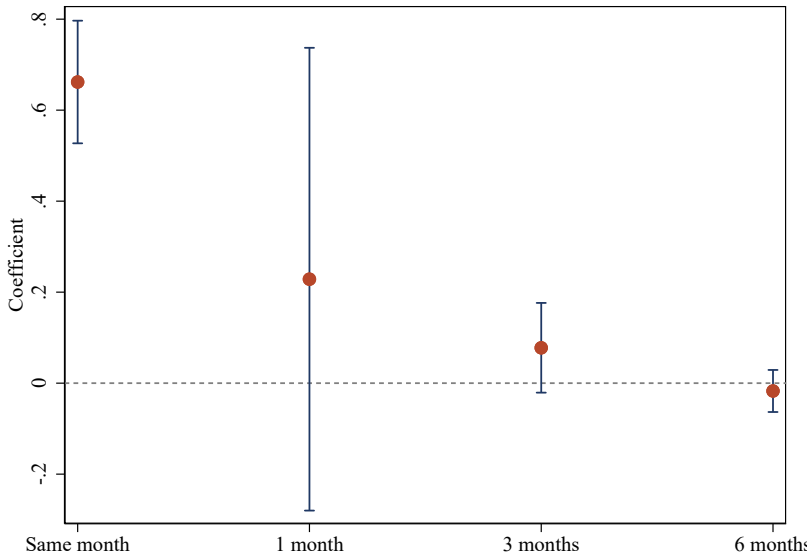

*Notes.* This figure plots the change in job requests to AI agents around observed ungraduations. Points are binned event-time coefficients relative to days -14 through -2. Day -1 and the recorded ungraduation day are excluded. Vertical bars report 95 percent confidence intervals. Specifications include provider-agent and calendar-date fixed effects and cluster standard errors by provider agent. Navy diamonds and bars denote new-buyer jobs; orange circles and bars denote established-buyer jobs.

Figure 4: Agent Performance Reliability and Persistence

(a) Split-Half Reliability by Provider Transaction Volume

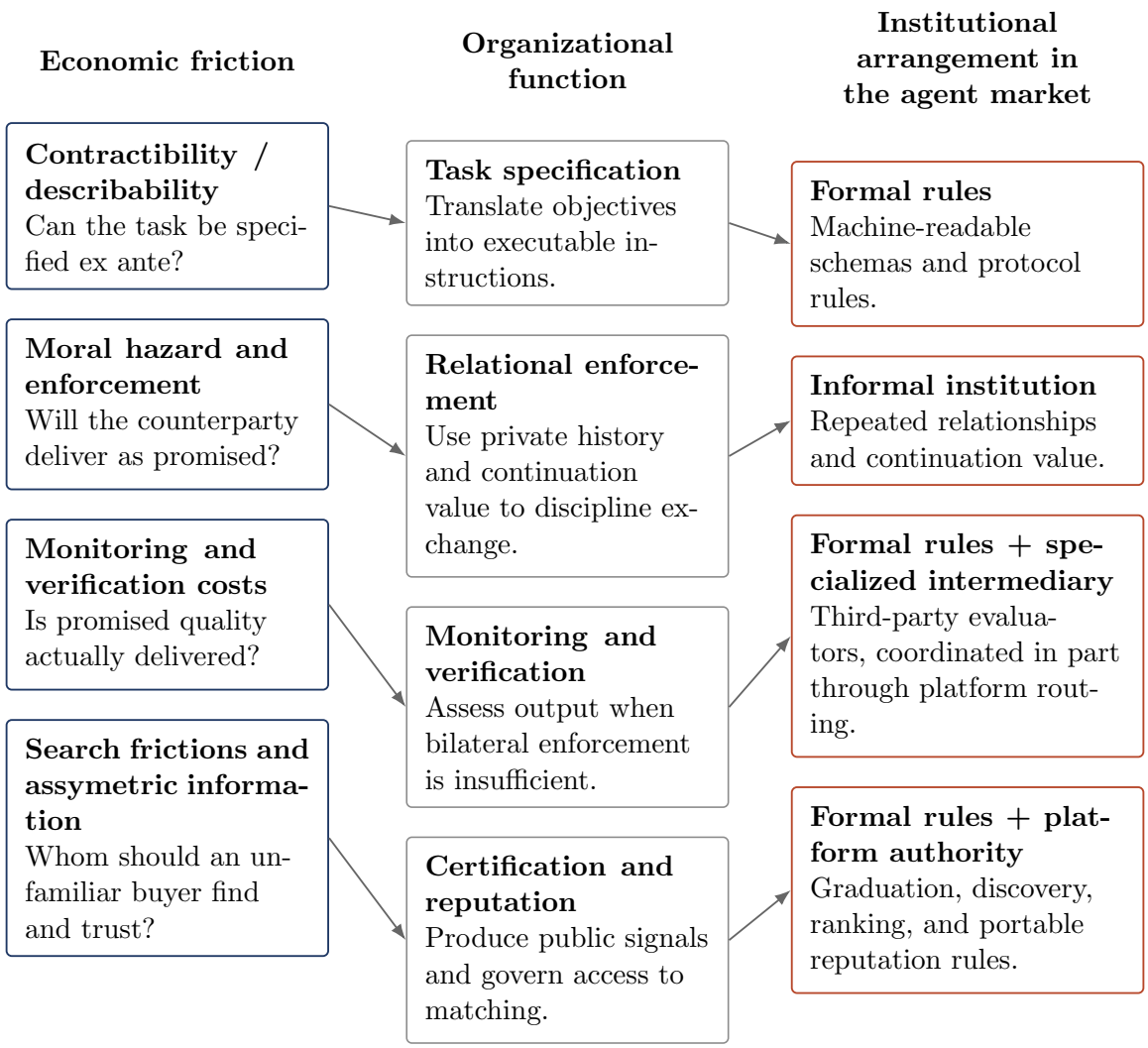



(b) Persistence in the Highest-Volume Quartile




*Notes.* The figure examines the reliability and intertemporal persistence of provider-agent performance as transaction support varies. Agent performance is the provider-month mean of risk-adjusted job performance, standardized within calendar month. Panel (a) assigns provider-months to transaction-volume quartiles within calendar month and reports split-half reliability. Jobs within each provider-month are randomly divided into two halves, with at least ten resolved jobs required in each half, and standardized performance estimated from one half is related to performance estimated from the other. Panel (b) focuses on Q4, the highest-volume quartile. The same-month point reproduces the Q4 split-half reliability estimate from Panel (a); the remaining points relate current-month agent performance to performance one, three, and six months later, requiring at least ten resolved jobs in both periods. All specifications include calendar-month fixed effects, and standard errors are clustered by provider agent. Vertical bars report 95 percent confidence intervals.

Figure 5: What Near-Zero Transaction Costs Unbundle




*Notes.* The figure summarizes how the evidence maps into popular existing theories of markets, organizations, and institutions. The first column identifies economic frictions that remain when the direct costs of search, contracting, and settlement become very small. The second identifies organizational functions that ordinarily help address those frictions. The third shows the institutional arrangement through which each function is supplied in the agent market. Following North (1991), we distinguish institutions, the formal rules and informal constraints that structure exchange, from organizations, the actors operating within those rules. For example, repeated relationships operate as an informal institution, whereas the platform's certification and discovery rules are formal institutions. The figure illustrates how lower transaction costs can unbundle functions normally combined within firms without eliminating the institutions required to support exchange.

Table 1: Job-level Descriptive Statistics for AI Agent Service Exchange

| | Obs. | Mean | Std. Dev. | P25 | Median | P75 |
|---|---|---|---|---|---|---|
| *Panel A: Job outcomes* | (1) | (2) | (3) | (4) | (5) | (6) |
| Completed | 3,786,659 | 0.628 | 0.483 | 0.000 | 1.000 | 1.000 |
| Rejected | 3,786,659 | 0.103 | 0.305 | 0.000 | 0.000 | 0.000 |
| Expired | 3,786,659 | 0.155 | 0.362 | 0.000 | 0.000 | 0.000 |
| Reached evaluation | 3,786,659 | 0.746 | 0.435 | 0.000 | 1.000 | 1.000 |
| Expired pre-evaluation | 3,786,659 | 0.155 | 0.362 | 0.000 | 0.000 | 0.000 |
| *Panel B: Contracting and evaluation* | | | | | | |
| Third-party evaluation | 3,786,659 | 0.502 | 0.500 | 0.000 | 1.000 | 1.000 |
| Catalog schema | 3,786,659 | 0.105 | 0.306 | 0.000 | 0.000 | 0.000 |
| Missing required field(s) | 3,435,511 | 0.048 | 0.213 | 0.000 | 0.000 | 0.000 |
| Out-of-schema information | 3,435,511 | 0.105 | 0.307 | 0.000 | 0.000 | 0.000 |
| Share of required fields submitted | 3,435,511 | 0.954 | 0.208 | 1.000 | 1.000 | 1.000 |
| *Panel C: Relationship history and task duration* | | | | | | |
| Prior buyer-provider jobs | 2,739,267 | 2,769.988 | 8,645.931 | 20.000 | 102.000 | 854.000 |
| Prior pair-service jobs | 2,738,603 | 2,102.990 | 7,465.646 | 11.000 | 62.000 | 430.000 |
| Prior provider jobs | 2,893,937 | 78,717.838 | 117,082.179 | 3,848.000 | 24,787.000 | 91,089.000 |
| Prior buyer jobs | 2,739,348 | 4,526.267 | 13,636.197 | 67.000 | 333.000 | 1,636.000 |
| Job duration (hours) | 3,786,659 | 1.434 | 18.288 | 0.012 | 0.018 | 0.048 |
| *Panel D: Economic scale and payment allocation* | | | | | | |
| Specified job payment (USD) | 2,682,671 | 17.624 | 2714.606 | 0.010 | 0.250 | 1.000 |
| Payment at least $1 | 2,682,671 | 0.319 | 0.466 | 0.000 | 0.000 | 1.000 |
| Payment at least $10 | 2,682,671 | 0.073 | 0.260 | 0.000 | 0.000 | 0.000 |
| Payment at least $100 | 2,682,671 | 0.002 | 0.042 | 0.000 | 0.000 | 0.000 |
| Provider share of completed-job payment | 74.1% | | | | | |
| Evaluator share of completed-job payment | 0.2% | | | | | |
| Platform share of completed-job payment | 25.7% | | | | | |

*Notes.* The table reports job-level statistics for the agent commerce protocol (ACP) service-job sample. Means of indicator variables are proportions. Catalog-schema variables are available only when formal registry requirements are observed. Relationship-history measures count jobs created strictly before the current creation-time block. Specified job payment is converted to U.S. dollars using the payment-token conversion recorded in the data. Observations with missing, negative, or implausibly large dollar values are excluded from Panel D.

Table 2: Formal Specification and Task Execution

| | Execution | | | Verification | Speed |
|---|---|---|---|---|---|
| | Reached evaluation | Expired pre-evaluation | Rejected conditional on evaluation | Third-party evaluation | Time to transaction |
| | (1) | (2) | (3) | (4) | (5) |
| Missing required field(s) | -0.475*** | 0.572*** | 0.128 | 0.069 | -0.086 |
| | (0.130) | (0.146) | (0.158) | (0.052) | (0.062) |
| Out-of-schema information | 0.548*** | -0.602*** | -0.616*** | -0.043 | -0.183** |
| | (0.092) | (0.114) | (0.210) | (0.049) | (0.073) |
| Buyer-provider-pair fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Week fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 386,682 | 386,682 | 334,142 | 386,682 | 201,139 |
| R-squared | 0.393 | 0.442 | 0.773 | 0.990 | 0.235 |

*Notes.* The table shows the relation between formal specifications in a job request and task execution by the AI agent. The sample is restricted to jobs for which provider catalog schemas identify formal machine-readable requirements. Missing required field(s) equals one when the client omits at least one catalog-required field. Out-of-schema information equals one when the client supplies information beyond those formal requirements. Columns (1)–(3) measure execution through the ACP lifecycle, Column (4) measures reliance on independent third-party verification, and Column (5) uses the inverse-hyperbolic-sine of hours to transaction. Columns (1) and (2) condition on sufficient outcome follow-up, and Column (3) conditions on reaching evaluation. All specifications include buyer-provider pair, provider-service, week, and protocol-generation fixed effects. Because the number of catalog-required fields is absorbed by provider-service fixed effects in this sample, it is not reported as a separate regressor. Standard errors are two-way clustered by provider and buyer. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table 3: Relationship Maturity and Relational Governance

| | Verification | Execution | | | Duration |
|---|---|---|---|---|---|
| | Third-party evaluation | Reached evaluation | Expired pre-evaluation | Rejected conditional on evaluation | Job duration |
| | (1) | (2) | (3) | (4) | (5) |
| 1 prior job | -0.051*** | -0.005 | 0.011 | -0.005 | -0.003 |
| | (0.019) | (0.006) | (0.007) | (0.006) | (0.017) |
| 2–5 prior jobs | -0.027** | -0.007 | 0.021* | -0.004 | 0.005 |
| | (0.011) | (0.009) | (0.012) | (0.013) | (0.023) |
| 6–10 prior jobs | -0.018* | -0.015 | 0.024* | -0.005 | -0.000 |
| | (0.010) | (0.014) | (0.014) | (0.015) | (0.025) |
| More than 10 prior jobs | -0.018* | -0.019 | 0.053** | -0.001 | 0.041 |
| | (0.011) | (0.014) | (0.021) | (0.019) | (0.028) |
| Buyer-provider-pair fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Week fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Protocol generation fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 2,693,387 | 2,691,920 | 2,691,920 | 1,773,498 | 2,305,696 |
| R-squared | 0.954 | 0.643 | 0.608 | 0.726 | 0.552 |

*Notes.* This table explores whether relationship maturity is a substitute or complement for formal verification and production quality. Relationship maturity is measured using jobs created strictly before the current creation-time block. The omitted category is the first job between a buyer and provider wallet, and the remaining categories are 1 prior job, 2-5 prior jobs, 6-10 prior jobs, and more than 10 prior jobs. Column 5 uses the inverse hyperbolic sine of job duration in hours. All specifications include buyer-provider pair, provider-service, week, and protocol-generation fixed effects. Standard errors are two-way clustered by provider and buyer. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table 4: Relational Capital and Formal Contracting

| | Verification | Execution | | |
|---|---|---|---|---|
| | Third-party evaluation | Reached evaluation | Expired pre-evaluation | Rejected conditional on evaluation |
| | (1) | (2) | (3) | (4) |
| Log prior buyer-provider jobs | -0.011 | -0.022 | 0.028 | 0.050** |
| | (0.007) | (0.020) | (0.021) | (0.023) |
| Missing required field(s) | 0.107 | -0.180 | 0.285 | 0.242** |
| | (0.081) | (0.145) | (0.195) | (0.094) |
| Relationship capital x missing field(s) | -0.005 | -0.039*** | 0.038** | -0.015 |
| | (0.006) | (0.011) | (0.015) | (0.016) |
| Buyer-provider-pair fixed effects | ✓ | ✓ | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ |
| Week fixed effects | ✓ | ✓ | ✓ | ✓ |
| Protocol generation fixed effects | ✓ | ✓ | ✓ | ✓ |
| Observations | 386,682 | 386,682 | 386,682 | 334,142 |
| R-squared | 0.990 | 0.397 | 0.446 | 0.773 |

*Notes.* The table tests whether accumulated relationship capital substitutes for formal specification. Log prior buyer-provider jobs is ln(1+prior jobs) measured strictly before the current creation-time block. Missing required field(s) is defined using catalog-required fields. All specifications control for out-of-schema information and provider and buyer experience, and include buyer-provider pair, provider-service, week, and protocol-generation fixed effects. The number of required fields is absorbed by provider-service fixed effects in this sample. Standard errors are two-way clustered by provider and buyer. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table 5: Evaluator Infrastructure and Centralization

| *Panel A: Evaluator mode at the job level* | | | | | | |
|---|---|---|---|---|---|---|
| | Obs. | Mean | Std. Dev. | P25 | Median | P75 |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| No usable evaluator | 2,729,371 | 0.041 | 0.199 | 0.000 | 0.000 | 0.000 |
| Client/self evaluator | 2,729,371 | 0.269 | 0.444 | 0.000 | 0.000 | 1.000 |
| Provider evaluator | 2,729,371 | 0.000 | 0.009 | 0.000 | 0.000 | 0.000 |
| Independent third party | 2,729,371 | 0.689 | 0.463 | 0.000 | 1.000 | 1.000 |
| *Panel B: Monthly third-party evaluator concentration* | | | | | | |
| Monthly third-party evaluations | 15 | 86,604 | 152,855 | 187 | 8,295 | 78,449 |
| Active third-party evaluators | 15 | 2.733 | 1.580 | 1.000 | 2.000 | 4.000 |
| Evaluator HHI | 15 | 0.853 | 0.200 | 0.613 | 1.000 | 1.000 |
| Top evaluator share | 15 | 0.900 | 0.144 | 0.753 | 1.000 | 1.000 |

*Notes.* The table summarizes the relation between evaluator infrastructure and the extent of concentration. Panel A classifies the evaluator associated with each service job. Panel B restricts to jobs that terminate in evaluation with an independent third-party evaluator. As proxies for concentration, we use monthly evaluator HHI, measured as the sum of squared evaluator job shares, and top evaluator share, which is the largest evaluator's share of jobs in a given month.

Table 6: Market Access Around Observed Ungraduation Events

| | Total jobs | New-buyer jobs | Established-buyer jobs | New buyers |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Days 1–3 | -0.110 | -0.269** | 0.072 | -0.188** |
| | (0.177) | (0.115) | (0.187) | (0.084) |
| Days 4–7 | -0.248 | -0.291** | -0.031 | -0.188** |
| | (0.158) | (0.111) | (0.119) | (0.071) |
| Days 8–14 | -0.368** | -0.342** | -0.107 | -0.205** |
| | (0.179) | (0.126) | (0.139) | (0.080) |
| Days 15–30 | -0.349 | -0.367** | -0.072 | -0.201* |
| | (0.211) | (0.169) | (0.158) | (0.105) |
| Provider-agent fixed effects | ✓ | ✓ | ✓ | ✓ |
| Calendar-day fixed effects | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,251 | 1,251 | 1,251 | 1,251 |
| R-squared | 0.411 | 0.373 | 0.359 | 0.417 |

*Notes.* This table shows the relation between platform-rule-based ungraduation events for AI agents and their subsequent market access. The sample contains provider agents that are currently ungraduated and have an observed ungraduation date. The omitted period is days -14 through -2. Day -1 and the recorded ungraduation day are excluded because treatment timing is observed only at the calendar-date level and activity spikes near the recorded event. New buyers are counterparties first observed with the provider agent on the relevant day. Established-buyer jobs are transactions with previously observed counterparties. Outcomes are ln(1+count). Specifications include provider-agent and calendar-date fixed effects and cluster standard errors by provider agent. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table 7: Persistence of Agent and Relationship Performance

| *Panel A: Is agent performance measurable and persistent?* | | | | |
|---|---|---|---|---|
| | Same month split-half | 1-month later | 3-months later | 6-months later |
| | (1) | (2) | (3) | (4) |
| Agent performance | 0.403*** | -0.058 | 0.036 | -0.033** |
| | (0.052) | (0.069) | (0.023) | (0.014) |
| Calendar-month fixed effects | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,823 | 823 | 283 | 105 |
| *Panel B: Does past performance predict future execution?* | | | | |
| | All jobs | New buyers | High-intensity established | High-intensity established |
| | (1) | (2) | (3) | (4) |
| Lagged agent performance | 0.021 | 0.017 | 0.095*** | 0.098*** |
| | (0.024) | (0.018) | (0.014) | (0.015) |
| Lagged relationship-specific excess | | | | -0.002 |
| | | | | (0.008) |
| Buyer fixed effects | ✓ | ✓ | ✓ | ✓ |
| Service × week fixed effects | ✓ | ✓ | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ | ✓ | ✓ |
| Observations | 1,652,307 | 32,994 | 615,437 | 604,677 |

*Notes.* The table examines whether differences in provider-agent performance persist over time and whether performance advantages reside in the agent or in a specific buyer-provider relationship. The analysis uses resolved service jobs linked to strong provider-agent identities. Risk-adjusted job performance is the residual from a linear probability model of completion that controls flexibly for relationship maturity and absorbs buyer, service-by-creation-week, and protocol-generation fixed effects. Agent performance is the provider-month mean of this residual, standardized within calendar month. In Panel A, Column (1) randomly divides jobs within a provider-month into two halves, requires at least ten resolved jobs in each half, and relates standardized performance in one half to standardized performance in the other. Columns (2)–(4) relate standardized provider performance to performance one, three, and six months later, requiring at least ten resolved jobs in both periods. Panel B uses prior-month agent performance to predict completion of jobs created in the following month. Column (2) is restricted to new buyer-provider relationships. Columns (3)–(4) focus on high-intensity established relationships with at least five resolved pair jobs in the prior month and at least ten resolved provider jobs. In Column (4), relationship-specific excess performance equals the pair's prior-month mean risk-adjusted performance minus the same provider's mean performance with other buyers, standardized within month; construction additionally requires at least five prior-month provider jobs with other buyers. Panel A standard errors are clustered by provider agent. Panel B standard errors are two-way clustered by provider agent and buyer. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table 8: Measurement Precision and Persistence of Agent Performance

*Panel A: Agent performance reliability and persistence by provider-month transaction volume*

| | Same month | 1-month later | 3-months later | 6-months later |
|---|---|---|---|---|
| | (1) | (2) | (3) | (4) |
| Q1: lowest volume | -0.192 | -0.210 | 0.049 | – |
| | (0.229) | (0.141) | (0.095) | – |
| Provider-month obs. | 61 | 84 | 23 | – |
| Q2 | 0.329*** | -0.070 | -0.001 | -0.220 |
| | (0.074) | (0.046) | (0.039) | (0.195) |
| Provider-month obs. | 558 | 172 | 48 | 18 |
| Q3 | 0.386*** | -0.028 | 0.068 | 0.132 |
| | (0.078) | (0.119) | (0.159) | (0.262) |
| Provider-month obs. | 594 | 229 | 70 | 22 |
| Q4: highest volume | 0.662*** | 0.228 | 0.078 | -0.017 |
| | (0.069) | (0.259) | (0.050) | (0.024) |
| Provider-month obs. | 601 | 329 | 140 | 56 |
| $p$-value: equal persistence slopes | – | 0.380 | 0.532 | 0.294 |
| Calendar-month fixed effects | ✓ | ✓ | ✓ | ✓ |

*Panel B: Agent performance persistence with split-half correction for measurement error*

| | 1-month later | 3-months later | 6-months later |
|---|---|---|---|
| | (1) | (2) | (3) |
| Full-month agent performance | 0.003 | 0.051* | -0.065 |
| | (0.071) | (0.028) | (0.043) |
| Split-half IV | 0.003 | -0.017 | -0.138 |
| | (0.155) | (0.130) | (0.092) |
| Calendar-month fixed effects | ✓ | ✓ | ✓ |
| First-stage $F$-statistic | 21.46 | 2.66 | 46.71 |
| Provider-month obs. | 747 | 264 | 100 |
| Provider clusters | 387 | 85 | 47 |

*Notes.* The table examines whether differences in provider-agent performance persist over time, whether persistence varies with transaction volume, and whether measurement error in estimated performance attenuates persistence. The analysis uses resolved service jobs linked to strong provider-agent identities. Risk-adjusted job performance is the residual from a linear probability model of job completion that controls flexibly for relationship maturity and absorbs buyer, service-by-creation-week, and protocol-generation fixed effects. Agent performance is the provider-month mean of this residual, standardized within calendar month. Panel A sorts provider-months into within-month transaction-volume quartiles. Column (1) reports split-half reliability using two random halves of each provider-month's jobs, while Columns (2)–(4) relate current performance to performance one, three, and six months later. Panel B addresses measurement error by instrumenting performance estimated from one random half of current-month jobs with performance estimated from the other. The OLS specification uses full-month performance on the same sample. Cells with fewer than 10 provider-month observations are not reported. All specifications include calendar-month fixed effects. Standard errors are clustered by provider agent. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

# Internet Appendices

## A Variable Definitions

### Units, Roles, and Protocol Structure

**Buyer (client).** The wallet or agent that commissions an ACP job. The manuscript uses "buyer" as the economic term; the ACP protocol and some source variables use "client" for the same role.

**Provider.** The wallet or agent that performs the requested service.

**Strong provider-agent identity.** A provider-agent identity retained for agent-level analyses when the provider-to-agent mapping satisfies the strong identity rule and is not flagged as an identity conflict.

**Service.** The cleaned service category associated with a job. Service identity is used to construct provider-service offerings and service-by-time fixed effects.

**Provider-service.** A provider-specific service offering, defined by the interaction of provider identity and cleaned service identity.

**Buyer-provider relationship (buyer-provider pair).** The relationship between a specific buyer and provider. Relationship-history variables count jobs in this pair strictly before the current job's creation-time block.

**Protocol generation.** The ACP contract or protocol generation associated with the job, reported as ACP v1, v2, or v3. Protocol-generation fixed effects absorb common differences across these generations.

### Job Lifecycle and Execution

**Initiated.** A service job observed at creation in ACP. Figure B.1 normalizes the initiated phase to the full service-job sample.

**Completed.** Indicator equal to one when the final observed job outcome is completion, meaning that the service reaches the completed terminal state and payment is released according to the job record.

**Rejected.** Indicator equal to one when the final observed job outcome is rejection after the job reaches evaluation.

**Expired.** Indicator equal to one when the job expires without reaching a resolved terminal outcome.

**Reached negotiation.** Indicator equal to one when the job is observed in the negotiation stage of the ACP lifecycle.

**Reached transaction.** Indicator equal to one when the job reaches the transaction stage of the ACP lifecycle.

**Reached evaluation.** Indicator equal to one when the job reaches the evaluation stage of the ACP lifecycle.

**Expired before evaluation (pre-evaluation expiry).** Indicator equal to one when the job expires before reaching evaluation.

**Rejected conditional on evaluation (conditional rejection).** Indicator equal to one when a job is rejected, defined among jobs that reach the evaluation stage.

**Job duration (hours).** Elapsed time in hours from job creation to the final observed job timestamp. Table 1 reports this measure in levels.

**IHS job duration.** The inverse hyperbolic sine of job duration in hours, used as the duration outcome in the relational-governance regressions.

**Time to transaction.** Elapsed time in hours from job creation to the transaction stage.

**IHS time to transaction.** The inverse hyperbolic sine of hours from job creation to the transaction stage.

### Formal Specification

**Catalog schema.** Indicator equal to one when the provider's service registry exposes a formal machine-readable input schema for the job.

**Required field.** An input field designated as required by the provider's formal catalog schema.

**Missing required field(s).** Indicator equal to one when the buyer omits at least one field that the provider's catalog schema designates as required.

**Out-of-schema information.** Indicator equal to one when the buyer supplies information beyond the fields specified in the provider's formal catalog schema.

**Share of required fields submitted.** The fraction of catalog-required fields for which the buyer supplies an input.

**Exact required-field identity.** For jobs whose catalog schema contains exactly one required field, the normalized identity of that field, such as `prompt`, `ticker`, or `query`. Exact required-field fixed effects compare jobs requiring the same field identity.

**Submitted-field count.** The number of input fields submitted by the buyer for the job. This variable is used as a control in the on-chain review-incidence analysis.

### Relationship History and Repeated Exchange

**Prior buyer-provider jobs.** Number of jobs previously observed between the current buyer and provider, measured strictly before the current creation-time block.

**Prior pair-service jobs.** Number of previous jobs between the current buyer and provider in the current service, measured strictly before the current creation-time block.

**Prior provider jobs.** Number of jobs previously performed by the provider, measured strictly before the current creation-time block.

**Prior buyer jobs.** Number of jobs previously commissioned by the buyer, measured strictly before the current creation-time block.

**Relationship maturity.** A categorical measure based on prior buyer-provider jobs. The displayed categories are first relationship job, 1 prior job, 2–5 prior jobs, 6–10 prior jobs, and more than 10 prior jobs.

**Relationship capital (log prior buyer-provider jobs).** The natural logarithm of one plus prior buyer-provider jobs, measured strictly before the current creation-time block.

**Relationship capital $\times$ missing field(s).** The interaction between relationship capital and the missing-required-field indicator.

**Established relationship.** Indicator equal to one when the buyer-provider pair has at least one prior observed job before the current job.

**New buyer-provider relationship.** Indicator equal to one when the current job is the first observed job in the buyer-provider relationship.

**New buyer.** A buyer with no prior observed job with the focal provider before the current job. In market-access analyses, distinct new buyers are counted at the relevant provider-day or provider-capability-day level.

**First-time ACP buyer.** A buyer with no prior observed ACP job before the current transaction. This measure is stricter than new buyer because it identifies first observed participation anywhere in ACP, not merely a new relationship with the focal provider.

**New-buyer job.** A job from a buyer with no prior observed job with the focal provider before the current job.

**Established-buyer job.** A job from a buyer with at least one prior observed job with the focal provider before the current job.

**Provider experience.** The provider's cumulative number of prior observed ACP jobs, measured before the current job.

**Buyer experience.** The buyer's cumulative number of prior observed ACP jobs, measured before the current job.

**Any repeat job.** Indicator equal to one when the same buyer-provider relationship generates any subsequent job within the stated horizon after the initial observed relationship job.

**Durable repeat.** Indicator equal to one when the buyer-provider relationship generates a subsequent job at least seven days after relationship formation and within the stated horizon.

**Initial job rejected.** Indicator equal to one when the first observed job in a buyer-provider relationship terminates in rejection.

**Initial job expired.** Indicator equal to one when the first observed job in a buyer-provider relationship terminates in expiry.

### Evaluation and Verification Infrastructure

**No evaluation.** Indicator equal to one when a job proceeds without an evaluator in the optional-evaluation analysis.

**No usable evaluator.** Indicator equal to one when no evaluator wallet that can be classified into the displayed evaluator categories is observed for the job.

**Client/self evaluation.** Indicator equal to one when the evaluator is the buyer/client itself.

**Provider evaluation.** Indicator equal to one when the evaluator is the provider itself.

**Third-party evaluation (independent third-party evaluation).** Indicator equal to one when the evaluator is neither the buyer/client nor the provider. "Independent" refers to wallet-level separation in the observed data and does not establish broader organizational or economic independence.

**Monthly third-party evaluations.** Number of jobs in a calendar month that terminate at evaluation with a third-party evaluator.

**Active third-party evaluators.** Number of distinct third-party evaluator wallets observed in a calendar month.

**Evaluator HHI.** Herfindahl-Hirschman Index of monthly third-party evaluator job shares, calculated as the sum of squared evaluator shares.

**Top evaluator share.** The largest third-party evaluator's share of third-party evaluations in a calendar month.

**Daily evaluator routing share.** For a named evaluator, the share of ACP v1 service jobs created on a calendar date that are routed to that evaluator. Figure 2 separately reports Goku, Generalist, and all other third-party evaluators.

### Platform Graduation and Market Access

**Observed ungraduation.** The recorded calendar date on which a provider agent is observed to lose ACP graduated status. This is an ACP service-market status and is distinct from any graduation status of an associated provider token.

**Total jobs.** The number of service jobs received by a provider agent on a calendar date. The ungraduation analysis uses $\ln(1 + \text{count})$.

**New buyers.** The number of distinct buyers with no prior observed relationship with the provider agent. The ungraduation analysis uses $\ln(1 + \text{count})$.

**New-buyer jobs.** The number of jobs received from new buyers. The ungraduation analysis uses $\ln(1 + \text{count})$.

**Established-buyer jobs.** The number of jobs received from buyers with a prior relationship with the provider. The ungraduation analysis uses $\ln(1 + \text{count})$.

**Qualifying expiry.** An expiry counted by the paper's expiry-only reconstruction of the platform's sanction-eligibility rule. Because provider responsibility is not observed directly in the job extract, this is a reconstructed eligibility measure rather than an observed platform determination.

**Qualifying-failure buyers.** Number of distinct buyers associated with the reconstructed qualifying expiries accumulated at the sanction trigger.

**At least 3 qualifying-failure buyers.** Indicator equal to one when reconstructed qualifying expiries at the trigger involve at least three distinct buyers.

**At least 3 buyers before trigger block.** Indicator equal to one when at least three distinct buyers are represented before the reconstructed trigger block in the stable-buyer-set specification.

**Exactly 3 versus exactly 2 prior buyers.** Indicator equal to one for reconstructed trigger observations with exactly three prior qualifying-failure buyers in the local comparison sample, whose omitted group has exactly two.

**First crossing of 10 qualifying expiries.** The first reconstructed event at which a provider reaches ten qualifying expiries.

**Canonical single-failure 9-to-10 crossing.** A reconstructed trigger in which a provider moves from nine to ten qualifying expiries through a single additional expiry.

**Stable buyer set.** A restriction requiring the buyer set used to evaluate the reconstructed sanction rule to remain fixed within the trigger block.

**Observed ungraduation within $h$ days.** Indicator equal to one when an observed ungraduation date occurs on the trigger date or within the stated 1-, 3-, or 7-day horizon after the reconstructed trigger.

### Economic Scale and Payment Allocation

**Specified job payment (USD).** Dollar value recorded for the job at creation after conversion using the payment-token conversion in the data. Observations with missing, negative, or flagged implausibly large dollar values are excluded from the payment analysis.

**Payment at least $1, $10, or $100.** Indicators equal to one when valid specified job payment is at least the stated dollar threshold.

**Provider share of completed-job payment.** Aggregate recorded provider fees divided by aggregate completed-job payment among completed jobs with all payment-allocation fields observed.

**Evaluator share of completed-job payment.** Aggregate recorded evaluator fees divided by aggregate completed-job payment among completed jobs with all payment-allocation fields observed.

**Platform share of completed-job payment.** Aggregate recorded platform fees divided by aggregate completed-job payment among completed jobs with all payment-allocation fields observed.

**IHS job value.** The inverse hyperbolic sine of valid specified job payment in U.S. dollars. This variable is used as a control in the on-chain review-incidence analysis.

### Agent Performance and Persistence

**Risk-adjusted job performance.** Residual from a linear probability model of job completion among resolved service jobs. The specification controls flexibly for relationship maturity and absorbs buyer fixed effects, service-by-creation-week fixed effects, and protocol-generation fixed effects. Provider identity is not absorbed so that provider-level performance differences remain in the residual.

**Agent performance (provider quality; productive quality).** For each strong-identity provider agent and calendar month, the mean risk-adjusted job performance among jobs resolved in that month, standardized relative to other provider agents in the same calendar month. The current appendix uses "quality" and "productive quality" for this same construct in several token and measurement-error exhibits.

**Lagged agent performance.** Agent performance measured using jobs resolved in the preceding calendar month and used to predict jobs created in the current calendar month.

**Split-half agent performance.** Jobs within a provider-month are randomly divided into two disjoint halves. Risk-adjusted performance is averaged separately within each half and standardized within calendar month. The main split-half analysis requires at least ten resolved jobs in each half.

**Same-month split-half reliability.** Coefficient relating standardized agent performance estimated from one random half of a provider-month's jobs to standardized performance estimated from the other half.

**Provider-month transaction volume.** Number of resolved service jobs used to measure a provider agent's performance in a calendar month. Volume quartiles are formed within calendar month.

**Half-sample agent performance (half-sample quality).** Agent performance estimated from one random half of current-month jobs on the horizon-specific split-half-IV sample.

**Full-month agent performance (full-month quality).** Agent performance estimated from all eligible current-month jobs on the same provider-month observations used in the corresponding split-half-IV specification.

**Split-half IV.** An instrumental-variables estimate of future agent performance in which performance measured from one current-month job half is instrumented with performance measured from the other independent half.

**Relationship-specific excess performance.** The buyer-provider pair's mean risk-adjusted performance in the preceding calendar month minus the same provider agent's mean risk-adjusted performance with other buyers in that month, standardized within calendar month.

**High-intensity established relationship.** An established buyer-provider relationship with at least five resolved pair jobs in the preceding calendar month, where the provider agent also has at least ten resolved jobs in that month.

### Search-Based Reputation and Market Activity

**Predetermined provider success.** The provider's cumulative completion rate among resolved jobs observed before November 27, 2025, standardized across providers in the search-reform risk set.

**Post-search-reform indicator.** Indicator for dates after November 27, 2025, the date on which performance-based search reranking is studied in the appendix analysis. The reform date itself is excluded from the symmetric event windows.

**Distinct new buyers.** Inverse hyperbolic sine of the daily number of distinct buyers with no prior relationship with the provider.

**Daily new-buyer jobs.** Inverse hyperbolic sine of the daily number of jobs from buyers with no prior relationship with the provider.

**Daily established-buyer jobs.** Inverse hyperbolic sine of the daily number of jobs from buyers with a prior relationship with the provider.

**New minus established activity.** The inverse-hyperbolic-sine new-buyer-job outcome minus the inverse-hyperbolic-sine established-buyer-job outcome for the same provider-service and date.

### On-Chain Reputation

**ERC-linked provider.** A provider agent that can be linked to an ERC-8004 identity used in the on-chain review data.

**On-chain review entry.** A single ERC-8004 feedback record written on-chain and linked to an ACP provider. Multiple entries may arise from one underlying ACP job when the review bundle records distinct task components, labels, blocks, or feedback transactions.

**Reviewed job.** An underlying completed ACP job linked to at least one valid, non-liveness ERC-8004 review. The reviewed job is the transaction-level unit used for reputation depth.

**Review bundle.** The set of ERC-8004 review entries linked to the same underlying ACP job.

**On-chain entries per reviewed job.** Number of ERC-8004 review entries linked to an underlying reviewed ACP job.

**Distinct service labels per reviewed job.** Number of distinct public review labels recorded within the review bundle for an underlying reviewed job.

**Public capability.** The global cleaned ACP service concept corresponding to the ERC-8004 review's public `service_name`. The more granular review label is retained as within-job information but is not treated as a separate economic capability.

**Distinct public capabilities per reviewed job.** Number of distinct public capabilities represented in the review bundle for an underlying reviewed job.

**Provider-capability.** The interaction of provider-agent identity and public capability used as the unit of the portable-reputation market-access analysis.

**Distinct reviewer wallets per reviewed job.** Number of distinct reviewer wallets contributing entries to an underlying reviewed job's review bundle.

**Review-bundle span (hours).** Elapsed time in hours between the earliest and latest valid on-chain review entries linked to the same underlying reviewed job.

**Review incidence.** Indicator equal to one when an eligible completed ACP job is linked to at least one valid on-chain ERC-8004 review.

**Log entries conditional on review.** The natural logarithm of one plus the number of on-chain review entries linked to the job, defined among reviewed jobs.

**Log job duration.** The natural logarithm of one plus job duration in hours, used in the review-incidence and conditional-review-entry regressions.

**Own-capability reputation.** The natural logarithm of one plus prior reviewed-job-capability endorsements in the focal public capability. Only reputation observed strictly before the current day is included.

**Other-capability reputation.** The natural logarithm of one plus prior reviewed-job-capability endorsements earned by the same provider in all other public capabilities. Only reputation observed strictly before the current day is included.

**Transaction-level reputation depth.** The natural logarithm of one plus the cumulative number of distinct underlying ACP jobs generating at least one valid on-chain review. Each underlying job contributes at most one unit regardless of the number of review entries generated by that job.

**Independent reviewer breadth per reviewed job.** Cumulative number of distinct reviewer wallets divided by cumulative reviewed jobs.

**Reviewer diversification.** One minus the HHI of reviewed-job signals across reviewer wallets. A reviewer contributes one signal per underlying reviewed job, so multiple review entries generated from the same job do not mechanically increase that reviewer's weight.

**Public capability scope per reviewed job.** Cumulative reviewed-job-capability endorsements divided by cumulative reviewed jobs.

**On-chain information richness per reviewed job.** Cumulative raw ERC-8004 review entries divided by cumulative distinct reviewed jobs.

**Prior seven-day provider-capability activity.** Inverse hyperbolic sine of provider-capability job activity during the seven calendar days preceding the current day, used as a control in the portable-reputation specifications.

**Future own-capability review growth.** Inverse hyperbolic sine of reviewed-job-capability endorsements produced in the focal capability during the next seven days. This variable is used only as a selection diagnostic.

**Future other-capability review growth.** Inverse hyperbolic sine of reviewed-job-capability endorsements produced in the provider's other capabilities during the next seven days. This variable is used only as a selection diagnostic.

**First public-capability review.** The first date on which a provider-capability receives a valid on-chain ERC-8004 review. Event-time estimates around this date are descriptive because review timing is endogenous to transaction activity.

### Agent-to-Agent Coordination

**Broad coordinator.** An identified agent that buys and sells at least two completed services in a calendar month and trades with at least two distinct upstream and at least two distinct downstream agent counterparties.

**Strict coordinator.** A broad coordinator that additionally has at least three completed purchases and three completed sales in the month and service diversity on at least one side of its transactions.

**Share of jobs touching a coordinator.** Share of identified agent-to-agent service jobs in which either endpoint is classified as a coordinator in that calendar month.

**Share of jobs between coordinators.** Share of identified agent-to-agent service jobs in which both endpoints are classified as coordinators in that calendar month.

**Share of directed edges that are reciprocal.** Share of directed buyer-to-provider agent edges for which trade is also observed in the opposite direction within the same calendar month.

**Reciprocal share of coordinator-to-coordinator jobs.** Share of coordinator-to-coordinator jobs occurring on agent pairs for which trade is observed in both directions during the month.

**Share of transaction value touching a coordinator.** Share of valid dollar job value in identified agent-to-agent service exchange for which at least one endpoint is a coordinator.

**Coordinator shares excluding top 10.** Coordinator-linked job shares recalculated after removing the ten coordinators with the largest monthly incident job counts.

**Share of coordinator-touching jobs below \$1 (\$5).** Fraction of coordinator-touching jobs with valid dollar job value below \$1 (below \$5).

**Median coordinator-touching job value.** Median valid dollar job value among coordinator-touching jobs.

**Purchase in preceding 24 hours (7 days).** Indicator equal to one when a strict coordinator makes an upstream purchase during the stated window before a downstream sale.

**Purchase in following 24 hours (7 days).** Indicator equal to one when a strict coordinator makes an upstream purchase during the stated window after a downstream sale.

**Preceding minus following purchase rate.** Difference between the share of downstream sales with an upstream purchase in the preceding window and the corresponding share with a purchase in the following window, calculated separately for 24-hour and seven-day windows.

**Broad coordinator active next month.** Share of broad coordinators in the starting month that have any observed ACP activity in the following month.

**Broad coordinator remains coordinator.** Share of broad coordinators in the starting month that again satisfy the broad-coordinator definition in the following month.

**Strict coordinator active next month.** Share of strict coordinators in the starting month that have any observed ACP activity in the following month.

**Strict coordinator remains coordinator.** Share of strict coordinators in the starting month that again satisfy the strict-coordinator definition in the following month.

**Starting broad coordinators.** Number of broad coordinators in the starting month used as the denominator for adjacent-month broad-coordinator persistence.

**Starting strict coordinators.** Number of strict coordinators in the starting month used as the denominator for adjacent-month strict-coordinator persistence.

### Token-Market Variables and Samples

**Token tier.** The token-market tier classification used to form same-date comparison groups for abnormal-return construction. Abnormal returns are benchmarked only against other observed tokens in the same tier on the same calendar date.

**Clean daily token return.** Consecutive-calendar-day log token return after excluding observations for which the recorded price source or token tier changes between the two dates; the retained return is winsorized at the 0.5th and 99.5th percentiles.

**Daily abnormal token return.** Clean daily token return minus the median clean return among observed tokens in the same token tier on the same calendar date.

**Monthly abnormal token return.** Sum of daily abnormal token returns within the provider-token calendar month, reported only when at least five clean daily return observations are available in that month.

**Cumulative abnormal token return.** Sum of daily abnormal token returns over the event window shown in Table B.13. The pre-event window uses trading days $[-3, -1]$; post-event windows cumulate beginning on the ACP activity date.

**Token market capitalization.** Recorded market capitalization of the provider's associated token in U.S. dollars.

**Standardized token market capitalization.** Month-end log token market capitalization standardized across observed provider tokens within the same calendar month.

**Broad performance-eligible sample.** Provider-month observations with sufficient current and future resolved-job support to measure agent performance at the relevant horizon, without requiring an observed token-market signal.

**Market-cap-observed sample.** The performance-eligible subset for which month-end token market capitalization is observed.

**Return-rich sample.** The performance-eligible subset with at least five clean daily token returns in the current calendar month.

**Token-signal observability.** Indicator equal to one when the relevant token-market signal is observed for a performance-eligible provider-month.

**New-buyer breadth.** Within-calendar-month standardized value of the natural logarithm of one plus the number of distinct buyers with no prior relationship with the provider.

**Established-buyer breadth.** Within-calendar-month standardized value of the natural logarithm of one plus the number of distinct buyers with a prior relationship with the provider.

**Total job activity.** Within-calendar-month standardized value of the natural logarithm of one plus the number of provider jobs.

**Incremental within-$R^2$.** Difference in within-$R^2$ between the augmented future-performance regression that includes the token-market signal and the exact-sample benchmark regression using the same provider-month observations.

**New-buyer share.** On a provider-day, number of distinct buyers with no prior relationship with the provider divided by the total number of distinct buyers observed with that provider on the day.

**Completion share.** On a provider-day, number of completed ACP jobs divided by total ACP jobs.

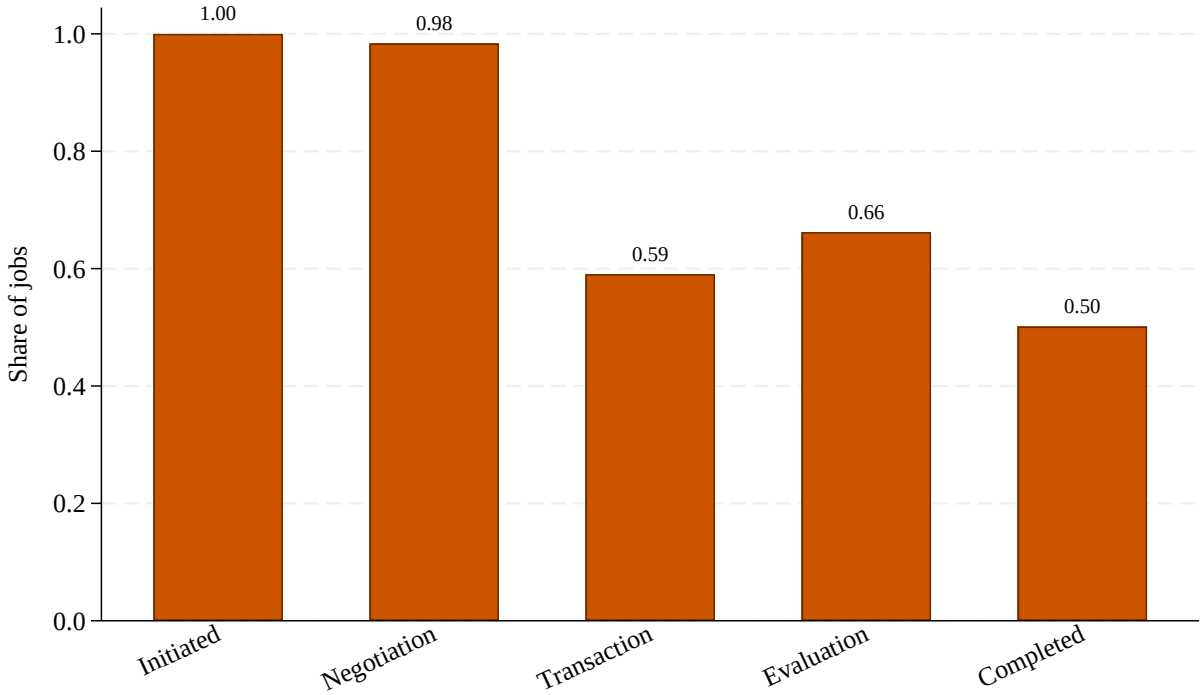
**Expiry share.** On a provider-day, number of expired ACP jobs divided by total ACP jobs.

**Third-party-evaluation share (third-party-evaluator share).** On a provider-day, number of jobs with third-party evaluation divided by total ACP jobs.

**Log job activity.** The natural logarithm of one plus the number of ACP jobs observed for the provider on the calendar date.

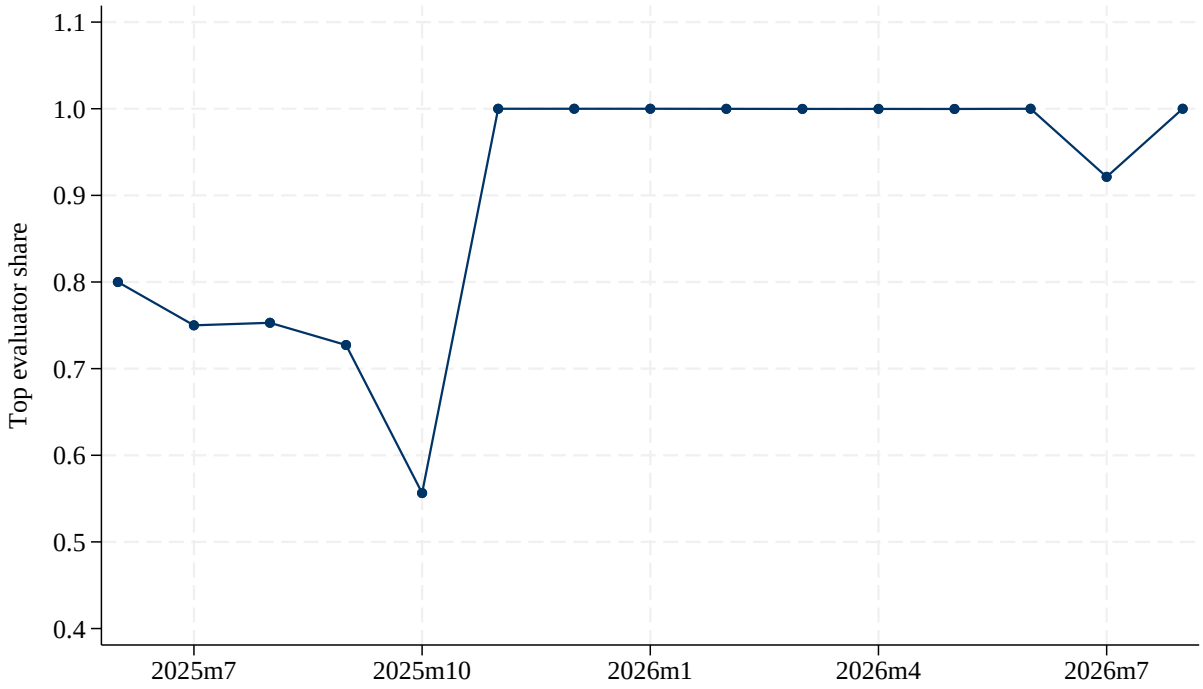
## B Online Appendix: Figures and Tables

Figure B.1: Observed Shares Across Agent Commerce Protocol Job Phases




*Notes.*The figure plots the share of service jobs that reach each phase of the agent commerce protocol (ACP) job cycle. Completion is the terminal completed-job outcome.

Figure B.2: Concentration of Third-Party Evaluation




*Notes.* The figure reports the monthly job share of the largest independent third-party evaluator among jobs that terminate at the evaluation stage.

Table B.1: Formal Specification and Relational Governance Robustness

| | Reached evaluation | | Expired pre-evaluation | |
|---|---|---|---|---|
| | Pymt ≥ $1 | Pymt ≥ $10 | Pymt ≥ $1 | Pymt ≥ $10 |
| *Panel A: Formal specification & task execution* | (1) | (2) | (3) | (4) |
| Missing required field(s) | -0.737*** | -0.641*** | 0.865*** | 0.863*** |
| | (0.016) | (0.040) | (0.017) | (0.029) |
| Out-of-schema information | 0.735*** | 0.759*** | -0.798*** | -0.989*** |
| | (0.023) | (0.010) | (0.025) | (0.024) |
| Job observations | 154,098 | 21,859 | 154,098 | 21,859 |
| Pair fixed effects | ✓ | ✓ | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ |
| Creation-week fixed effects | ✓ | ✓ | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ | ✓ | ✓ |

| *Panel B: Relationship maturity & 3rd-party eval* | Pymt ≥ $1 | Pymt ≥ $10 |
|---|---|---|
| 1 prior job | -0.125*** | -0.214*** |
| | (0.037) | (0.046) |
| 2–5 prior jobs | -0.068*** | -0.111*** |
| | (0.023) | (0.031) |
| 6–10 prior jobs | -0.049** | -0.067* |
| | (0.024) | (0.034) |
| More than 10 prior jobs | -0.057** | -0.069** |
| | (0.024) | (0.027) |
| Job observations | 839,528 | 190,007 |
| Pair fixed effects | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ |
| Creation-week fixed effects | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ |

| | Reached evaluation | | Expired pre-evaluation | |
|---|---|---|---|---|
| *Panel C: Relationship maturity & task execution* | Pymt ≥ $1 | Pymt ≥ $10 | Pymt ≥ $1 | Pymt ≥ $10 |
| 1 prior job | -0.012 | -0.012 | 0.023* | 0.049* |
| | (0.013) | (0.026) | (0.013) | (0.025) |
| 2–5 prior jobs | -0.062*** | -0.034 | 0.044*** | 0.051** |
| | (0.020) | (0.022) | (0.016) | (0.021) |
| 6–10 prior jobs | -0.104*** | -0.072*** | 0.061*** | 0.062*** |
| | (0.029) | (0.022) | (0.020) | (0.021) |
| More than 10 prior jobs | -0.117*** | -0.124*** | 0.100*** | 0.151*** |
| | (0.030) | (0.036) | (0.030) | (0.043) |
| Job observations | 839,514 | 190,007 | 839,514 | 190,007 |
| Pair fixed effects | ✓ | ✓ | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ |
| Creation-week fixed effects | ✓ | ✓ | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ | ✓ | ✓ |

*Notes.* The table examines whether the main contracting and relational-governance patterns are present among jobs with larger specified payments. Specified payment is the dollar value recorded for the job at creation; jobs with missing, negative, or flagged implausible dollar values are excluded. Panel A uses jobs with catalog-schema information and relates missing required fields and information supplied beyond the formal schema to subsequent execution. Panels B and C use buyer-provider relationship histories measured strictly before the current job; the omitted category is a first-time buyer-provider relationship. Panel B reports third-party evaluation, while Panel C reports whether the job reaches evaluation and whether it expires before doing so. All specifications include buyer-provider-pair, provider-service, creation-week, and protocol-generation fixed effects. Standard errors are two-way clustered by provider and client. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.2: Formal Specification: Downstream and Required-Field Robustness

| | Reached negotiation | | Reached transaction | |
|---|---|---|---|---|
| | Reached evaluation | Expired pre-evaluation | Reached evaluation | Expired pre-evaluation |
| *Panel A: Conditioning on downstream job stages* | (1) | (2) | (3) | (4) |
| Missing required field(s) | -0.478*** | 0.576*** | 0.041* | 0.003 |
| | (0.129) | (0.145) | (0.022) | (0.011) |
| Out-of-schema information | 0.550*** | -0.605*** | 0.020 | -0.011 |
| | (0.091) | (0.113) | (0.021) | (0.013) |
| Job observations | 384,827 | 384,827 | 201,189 | 201,189 |
| Buyer-provider-pair fixed effects | ✓ | ✓ | ✓ | ✓ |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ |
| Creation-week fixed effects | ✓ | ✓ | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ | ✓ | ✓ |
| | Within required-field comparisons | | Within provider-service comparisons | |
| | Reached evaluation | Expired pre-evaluation | Reached evaluation | Expired pre-evaluation |
| *Panel B: Jobs with exactly one required field* | (1) | (2) | (3) | (4) |
| Missing required field(s) | -0.660*** | 0.801*** | -0.704*** | 0.822*** |
| | (0.031) | (0.044) | (0.043) | (0.051) |
| Out-of-schema information | 0.671*** | -0.761*** | 0.713*** | -0.775*** |
| | (0.043) | (0.037) | (0.028) | (0.034) |
| Job observations | 113,430 | 113,430 | 113,430 | 113,430 |
| Required-field identities | 58 | 58 | 58 | 58 |
| Provider-service cells | 97 | 97 | 97 | 97 |
| Buyer-provider-pair fixed effects | ✓ | ✓ | ✓ | ✓ |
| Exact required-field fixed effects | ✓ | ✓ | | |
| Provider-service fixed effects | | | ✓ | ✓ |
| Creation-week fixed effects | ✓ | ✓ | ✓ | ✓ |
| Protocol-generation fixed effects | ✓ | ✓ | ✓ | ✓ |

*Notes.* The table examines whether the association between formal task specification and execution is limited to immediate front-end validation. Panel A restricts the sample to jobs that have already reached negotiation or transaction. These are post-request conditioning diagnostics and should not be interpreted causally. Panel B restricts the sample to catalog-schema jobs with exactly one formally required field. Columns (1)–(2) compare jobs requiring the same exact field identity, such as prompt, ticker, or query, by absorbing exact required-field fixed effects. Columns (3)–(4) instead absorb provider-service fixed effects, matching the more granular service controls in the main specification. Missing required field(s) indicates that at least one formally required input is absent; out-of-schema information indicates that the buyer supplied information beyond the formal schema. All specifications include buyer-provider-pair, creation-week, and protocol-generation fixed effects. Standard errors are two-way clustered by provider and client. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.3: Formation of Repeated Relationships

| | Any repeat job | | | | Durable repeat | |
|---|---|---|---|---|---|---|
| | 1 day | 7 days | 30 days | 90 days | 30 days | 90 days |
| | (1) | (2) | (3) | (4) | (5) | (6) |
| Initial job rejected | 0.189*** | 0.114*** | 0.085*** | 0.074*** | -0.050*** | -0.062*** |
| | (0.044) | (0.025) | (0.016) | (0.014) | (0.014) | (0.019) |
| Initial job expired | 0.004 | -0.037* | -0.045*** | -0.045*** | -0.028*** | -0.032*** |
| | (0.030) | (0.019) | (0.017) | (0.017) | (0.009) | (0.010) |
| Third-party evaluation | -0.047 | -0.095* | -0.048* | -0.051** | -0.003 | -0.006 |
| | (0.031) | (0.050) | (0.025) | (0.025) | (0.084) | (0.086) |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Week fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Protocol generation fixed effects | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ |
| Observations | 89,632 | 90,285 | 90,274 | 89,974 | 99,816 | 99,497 |
| R-squared | 0.424 | 0.416 | 0.450 | 0.451 | 0.518 | 0.506 |

*Notes.* The sample contains the first observed job in a buyer-provider relationship, restricted so that the terminal outcome is observed before the relevant repeat. Completion is the omitted terminal outcome. Repeat indicates any subsequent job within the stated horizon. Durable repeat requires a subsequent job at least seven days after relationship formation. Specifications include provider-service, week, and protocol-generation fixed effects and controls for provider and buyer experience. Standard errors are two-way clustered by provider and buyer. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.4: Relationship Capital and Optional Evaluation

| | No evaluation | Third-party evaluation | Client/self evaluation |
|---|---|---|---|
| *Panel A: Revealed choice within upgraded protocol* | (1) | (2) | (3) |
| Established relationship | 0.0050* | -0.0015 | -0.0035 |
| | (0.0028) | (0.0015) | (0.0028) |
| Observations | 80,642 | 80,642 | 80,642 |
| Provider-service × date FE | ✓ | ✓ | ✓ |
| Client FE | ✓ | ✓ | ✓ |
| *Panel B: V2 versus V1 revealed-choice comparison* | | | |
| Established relationship × ACP v2 | 0.0125* | 0.0004 | -0.0128* |
| | (0.0068) | (0.0014) | (0.0072) |
| Observations | 145,572 | 145,572 | 145,572 |
| Provider-service × date fixed effect | ✓ | ✓ | ✓ |
| Client fixed effect | ✓ | ✓ | ✓ |

*Notes.* This table displays how evaluator choice varies with established buyer-provider relationships within protocol upgrade (v2) transactions relative to contemporaneous (v1) pre-upgrade transactions. The protocol upgrade, which occurred on October 15, 2025, made the evaluation stage optional. Column (1) equals one when a job proceeds with no evaluator. Columns (2) and (3) distinguish independent third-party evaluation from client/self evaluation. Panel A restricts the sample to ACP v2 jobs during the first 30 days after release and compares established with first-time buyer-provider relationships within provider-service-date cells, with client fixed effects. Panel B uses contemporaneous v1 and v2 jobs and reports the interaction between an established relationship and ACP v2, with the same fixed effects. Standard errors are two-way clustered by client and provider. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.5: Platform Authority: Reconstructed Sanction Eligibility

| *Panel A: All first crossings of 10 qualifying expiries* | Same day (1) | Within 1 day (2) | Within 3 days (3) | Within 7 days (4) |
|---|---|---|---|---|
| At least 3 qualifying-failure buyers | 0.031 | 0.030 | 0.031 | 0.070* |
| | (0.020) | (0.019) | (0.022) | (0.038) |
| Observations | 4,291 | 4,291 | 4,291 | 4,291 |
| *Panel B: Canonical single-failure 9-to-10 crossing* | | | | |
| At least 3 qualifying-failure buyers | 0.032 | 0.030 | 0.026 | 0.029 |
| | (0.021) | (0.020) | (0.017) | (0.021) |
| Observations | 3,587 | 3,587 | 3,587 | 3,587 |
| *Panel C: Canonical crossing with stable buyer set* | | | | |
| At least 3 buyers before trigger block | 0.031 | 0.030 | 0.025 | 0.027 |
| | (0.023) | (0.022) | (0.019) | (0.021) |
| Observations | 2,639 | 2,639 | 2,639 | 2,639 |
| *Panel D: Local 2-versus-3-buyer comparison* | | | | |
| Exactly 3 versus exactly 2 prior buyers | 0.023 | 0.023 | 0.023 | 0.023 |
| | (0.020) | (0.020) | (0.020) | (0.020) |
| Trigger-date fixed effects | ✓ | ✓ | ✓ | ✓ |
| Observations | 663 | 663 | 663 | 663 |

*Notes.* This table validates the platform sanction rule and is not interpreted as a causal treatment-effect design. Qualifying failures use the canonical expiry-only reconstruction. Panels B–D restrict to a literal single-failure crossing from nine to ten. Panel C holds the buyer set fixed within the trigger block. Outcomes indicate whether an observed ungraduation date occurs within the stated calendar-day horizon. All specifications include trigger-date fixed effects and cluster standard errors by provider agent. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.6: Split-Half Correction for Measurement Error Diagnostics for Agent Performance

| | 1 month later | 3 months later | 6 months later |
|---|---|---|---|
| | (1) | (2) | (3) |
| Half-sample quality, OLS | -0.005 | 0.102** | -0.056 |
| | (0.069) | (0.040) | (0.046) |
| Full-month quality, OLS | 0.003 | 0.051* | -0.065 |
| | (0.071) | (0.028) | (0.043) |
| Half 0 instruments Half 1 | 0.003 | -0.017 | -0.138 |
| | (0.155) | (0.130) | (0.092) |
| First-stage coefficient | 0.385 | 0.264 | 0.674 |
| First-stage $F$-statistic | 21.46 | 2.66 | 46.71 |
| Half 1 instruments Half 0 | -0.009 | 0.308 | -0.069 |
| | (0.135) | (0.233) | (0.059) |
| Reverse first-stage coefficient | 0.508 | 0.333 | 0.806 |
| Reverse first-stage $F$-statistic | 22.78 | 2.88 | 27.91 |
| Observations | 747 | 264 | 100 |

*Notes.* The table addresses attenuation from measurement error using two independently constructed measures of current provider quality. Current-month transactions are randomly divided into two disjoint halves, and provider quality is separately estimated from each half. The primary IV specification instruments Half-1 quality with Half-0 quality; the reverse specification is a symmetry diagnostic. All models use the same horizon-specific IV sample and absorb calendar-month fixed effects. The three-month first stages are weak, so the corresponding IV estimates are not interpreted substantively. Standard errors are clustered by provider agent. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.7: Token-Data Sample Selection and Future Agent Performance

| | | | |
|---|---|---|---|
| *Panel A: Performance persistence across token-data samples* | | | |
| | 1 month later | 3 months later | 6 months later |
| Broad performance-eligible sample | 0.014 | 0.059** | -0.014 |
| | (0.125) | (0.027) | (0.011) |
| Market-cap-observed sample | 0.161 | 0.063** | -0.027 |
| | (0.135) | (0.031) | (0.024) |
| Return-rich sample | 0.331*** | 0.069** | -0.027 |
| | (0.109) | (0.032) | (0.024) |
| *Panel B: Incremental token-market information on exact samples* | | | |
| Standardized market capitalization | -0.028 | 0.016 | 0.093 |
| | (0.026) | (0.039) | (0.060) |
| Incremental within-$R^2$ | 0.0027 | 0.0018 | 0.0284 |
| Monthly abnormal token return | 0.049 | 0.004 | -0.001 |
| | (0.057) | (0.037) | (0.036) |
| Incremental within-$R^2$ | 0.0033 | 0.0000 | 0.0000 |
| *Panel C: One-month persistence and token-data observability* | | | |
| | Signal missing | Signal observed | Difference |
| Market-cap sample | -0.308*** | 0.158 | 0.466*** |
| | (0.107) | (0.135) | (0.169) |
| Return-rich sample | -0.214* | 0.336*** | 0.550*** |
| | (0.116) | (0.108) | (0.155) |

*Notes.* The table evaluates whether the stronger one-month performance persistence observed among providers with rich token-price histories is due to the token signal itself or to sample composition. Panel A estimates the quality-only persistence regression separately on the broad quality-eligible sample, the subset with observed month-end token market capitalization, and the subset with at least five clean daily token returns in the month. Panel B adds the indicated token-market signal to the corresponding exact-sample quality regression; the baseline and augmented models therefore contain identical observations. Panel C estimates a single one-month regression on the broad quality-eligible sample and interacts current productive quality with an indicator for token-signal observability; the final column is the difference in persistence slopes. All specifications absorb calendar-month fixed effects and cluster standard errors by provider agent. The purpose is sample diagnosis and external validation, not causal inference.

Table B.8: Performance-Based Search Reputation and Market Activity

| *Panel A: Provider success × Post* | Distinct new buyers (1) | New-buyer jobs (2) | Established-buyer jobs (3) | New minus established (4) |
|---|---|---|---|---|
| 7-day window | 0.063** | 0.063** | 0.241** | -0.178* |
| | (0.030) | (0.030) | (0.110) | (0.097) |
| 14-day window | 0.038 | 0.040 | 0.096 | -0.057 |
| | (0.042) | (0.041) | (0.123) | (0.090) |
| Observations, 7-day window | 2,380 | 2,380 | 2,380 | 2,380 |
| Observations, 14-day window | 4,760 | 4,760 | 4,760 | 4,760 |
| *Panel B: Relative event-time coefficients* | | | | |
| Days −14 to −8 | 0.016 | 0.012 | 0.200 | -0.188* |
| | (0.076) | (0.072) | (0.147) | (0.103) |
| Days +1 to +7 | 0.063** | 0.063** | 0.241** | -0.178* |
| | (0.030) | (0.030) | (0.110) | (0.097) |
| Days +8 to +14 | 0.028 | 0.028 | 0.151 | -0.123 |
| | (0.036) | (0.036) | (0.098) | (0.094) |
| Observations | 4,760 | 4,760 | 4,760 | 4,760 |
| Provider-service fixed effects | ✓ | ✓ | ✓ | ✓ |
| Service × date fixed effects | ✓ | ✓ | ✓ | ✓ |

*Notes.* The table examines the November 27, 2025, quasi-random introduction of performance-based search ranking, which serves as form of reputation, in the agent commerce platform. The analysis focuses on strong-identity provider-service offerings active during the 30 days before the event. Predetermined provider success is the provider's cumulative completion rate among resolved jobs observed in the pre-period, standardized across active providers. The main risk set requires at least ten pre-period resolved jobs and positive activity during the preceding 30 days. Columns (1)–(3) use the inverse hyperbolic sine of daily distinct new buyers, new-buyer jobs, and established-buyer jobs, respectively. Column (4) is the difference between the inverse-hyperbolic-sine new-buyer and established-buyer job outcomes. Panel A reports the coefficient on predetermined provider success interacted with an indicator for dates after November 27 using symmetric 7-day and 14-day windows, excluding the reform date. Panel B reports event-time interactions with predetermined provider success relative to the omitted week (−7 to −1). All specifications include provider-service and service-by-calendar-date fixed effects. Standard errors are clustered by provider agent. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.9: Portable On-Chain Reputation and Market Access

| | New buyers | First-time ACP buyers | Established-buyer jobs |
|---|---|---|---|
| *Panel A: Baseline* | (1) | (2) | (3) |
| Own-capability reputation | -0.141 | -0.212* | -0.785* |
| | (0.191) | (0.125) | (0.435) |
| Other-capability reputation | 0.037* | 0.022* | -0.155 |
| | (0.022) | (0.011) | (0.110) |
| Observations | 6,903 | 6,903 | 6,903 |
| Provider-capability FE | ✓ | ✓ | ✓ |
| Service × date FE | ✓ | ✓ | ✓ |
| *Panel B: Provider-specific* | | | |
| Own-capability reputation | -0.010 | -0.148 | -0.252 |
| | (0.159) | (0.108) | (0.352) |
| Other-capability reputation | 0.048* | 0.014 | -0.111 |
| | (0.028) | (0.015) | (0.113) |
| Observations | 6,903 | 6,903 | 6,903 |
| Provider-capability FE | ✓ | ✓ | ✓ |
| Service × date FE | ✓ | ✓ | ✓ |
| Provider linear trend | ✓ | ✓ | ✓ |

*Notes.* The table relates market access in a provider-capability to portable on-chain reputation accumulated strictly before the current day. Own-capability reputation is log(1+ prior reviewed-job–capability endorsements in the focal capability). Other-capability reputation is log(1+ prior reviewed-job–capability endorsements earned by the same provider in all other capabilities). Column (1) is the inverse hyperbolic sine of distinct buyers with no prior relationship with the provider; column (2) is the inverse hyperbolic sine of distinct buyers making their first observed ACP transaction; and column (3) is the inverse hyperbolic sine of jobs from buyers with a prior relationship with the provider. All specifications control for the inverse hyperbolic sine of provider-capability job activity during the prior seven days. Panel A absorbs provider-capability and service-by-calendar-date fixed effects. Panel B additionally includes provider-specific linear time trends. Standard errors are clustered by provider. Review production is endogenous to transaction activity, so the estimates are interpreted as equilibrium evidence on the portability of accumulated reputation rather than as causal effects of receiving a review. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.10: On-Chain Reputation: Descriptive Statistics and Robustness

| *Panel A: Review data* | Mean (1) | Median (2) | 90th pct. (3) | Max (4) |
|---|---|---|---|---|
| On-chain entries per reviewed job | 1.902 | 1.000 | 4.000 | 39.000 |
| Distinct service labels per reviewed job | 1.665 | 1.000 | 3.000 | 9.000 |
| Distinct public capabilities per reviewed job | 1.014 | 1.000 | 1.000 | 3.000 |
| Distinct reviewer wallets per reviewed job | 1.002 | 1.000 | 1.000 | 2.000 |
| Review-bundle span (hours) | 1.272 | 0.000 | 0.017 | 186.250 |
| On-chain reviews / unique reviewed jobs | 812 / 427 | | | |
| ERC-linked providers / provider-capabilities | 55 / 342 | | | |

| *Panel B: Review incidence* | Review incidence all ERC jobs | Review incidence post-first review | Log entries conditional all reviewed jobs | Log entries conditional post-first review |
|---|---|---|---|---|
| New buyer-provider relationship | 0.003* | 0.005* | -0.173 | -0.173 |
| | (0.002) | (0.002) | (0.179) | (0.179) |
| Third-party evaluation | 0.006 | 0.007 | 0.000 | 0.000 |
| | (0.005) | (0.007) | (0.000) | (0.000) |
| Log job duration | -0.000 | -0.000 | 0.676 | 0.676 |
| | (0.000) | (0.000) | (0.537) | (0.537) |
| Observations | 253,129 | 198,909 | 319 | 317 |

| *Panel C: Composition of accumulated reputation* | New buyers | First-time ACP buyers | Established-buyer jobs |
|---|---|---|---|
| Transaction-level reputation depth | 0.033 | -0.005 | -0.190* |
| | (0.032) | (0.015) | (0.111) |
| Independent reviewer breadth / reviewed job | -0.001 | -0.019 | 0.127 |
| | (0.201) | (0.138) | (0.576) |
| Reviewer diversification (1 − HHI) | -0.072 | -0.032 | 0.330 |
| | (0.204) | (0.164) | (0.484) |
| Public capability scope / reviewed job | 4.030*** | 0.654 | 1.695 |
| | (0.624) | (0.441) | (1.380) |
| On-chain information richness / reviewed job | 0.036 | 0.046 | 0.455 |
| | (0.080) | (0.049) | (0.323) |

| *Panel D: Future-review-growth selection diagnostic* | New buyers | First-time ACP buyers |
|---|---|---|
| Next-seven-day own-capability review growth | 0.195*** | 0.093*** |
| | (0.063) | (0.034) |
| Next-seven-day other-capability review growth | -0.040** | -0.018* |
| | (0.017) | (0.010) |

| *Panel E: First public-capability review* | New buyers | First-time ACP buyers | Established-buyer jobs |
|---|---|---|---|
| Days −30 to −15 | -0.057 | 0.151 | -0.831 |
| | (0.743) | (0.252) | (0.638) |
| Days −14 to −8 | -0.410 | -0.092 | -0.598 |
| | (0.337) | (0.143) | (0.482) |
| Days +1 to +7 | 0.890*** | 0.365 | -0.169 |
| | (0.227) | (0.243) | (0.299) |
| Days +8 to +14 | 0.930*** | 0.497** | 0.141 |
| | (0.260) | (0.214) | (0.501) |
| Days +15 to +30 | 0.842*** | 0.209** | 0.636 |
| | (0.173) | (0.086) | (0.700) |
| Observations | 1,076 | 1,076 | 1,076 |

*Notes.* This table summarizes on-chain reputation data and considers the robustness of our analysis of its introduction. Panel A displays summary statistics on the underlying reviewed ACP jobs. Multiple on-chain entries generated by the same job are retained as information richness, and reviewer breadth is based on distinct reviewer wallets. Panel B restricts the review-incidence denominator to completed jobs of providers observed in the ERC-8004 crosswalk; columns (3)–(4) condition on receiving an on-chain review. Review-incidence regressions absorb provider and service-by-week fixed effects and include submitted-field count and IHS job value controls when available. Panel C includes provider-capability fixed effects, service-by-date fixed effects, provider-specific linear trends, and prior seven-day provider-capability activity. Each composition measure is entered separately together with log transaction-level reputation depth. The capability-scope estimate should be interpreted cautiously because multi-capability reviewed jobs are rare. Panel D adds reputation that will be produced during the next seven days; significant coefficients diagnose endogenous reputation production. Panel E reports the unrestricted first-public-capability-review event-time exercise, with days −7 to −1 omitted. It is a temporal diagnostic, not a causal design. Mature-service variants are not reported because the identifying sample falls to four provider clusters. Standard errors are clustered by provider in Panels C–E and two-way clustered by provider and client in Panel B. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.11: Agent-to-Agent Coordination and Reciprocal Exchange

| | February 2026 (1) | March 2026 (2) | April 2026 (3) |
|---|---|---|---|
| *Panel A: Prevalence and reciprocal core* | | | |
| Share of jobs touching a coordinator | 0.693 | 0.948 | 0.252 |
| Share of jobs between coordinators | 0.408 | 0.852 | 0.188 |
| Share of directed edges that are reciprocal | 0.162 | 0.768 | 0.441 |
| Reciprocal share of coordinator-to-coordinator jobs | 0.730 | 0.750 | 0.928 |
| Share of transaction value touching a coordinator | 0.373 | 0.689 | 0.134 |
| Share of jobs touching a coordinator, excluding top 10 | 0.422 | 0.906 | 0.070 |
| Share of jobs between coordinators, excluding top 10 | 0.203 | 0.811 | 0.030 |
| *Panel B: Transaction value* | | | |
| Share of coordinator-touching jobs below 1 | 0.765 | 0.999 | 0.960 |
| Share of coordinator-touching jobs below 5 | 0.997 | 0.999 | 1.000 |
| Median coordinator-touching job value () | 0.015 | 0.010 | 0.015 |
| *Panel C: Strict-coordinator procurement around downstream sales* | | | |
| Purchase in preceding 24 hours | 0.771 | 0.703 | 0.867 |
| Purchase in following 24 hours | 0.769 | 0.657 | 0.913 |
| Preceding minus following, 24 hours | 0.002 | 0.046 | -0.045 |
| Purchase in preceding 7 days | 0.953 | 0.943 | 0.999 |
| Purchase in following 7 days | 0.861 | 0.904 | 0.976 |
| Preceding minus following, 7 days | 0.093 | 0.039 | 0.023 |
| *Panel D: Persistence of the coordinator role* | | | |
| | February → March | March → April | |
| Broad coordinators active next month | 0.800 | 0.147 | |
| Broad coordinators remain coordinators | 0.355 | 0.056 | |
| Strict coordinators active next month | 0.854 | 0.156 | |
| Strict coordinators remain coordinators | 0.365 | 0.058 | |
| Starting broad coordinators | 110 | 251 | |
| Starting strict coordinators | 96 | 224 | |

*Note.* The table describes the temporary coordination structure that emerges in identified agent-to-agent service exchange. A broad coordinator buys and sells at least two completed services in the month and trades with at least two distinct upstream and downstream agent counterparties; the strict definition additionally requires at least three purchases and three sales and service diversity on at least one side. Panel A reports the prevalence of coordinator-linked exchange and reciprocity. The top-10 rows remove the ten coordinators with the largest monthly incident job counts. Panel B documents the very small dollar values of coordinator-linked jobs. Panel C compares whether a strict coordinator's downstream sale is surrounded by an upstream purchase before versus after the sale; similar preceding and following rates indicate that high transaction frequency, rather than directional input-output sequencing, explains much of the raw timing pattern. Panel D reports adjacent-month role persistence. The analysis is descriptive and does not interpret network position as exogenous.

Table B.12: Token-Market Valuation, Market Access, and Productive Capability

| *Panel A: What is associated with monthly abnormal token returns?* | (1) | (2) | (3) |
|---|---|---|---|
| Current productive quality | -0.035 | -0.058 | -0.047 |
| | (0.088) | (0.086) | (0.087) |
| New-buyer breadth | | 0.087* | 0.010 |
| | | (0.048) | (0.048) |
| Established-buyer breadth | | -0.071*** | -0.087*** |
| | | (0.026) | (0.029) |
| Total job activity | 0.061 | | 0.133** |
| | (0.047) | | (0.064) |
| Provider fixed effects | ✓ | ✓ | ✓ |
| Calendar-month fixed effects | ✓ | ✓ | ✓ |
| Observations | 425 | 425 | 425 |
| Provider clusters | 96 | 96 | 96 |

*Panel B: Does market valuation add information about future productive quality?*

| | 1 month later | 3 months later | 6 months later |
|---|---|---|---|
| | (1) | (2) | (3) |
| Current productive quality, exact-sample benchmark | 0.161 | 0.063** | -0.027 |
| | (0.135) | (0.031) | (0.024) |
| Current productive quality, with market valuation | 0.161 | 0.066** | -0.006 |
| | (0.133) | (0.032) | (0.023) |
| Standardized token market capitalization | -0.028 | 0.016 | 0.093 |
| | (0.026) | (0.039) | (0.060) |
| Incremental within-$R^2$ from market valuation | 0.0027 | 0.0018 | 0.0284 |
| Calendar-month fixed effects | ✓ | ✓ | ✓ |
| Observations | 464 | 180 | 70 |
| Provider clusters | 224 | 50 | 28 |

*Notes.* Panel A examines the relation between token-market repricing and productive quality versus market access. The outcome is the provider token's monthly abnormal return, constructed from clean daily returns relative to other tokens in the same tier and requiring at least five clean return days in the month. Productive quality is the provider-month risk-adjusted completion measure used in the persistence analysis. New-buyer breadth and established-buyer breadth are the within-month standardized logs of one plus the number of distinct new and established buyers, respectively; total job activity is defined analogously. All Panel A specifications require at least ten resolved provider jobs and include provider and calendar-month fixed effects. Panel B asks whether an independent market signal contains information about future productive quality beyond current measured quality. The exact-sample benchmark and augmented regression use identical provider-month observations at each horizon. Market capitalization is standardized within calendar month. Incremental within-$R^2$ is the difference between the augmented and exact-sample benchmark specifications. These estimates are predictive associations and are not interpreted causally. Standard errors are clustered by provider agent. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

Table B.13: Token Returns Around Provider Market Access and Governance

| | Pre-event $[-3,-1]$ (1) | Days 0–1 $[0,1]$ (2) | Days 0–3 $[0,3]$ (3) | Days 0–7 $[0,7]$ (4) |
|---|---|---|---|---|
| New-buyer share | 0.010 | 0.025** | 0.037** | 0.044** |
| | (0.013) | (0.012) | (0.015) | (0.021) |
| Completion share | 0.014 | 0.028** | 0.026 | 0.063** |
| | (0.015) | (0.014) | (0.019) | (0.030) |
| Expiry share | -0.018 | -0.005 | -0.010 | 0.002 |
| | (0.018) | (0.016) | (0.023) | (0.033) |
| Third-party-evaluator share | -0.010 | -0.029** | -0.056*** | -0.066** |
| | (0.021) | (0.013) | (0.020) | (0.032) |
| Log job activity | 0.011*** | 0.012*** | 0.006 | 0.006 |
| | (0.004) | (0.003) | (0.004) | (0.005) |
| Token fixed effects | ✓ | ✓ | ✓ | ✓ |
| Tier × date fixed effects | ✓ | ✓ | ✓ | ✓ |
| Observations | 4,568 | 4,896 | 4,657 | 4,379 |

*Notes.* The table displays the relation between provider-day market-access and governance characteristics and cumulative abnormal token returns. The sample is restricted to provider-days with at least five ACP jobs. Abnormal returns are measured relative to other tokens in the same tier on the same date. Column (1) uses the three trading days preceding the ACP activity date; Columns (2)–(4) cumulate returns beginning on the ACP activity date. Because token repricing is already visible before some observed new-buyer expansion, these estimates are interpreted as descriptive information aggregation rather than causal price responses. Standard errors are two-way clustered by token and calendar date. Statistical significance of 10%, 5%, and 1% is denoted by *, **, and ***, respectively.

# C Online Appendix: Theory Details

## C.1 Platform Routing and the Organization of Verification

The baseline framework treats external verification as a governance technology that can be supplied outside the buyer–provider relationship. This appendix adds a simple routing margin that distinguishes two separate questions: whether a transaction uses external verification, and, conditional on verification, which evaluator performs it. The distinction is useful because a market may decentralize the first decision while centralizing the second.

Consider a mass $M$ of transactions for which external verification is used. Each transaction can be routed to evaluator $e \in \mathcal{E}$. Let $n_e$ denote evaluator $e$'s verification volume, and let its per-transaction verification cost be

$$k_e(n_e) = \bar{k}_e - g(n_e), \qquad g'(\cdot) \geq 0, \tag{C.1}$$

where $g(n_e)$ captures standardization, tooling, or other scale economies in evaluating machine output. In addition, selecting and integrating a non-default evaluator requires a routing or compatibility cost $\tau > 0$. If the platform designates evaluator $d$ as the default, a transaction routed to $d$ avoids this cost. The effective verification cost is therefore

$$\widetilde{k}_e = k_e(n_e) + \tau \mathbf{1}\{e \neq d\}. \tag{C.2}$$

A buyer that has already decided to use external verification chooses the evaluator that minimizes expected verification cost net of any evaluator-specific quality advantage. Let $a_e$ denote the value of evaluator $e$'s accuracy or specialization. Conditional on using verification, evaluator $e$ is chosen to maximize

$$a_e - \widetilde{k}_e. \tag{C.3}$$

The platform's designation of a default therefore changes the allocation of verification even when evaluator technologies themselves do not change.

This extension generates two conceptually distinct sources of concentration. First, concentration may arise endogenously because $g'(\cdot) > 0$: greater volume lowers the cost of verification, generating the familiar scale-economy force in Prediction 3.2. Second, concentration may arise because the platform coordinates transactions on a common verifier by reducing the routing cost of the default. The latter mechanism does not require the default evaluator to be intrinsically superior. If

$$\tau > \max_{e \neq d} \left\{ (a_e - a_d) - [k_e(n_e) - k_d(n_d)] \right\}, \tag{C.4}$$

then all otherwise comparable transactions prefer the default evaluator. Verification can therefore become highly concentrated even when multiple evaluators remain technologically feasible.

The platform's role is consequently one of *meta-governance*. Buyers and providers may choose whether to rely on formal verification, but the platform can organize the infrastructure through which verification is supplied. This distinction matters because it separates bilateral governance from infrastructure governance. Relationship capital can reduce the extensive-margin demand for verification, as in Prediction **??**, while a platform default can

determine the intensive-margin identity of the verifier among transactions that remain externally evaluated.

The model also implies that evaluator concentration need not indicate durable evaluator-specific capability. Suppose two evaluators $d$ and $d'$ have similar $a_e$ and $k_e(\cdot)$, and the platform can change the default at low switching cost. A change in the default from $d$ to $d'$ then produces a discrete reallocation of verification activity even if the underlying verification technology is unchanged. In the limiting case, the incumbent evaluator can lose nearly all routed volume immediately when the default changes. Thus, an abrupt evaluator handoff is consistent with verification being a standardized, modular input whose market position is conferred by platform routing rather than by persistent evaluator-specific advantage.

This yields the following extension of Prediction 3.2.

[Verification concentrates through scale or routing] Conditional on external verification being used, evaluation activity concentrates when either evaluator-level scale economies or platform-level routing advantages are sufficiently strong. A platform default can generate concentrated verification without requiring a natural monopoly in evaluator technology. Accordingly, a change in the default may produce a sharp reallocation of evaluator market share without a commensurate change in the underlying use of verification or in transaction outcomes.

The empirical distinction is observable. Scale economies alone predict concentration, but do not by themselves predict that evaluator identity should change almost discretely following a platform routing change. Platform coordination predicts that evaluator shares can shift sharply when the default changes, while the relationship between bilateral history and the decision to use verification can remain distinct. The October 2025 evaluator transition therefore informs the mechanism behind concentration rather than merely documenting its level.