Financial Economics Insights

Article

"Job Envelope" Pricing Framework for AI Agents: A Comparative Perspective with Electricity, Telecom and Cloud Pricing Evolution

Jingyao (Lux) Zhao [1,*]

[1] Harvard University, Cambridge, USA
* Correspondence: Jingyao (Lux) Zhao, Harvard University, Cambridge, USA

Abstract: In recent years, artificial intelligence productization has expanded beyond Large Language Models (LLMs) toward agentic applications that embed autonomous reasoning, planning, and action into customer-facing workflows. Unlike LLMs, which can be priced as infrastructural utilities based on token consumption, AI agents operate as task-oriented systems that coordinate tools, memory, and retries over time to achieve domain-specific goals. As a result, existing usage- or outcome-based pricing models fail to fully capture the cost structure and value creation mechanisms of agentic systems. This paper examines the emerging landscape of AI agent pricing through a comparative lens, drawing parallels with the historical evolution of pricing in electricity, telecommunications, and cloud computing. Across these markets, pricing structures converged toward multi-part tariffs that aligned with underlying cost causation, capacity constraints, and quality of service considerations as technologies commoditized and diffused. Building on these insights, this paper proposes pricing per job envelope as a new paradigm for AI agents. This paper formalizes a three-part tariff consisting of a fixed envelope fee, allowance-based activity pricing, and optional quality-of-service modifiers. This framework aligns with established pricing models for knowledge work, such as consulting engagements, while leveraging automation and telemetry to enforce boundaries more precisely. The job envelope framework provides a scalable and economically robust foundation for pricing agentic systems as they move toward widespread enterprise adoption. Ultimately, this comparative analysis and the resulting multi-part tariff structure offer critical strategic guidance for developers and enterprises seeking to sustainably monetize and deploy next-generation autonomous artificial intelligence solutions.

Keywords: ai agents; pricing models; multi-part tariffs; autonomous systems; enterprise ai; cloud pricing

Received: 14 February 2026
Revised: 29 March 2026
Accepted: 10 April 2026
Published: 16 April 2026

1. Introduction

In recent years, AI productization has expanded beyond Large Language Models (LLMs) into agents, or agentic applications. These agentic applications, distinct from the models themselves, are no longer mere resources of productivity but often customer-facing products that integrate into existing workflows. Companies in customer service, sales, legal, and IT have developed a new generation of enterprise-facing applications that offer domain- or function-specific agents as an essential part of their products [1].

Unlike LLMs, which process tokens and are charged based on token input, output, or cache, agent stacks are more complex [1]. The most common type of agent architecture includes orchestration, tool-calling, retries, and memory: a ReAct-style control loop where an LLM iteratively performs stateful planning, selects tools or sub-agents, executes actions, incorporates observations into a shared context or memory, and repeats until a termination condition is met. While typically "agentic workflows" involve orchestration

under explicit human instructions, this paper focuses on agents in the strict sense, which includes an autonomous orchestration or "planning" layer.

Pricing for AI agents is a complex and emerging topic, inevitably diverse given the range of applications and services agents are beginning to be employed in the economy. If LLMs resemble electricity, then agents resemble the process where electricity becomes embedded in various use cases: lighting, factories, locomotion, etc. Therefore, pricing for agents naturally evolves with the units of work it carries out, rather than the units of utility [2]. This paper aims to compare the state of agent pricing with the historical developments of the electricity, telecom, and cloud markets, as well as propose a new paradigm of AI agent pricing for the future. Many of these mature industries only developed complex pricing structures after wider adoption and massive scale of demand, which is yet to come for agents but ultimately inevitable. This paper aims to provide a vision for the next chapter of agents pricing, based on a comparative perspective from more mature markets.

2. Comparative Perspectives: Pricing Evolution in Electricity, Telecom and Cloud

The pricing evolution of new technologies progresses through various stages of value prioritization, and only over time is the real measurement of value optimized. In this section, we examine the evolution of pricing for electricity, telecom, and cloud markets as they commoditize over time and adapt to the role of technology in society [3]. This includes the phenomena of non-coincidental peaks across the stack in electricity, the emergence of true value unit and Quality of Service pricing in telecom, and the bifurcation of storage and operational capacity in cloud services.

2.1. Electricity

Electricity emerged in the late nineteenth century as a practical industrial technology when inventors developed reliable generators, motors, and distribution systems [4]. In the late 1890s and early 1900s, watt-hour meters became commercially viable for measuring actual consumption. However, as electricity use expanded beyond lighting and demand increased on an unprecedented scale, peak capacity became an integral part of pricing. The 1892 Hopkinson Tariff was structured around two components: running cost and "peak load standing cost," the latter referring to the maximum possible demand from each customer, i.e., the cost of the electricity utility being ready to supply electricity, independent of customer usage. Measurements such as the Wright Demand Indicator were invented to help measure this maximum demand concept. This marked the first attempt at aligning utility pricing with underlying cost causations.

In the 1940s and 1950s, the cost causation of electricity gained renewed interest in the form of time-of-use rates, as it became evident that throughout the day and for each region, there are significant peaks and troughs in consumption aligned with the natural workday. The French Tarif Vert divided energy charges into winter and summer seasons, with different peak and trough schedules [5]. This time-of-use style of pricing represented further cost alignment improvement but soon showed its own shortcomings:

"the effectiveness of a time-of-use rate with only volumetric charges to provide economically efficient signals for capacity installation would depend on whether the peaks for generation, transmission, and distribution systems occur in the same time period... the peak demand for the distribution network occurs at a different time period than either the generation or transmission systems' peaks."

In other words, the more granular components of electricity cost are divided into generation, transmission, and distribution, which all have different cadences of peak and trough periods, resulting in an imperfect categorization of time-of-use pricing [1]. Besides the lack of supply-side cost component peak alignment, demand-side alignment is also challenging, since the system peak across a network might not align with each building or each household's local peak, resulting in noncoincident peaks. Despite its imperfections, time-of-use pricing is still widely adopted around the world today, with various modifications including critical-peak pricing during "critical event days," or default rates with customer flexibility to opt out.

The most advanced and modular experimentations include Illinois ComEd, which has a multi-part pricing structure: real-time energy pricing, coincident peak generation capacity (kwh), and a flat fee for transmission and distribution. Broadly speaking, this is the closest pricing method to capturing the real underlying cost of electricity operation: inclusive of generation, transmission, and distribution, as well as peak capacity [2]. Over 140 years of development in electricity pricing shows that the pricing of a new technology will align over time to its cost causation as it commoditizes, which is often modular and complex. While there are no perfect pricing methods, the most sustainable ones in the long term must take into account the unit economics of each cost component.

2.2. Telecom

Telephony, in its early phase in the late 19th and early 20th centuries, consisted of a simple business model under the "circuit-switched" network: for each voice call made, one dedicated path is maintained (usually routed by a professional) which stays active during silences and pauses. Very soon, Bell Telephone Company and the earliest telephone companies started charging based on duration and distance, the underlying cost causations. Early telephone call pricing was therefore predominantly minute-level billing, consisting of "two-part tariffs": fixed pricing and usage-based pricing, often in the form of prepaid services consisting of peak vs. off-peak minutes as well as distance differentiations [6].

In the 1960s, military-oriented research began to explore possibilities of a distributed network, which would presumably still survive a nuclear attack. The National Physical Laboratory in the UK simultaneously experimented with packet-switching, where data is broken down into packets, which are each routed independently through the network. A few years later, the U.S. Department of Defense built ARPANET, the first operational packet-switched network at scale. This technical innovation of packet-switching enabled more flexible routing than fixed circuit networks and allowed computers to send unprecedentedly high volumes of data, while voice carriers still functioned on a circuit-switched basis [7].

As consumer and enterprise demand for data grew faster than their demand for voice, more and more voice carriers and telecom incumbents decided to switch over to packet-switched networks and later the data-centric model: data plans (MB / GB) started to emerge, as smartphones also fundamentally changed the way users perceived value -- in data across various apps, and no longer in voice minutes [8].

Therefore, a combination of technological disruption in networks and a disruption-led new paradigm of user behavior led to a transformation in the perception of "unit of work," which then forced pricing schemes to change [9]. What might appear as the natural unit of work from the perspective of UI, which in this case was voice but for AI so far has been prompts and tokens, might not always end up being the true "unit of work." As agentic use cases develop, traces, completed tasks, and active workflows all start to emerge as plausible "units of work."

After the switch to data-centric pricing, major carriers faced fierce competition; the 3G/4G/5G evolution also led to the introduction of Quality of Service pricing, and the phenomenon of throttling, where data is unlimited theoretically, but users see a drop in speed or consistency after a certain threshold [10]. This is viable fundamentally due to the non-binary nature of data, whose speed is an enabling characteristic as opposed to a critically limiting one, such as uptime for electricity. As we will see in cloud pricing as well, quality of service pricing is crucial as a dimension to pricing, especially in a high-demand but increasingly commoditized category.

2.3. Cloud

Without delving too much into the history of computing, cloud as a technological innovation also evolved from the ARPANET and the packet-switching networks, as well as the connection of networks with each other later on, in the age of the internet [11]. At the turn of the 20th century, Amazon started to branch out from selling books to an "everything store," which meant that developers needed to code a vast amount of features

for every type of item and their personalization. This problem also exacerbated ironically as their customer base quickly scaled out; however, the previously monolithic architecture, where one single code base was now exposed to many risk surfaces, could not communicate with other services outside of Amazon. In response, in 2002, Jeff Bezos mandated the entire company to expose data only through externalizable APIs, even in internal communications, which led to an exploding amount of redundant building of storage, databases, compute, etc. Out of this dire internal need to share infrastructure, AWS was born in 2006 as a microservices, rentable "cloud infrastructure" where users could pay a monthly or annual fee to avoid maintaining their own infrastructure and data centers.

AWS services mostly included:

- Simple Storage Service (S3) for online storage, priced based on storage size
- Elastic Compute Cloud (E2) for on-demand processing, memory, and networking, priced pay-as-you-go based on usage
- Relational Database (RDS) for database service, priced based on instance cost, i/o operations, data transfers and any fees from the above

Compared to traditional on-premises cost structures, where each company needed to host their own data center and infrastructure, the new cloud model allowed businesses to rent out infrastructure largely on a pay-as-you-go model, significantly lowering the barrier to entry in software, although bringing up the cost of goods sold, since infrastructure was now scalable with revenue. With a huge total addressable market in sight, Microsoft, IBM, Oracle, and other tech players quickly piled on and rolled out their own cloud services, launching a "price war" that quickly reduced the marginal cost of cloud [12]. Through the 2010s, AWS cut its EC2 cost 107 times, S3 price was cut 80%, and overall pricing decreased 6-10 times (As shown in Table 1).

Table 1. Historical Price Reductions in Key Cloud Infrastructure Components.

Service	"Launch" price	Recent Price (2020s)	Rough drop
S3 Standard storage	$0.15 / GB-month	$0.023 / GB-month	~6.52×
EC2 compute	$0.10 / hour	$0.0104 / hour	~9.62×
RDS compute	$0.11 / hour	$0.017 / hour	~6.47×

Despite price cuts and the ongoing commoditization of cloud and the multi-faceted sources of cost, a closer look at the pricing structure of cloud providers reveals that there are inherently two main types of pricing:

1. Capacity: charged for resources that are provisioned or reserved over time, regardless of how much work they do, e.g., compute, storage, and databases
2. Activity: the cost of any ongoing activity based in the cloud, including read/write, data egress fees, and other data movement fees

Over time, competitive capacity pricing adopted discounts on unused capacity, spot capacity, prepayment of capacity bills, and volume discounts to lock in 1-3 years of revenue. Google GCP also introduced consistent use discounts, which essentially is a volume- and time-based discount without lock-in of upfront cash payment [13]. In essence, these capacity plans are very much similar to the data plans in telecom, where capacity-based discounts incentivize more sticky and lasting user behaviors.

On the activity side, pricing has persisted slightly better than capacity pricing, although in recent years commoditization might still be ongoing given the latest EU Data Act stipulated that in the event of a switch of provider, customers can only be charged for the actual cost incurred on the original provider for a switch ("at-cost"), from 2027 onwards. Google also recently went further beyond this act, waiving egress fees altogether in certain use cases [14].

Also similar to the telecom industry, quality of service is an important lever in pricing and achieved through service level agreements across multiple dimensions: availability/uptime guarantees, consistency (best if dedicated host), latency guarantees (through reserved capacity and premium routing), durability (highest being 11-12 nines).

Broken down into cost causations, below are the main cost drivers that are each being monetized:

1. Compute
2. Storage
3. Data movement (e.g., data egress)
4. Usage (e.g., API calls, database read/write)
5. Capacity reserved
6. Quality of Service

The takeaways for agentic pricing from cloud pricing are two-fold: 1) the dichotomy of capacity vs [15]. activity pricing: while activity pricing might be most visible in everyday operations, as agents maintain some level of the "system of record" or memory operations, capacity-oriented pricing should be introduced; 2) quality of service might be a hidden dimension that requires exploring, especially for pricing models with multi-layered cost causation.

While cloud is not the perfect example for agentic pricing since it is ultimately also one of the cost drivers itself of the agent stack, the commoditization of cloud and evolution of cloud pricing illustrates the essential dimensions of pricing under competition. At the same time, the gross margin pressure of cloud on cloud-based enterprises has re-emerged as an issue for CIOs and CTOs, especially as public companies optimize for margins: cloud allows companies to scale effectively and fast, but as they acquire scale, they become judged upon a different set of metrics.

3. Survey of Existing Pricing Structures

3.1. Existing methods

The contemporary pricing landscape for AI agents can be organized into three dominant pricing archetypes, distinguished by what they consider as the unit of value:

3.1.1. Token- or direct usage- based pricing

This category prices agents based on the direct token usage or API calls cost associated with underlying models or tools. Examples include Factory, Cursor, and Writer, which are priced according to token economics. Regardless of the packaging, which could vary by seats, monthly fees, or other structures, the limits in each plan revolve around the actual token usages that resemble utility pricing of infrastructure generation cost.

3.1.2. Workflow or activity-based pricing

This category prices agents based on abstracted, measurable execution units such as workflow steps, runs, or operations, abstracting away underlying model tokens. Examples include Cognition ("Agent Compute Units"), Rox ("actions"), and Clay (credits system). These systems resemble cloud-era workflow engines: users are charged for activity flowing through the system, not for cognitive output or business outcomes. This model dominates developer- and automation-oriented platforms because it offers predictability and aligns pricing with observable system load [16].

3.1.3. Outcome-based pricing

Outcome-based models price agents based on business-relevant results, such as resolved tickets, completed workflows, or delivered tasks, rather than internal activity. Examples include customer support agents who are priced per resolved conversation or ticket. Professional service AI agents, while not specifically charging on outcome, challenge the traditional seat-based model as buyers evaluate these service agents in terms of labor-replacement value. This outcome-based approach treats agents as labor or service substitutes, integrating retries, planning depth, and tool usage into a single externally visible unit of value.

3.2. Comparisons with Other Markets

Despite their novelty, agent pricing models exhibit strong thematic continuity with earlier infrastructure markets. Similar to electricity markets, where generation, transmission, and distribution peak at different times, agent systems experience non-

coincident stress across layers. Model inference, tool invocation, memory access, and logging peak independently. Current pricing bundles these layers, but throttling, concurrency limits, and tiering already function as implicit non-coincident cost controls.

Meanwhile, the transition from voice-centric to data-centric telecom pricing mirrors the shift from interaction-based to workflow-based agent pricing. Workflow operations, runs, and credits function analogously to packetized data: bursty, asynchronous, and statistically multiplexed [13]. Platforms like Cognition and Clay flow through a system rather than as reserved sessions, echoing packet-switched network economics.

Finally, agent pricing also closely parallels cloud pricing's separation of capacity (reserved concurrency, memory footprint, seats) from activity (executions, tool calls, operations). Hybrid models, such as seat-based access with usage caps, closely resemble cloud reserved instances combined with usage-based overages [16].

While agents inherit much from prior markets, they also introduce fundamentally new economic dimensions not captured by compute, storage, bandwidth, or QoS alone.

3.2.1. Autonomy and uncertainty risk

Agents can take irreversible actions in production systems, creating downside risk absent in traditional cloud services. Outcome-priced vendors implicitly price this risk by bundling guarantees, escalation paths, and failure handling into their fees. This resembles insurance or bonded services more than infrastructure metering.

Unlike deterministic workloads, agents may require multiple attempts, replanning cycles, or corrective actions [14]. Platforms absorb this variance internally when pricing per outcome, introducing a pricing dimension tied to uncertainty and search, not just execution.

3.2.2. Planning depth and cognitive complexity

Agents differ not only in their runtime but also in their planning depth, branching, and strategy revision. This "planning depth" is more akin to search complexity than mere computation or storage. Credit-based systems, such as Clay, partially reflect this by imposing higher charges for complex or multi-step actions [13]. While this concept is intrinsically related to "quality of service" in previous markets, it encompasses more dimensions than simply high or low quality. The depth and complexity required often depend on the nature and scale of the task, as well as human definitions of good versus bad.

3.2.3. Swarms: Coordination overhead in multi-agent systems

When multiple agents communicate, delegate, or negotiate, coordination itself becomes a cost driver. This overhead, absent in electricity, telecom, or basic cloud models, emerges in agent platforms that support multi-agent workflows, or "agent swarms," and is typically hidden inside per-run or per-outcome pricing.

Therefore, current agent pricing models combine inherited infrastructure logics—non-coincident cost recovery, flow-based usage, and capacity reservation—with novel dimensions of autonomy risk, epistemic uncertainty, planning complexity, and coordination overhead. This indicates that agent markets are evolving beyond traditional utility pricing into hybrid regimes that blend infrastructure metering with service and risk allocation [17].

4. Pricing per "Job Envelope": A "Unit of Work" Based Pricing Proposal for Agents

4.1. Conceptual Architecture

While the pricing for agents is rapidly evolving, the new dimensions necessitate a fresh pricing paradigm, akin to the transformations seen in electricity pricing, packet-switch data pricing in telecom, and cloud capacity- and activity-based pricing. The key distinction lies in the unlocking of intelligent, non-determinant exploration, as well as the overhead of collaboration, as agent swarms become the norm in enterprise adoption. In essence, there is a need for an updated pricing paradigm for intelligence.

As agent deployments transition from single, linear executions to swarms of cooperating agents, existing pricing models begin to falter [16]. Pure outcome-based pricing becomes fragile in swarm settings because outcomes are emergent rather than deterministic: multiple internal agents may explore dead ends, retry actions, or abandon partial plans before converging on a solution. In such systems, marginal cost variance increases sharply, and vendors absorb unbounded epistemic and coordination risk. Conversely, pricing purely on low-level usage metrics—such as tokens or individual steps—penalizes exploration and delegation, discouraging precisely the behaviors that make agent swarms valuable. This tension mirrors earlier infrastructure transitions where neither per-session nor per-unit pricing could adequately support systems characterized by burstiness, uncertainty, and shared resources.

The key structural insight is that agent swarms introduce a third layer of economic value beyond inputs and outputs: coordinated problem-solving over time. Unlike traditional workloads, swarm-based agents generate value not through a single model invocation or a single observable outcome, but through a bounded process of planning, delegation, execution, observation, and revision. This suggests that the appropriate pricing primitive is neither the token nor the business outcome, but an intermediate construct: a job envelope. A job envelope is a bounded, goal-directed execution context within which multiple agents are permitted to coordinate, explore alternatives, retry actions, and share memory, subject to predefined economic limits. It functions as the fundamental unit of work in swarm-based systems.

Building on this insight, I propose an experimental structure for agent swarms in the format of a three-part tariff:

1. A fixed base fee per job envelope covers orchestration overhead, baseline planning, and a limited amount of coordination and retry activity. This fee provides predictability for users while partially internalizing risk for vendors, analogous to fixed customer charges in electricity or request fees in cloud services.
2. Each job envelope includes a budgeted allowance for swarm activity—such as agent steps, tool calls, memory operations, or internal messages—with incremental charges applied to overages. This flow-based component prices coordination and exploration without exposing users to fine-grained token accounting, while simultaneously capping vendor downside.
3. (Optional) Outcome- or quality-of-service modifiers allow users to pay premiums for guarantees such as latency bounds, success rates, auditability, or human escalation. These modifiers sit atop the core tariff rather than replacing it, ensuring that outcome pricing enhances rather than destabilizes the underlying economic model.

This structure balances growth and margin protection in a way that neither pure usage-based nor pure outcome-based pricing can achieve. Predictable base fees reduce adoption friction and encourage experimentation; activity budgets prevent unbounded swarm behavior and margin erosion; and QoS premiums align pricing with high-stakes use cases that demand reliability and accountability. The logic closely parallels the evolution of cloud pricing toward reserved capacity plus burstable usage, and telecom pricing toward data plans with caps and priority tiers. Importantly, however, the job-envelope model reflects dimensions unique to agent systems, including epistemic uncertainty, planning depth, and coordination overhead—factors that have no direct analog in earlier utility markets [18].

In this sense, swarm-based agent pricing represents both a continuation and an extension of historical infrastructure economics. Like electricity, it must account for non-coincident bottlenecks across layers; like telecom, it must accommodate bursty, packetized flows; and like cloud, it must separate standing capacity from activity. Yet agents also introduce fundamentally new considerations—autonomy risk, cognitive depth, and coordination complexity—that necessitate a higher-level unit of work [19]. As agent swarms become the dominant deployment pattern, sustainable pricing will therefore converge on job-envelope-based models that explicitly price bounded coordination rather than raw computation or final outcomes alone.

4.2. Semi-formal Proposal

The total price charged for a job envelope j can be expressed as follows:

$$P_j = F_j + \Sigma_k[\ \alpha_k \cdot max(0, x_{\{j,k\}} - \bar{x}_k\ )\ ] + \Sigma_m\ [\ \beta_m \cdot q_{\{j,m\}}]$$

This expression breaks down pricing into three components: a fixed envelope fee, activity-based overages, and optional quality-of-service modifiers [17].

The fixed envelope fee F_j reflects the cost of admitting and maintaining a job in the system. It depends on the task class and scope:

$$F_j = f\ (g_j\ , s_j)$$

where g_j denotes the type of goal and s_j denotes the job's scope (for example, time horizon, systems accessed, or risk level). This fee recovers orchestration costs, baseline planning, and a bounded amount of coordination and retry activity. Economically, it functions as a capacity and risk-sharing charge, similar to fixed access fees in cloud or utility pricing.

Within each job envelope, the system tracks coarse-grained activity metrics such as agent steps, tool invocations, memory operations, or inter-agent messages. Let $x_{\{j,k\}}$ denote the realized usage of activity type k, and let $\bar{x}_k$ denote the included allowance for that activity [20].

Charges apply only when usage exceeds the allowance:

$$overage_k = \alpha_k \cdot max(0, x_{\{j,k\}} - \bar{x}_k\ )$$

This structure prices coordination and exploration rather than raw compute, allowing agents to explore and delegate within reasonable bounds while preventing unbounded cost exposure for the provider.

Some jobs require higher reliability or stronger guarantees. These requirements are captured through quality-of-service modifiers $q_{\{j,m\}}$ such as success guarantees, latency bounds, auditability, or human escalation. Each modifier is priced separately using coefficient β_m.

Importantly, these modifiers supplement the envelope and activity pricing rather than replacing it [21]. This avoids the instability of pure outcome pricing, where the provider bears all uncertainty associated with retries and exploration.

From the provider's perspective, expected pricing must exceed expected cost:

$$E[P_j] \geq E[C_j] + \varepsilon$$

where ε represents a target margin. The envelope fee covers fixed overhead, allowances absorb typical variability, and overages and QoS modifiers protect against tail risk. At the same time, marginal prices within the allowance region remain low, preserving incentives for exploration and effective coordination [22].

In summary, this framework generalizes prior infrastructure pricing models by introducing a bounded unit of work that internalizes uncertainty and coordination. It preserves the separation between capacity and activity seen in cloud pricing, incorporates flow-based usage similar to packet-switched networks, and explicitly prices risk and reliability in a way that traditional token- or outcome-based models cannot.

5. Comparison to Knowledge Work Pricing

5.1. Similarities

The proposed job envelope framework closely resembles how complex knowledge work, such as management consulting, legal services, and systems integration, has traditionally been priced. In these domains, work is rarely priced purely on outcomes or granular effort. Instead, pricing reflects a bounded engagement within which skilled professionals are granted discretion to explore alternatives, revise hypotheses, and coordinate across teams. This structure arises because the path to a successful outcome is uncertain beforehand, and efficient performance requires flexibility rather than strict adherence to predefined steps [23].

A consulting engagement is typically defined by a goal, a scope, and a set of constraints, such as duration, staffing levels, and budget. Within these bounds, consultants are expected to exercise judgment: pursuing lines of analysis that may or may

not pan out, iterating on recommendations, and reallocating effort as new information emerges. Clients do not pay per spreadsheet cell, per meeting, or per analytical iteration; they pay for the right to deploy expertise within a defined envelope. This is directly analogous to a job envelope in an agent system, which defines the goal and scope of work while allowing internal coordination and exploration without micromanaging each intermediate action [24].

The fixed envelope fee in agent pricing mirrors the role of project fees or retainers in professional services. These fees recover ramp-up costs, baseline availability, and the expected level of exploratory work required to address the problem. As in consulting, the fee is sensitive to task complexity and scope rather than to the precise amount of activity performed. It reflects an understanding that some degree of iteration and rework is not only inevitable but essential to producing high-quality outcomes.

Professional services contracts also implicitly include allowances for effort and iteration. When work expands materially beyond the original scope—because the problem is more complex than anticipated or because new requirements emerge—pricing typically shifts through change orders, expanded engagements, or additional workstreams. This parallels the activity allowance and overage structure in the job envelope framework. Normal levels of coordination and iteration are included by design, while exceptional expansion is priced separately [23]. In both cases, the objective is to avoid penalizing productive exploration while protecting the provider from unbounded cost exposure.

High-stakes consulting engagements often incorporate additional mechanisms for risk and quality management, such as senior oversight, responsiveness guarantees, escalation rights, or performance-linked fees. These features correspond to quality-of-service modifiers in agent pricing, which allow clients to pay explicitly for higher reliability, faster turnaround, auditability, or human intervention. Importantly, these elements are layered on top of the base engagement rather than replacing it, preserving economic stability while enabling differentiated service levels.

5.2. Key Differences & Innovation

The primary distinction between human knowledge work and agent swarms lies in observability and enforcement rather than economic structure. Agent systems generate detailed telemetry, tracking steps, retries, tool usage, and coordination, allowing for automatic and real-time enforcement of allowances and limits [3]. In contrast, human engagements rely on initial scoping and subsequent renegotiation. However, the underlying logic remains the same: both price bounded discretion under uncertainty rather than deterministic execution.

This comparison indicates that job envelope pricing is not an artificial construct imposed by technical necessity but rather a rediscovery of a well-established economic pattern adapted to autonomous systems. As agents increasingly perform tasks historically managed by knowledge workers, it is natural for their pricing to converge toward engagement-style models that balance flexibility, accountability, and risk-sharing, rather than purely computational or outcome-based schemes.

In the broader AI landscape, job envelope pricing represents a shift from pricing execution to pricing optionality [13]. The envelope grants the system a controlled right to explore, retry, and coordinate in pursuit of a goal, with economic limits transparent to both provider and user. This reframes pricing from a retrospective accounting of consumed resources to a forward-looking allocation of decision-making capacity under uncertainty. In this sense, job envelope pricing constitutes a novel hybrid between infrastructure tariffs and professional services contracts, uniquely suited to autonomous, adaptive systems whose value derives as much from their ability to search and adapt as from their raw computational throughput.

6. Concluding Remarks

This paper has argued that the emergence of AI agents as customer-facing, workflow-integrated products necessitates a rethinking of how artificial intelligence is priced. While Large Language Models can be reasonably priced as infrastructural utilities—metered by token input, output, or cache—agentic systems operate at a higher level of abstraction. They embed intelligence into concrete tasks, domains, and workflows, and their economic value derives not from raw computation but from coordinated, stateful problem solving over time. As a result, pricing agents purely on the basis of model-level usage obscures both the true cost structure of agent stacks and the value they deliver to users.

By examining the historical evolution of pricing in electricity, telecom, and cloud markets, this paper highlights a recurring pattern: as technologies commoditize and diffuse across use cases, pricing evolves away from surface-level units of consumption toward structures that reflect underlying cost causation, capacity constraints, and quality of service. Electricity pricing moved beyond volumetric energy charges to incorporate peak demand and non-coincident cost recovery; telecom pricing shifted from voice minutes to data-centric units and quality of service differentiation; cloud pricing bifurcated into capacity-based and activity-based components with layered service guarantees. These markets illustrate that mature pricing systems tend to be modular, multi-part, and closely aligned with how value is actually produced and constrained.

To address these challenges, the paper proposed pricing per job envelope as a new unit-of-work-based paradigm for agent systems. A job envelope represents a bounded, goal-directed execution context within which agents are permitted to plan, explore, coordinate, and retry subject to predefined economic limits. Building on this construct, the paper introduced a three-part pricing structure consisting of (i) a fixed envelope fee to recover orchestration and baseline risk, (ii) allowance-based activity pricing to account for coordination and exploration without micromanagement, and (iii) optional quality-of-service or outcome modifiers to price reliability, guarantees, and liability transfer. This structure balances growth and experimentation incentives with margin protection and cost transparency, while remaining legible to enterprise buyers.

Importantly, the job envelope framework also aligns agent pricing with long-standing practices in knowledge work pricing, such as consulting and legal services, where engagements are scoped ex ante and priced to allow bounded discretion under uncertainty. The innovation lies not in inventing an entirely new economic logic, but in formalizing and operationalizing this logic through automation, telemetry, and real-time enforcement. Job envelope pricing thus represents a shift from pricing execution to pricing optionality: the controlled right to explore and decide in pursuit of a goal.

As agent adoption remains in its early stages, pricing models will inevitably continue to evolve. However, history suggests that sustainable pricing regimes emerge only after periods of experimentation, scale, and competitive pressure. By situating agent pricing within a comparative framework drawn from more mature markets, and by articulating job envelope pricing as a principled response to the unique economics of agentic systems, this paper aims to provide a conceptual foundation for the next chapter of AI agent pricing—one that is robust to autonomy, coordination, and uncertainty, and capable of supporting agents as a durable layer of the modern digital economy.

References

1. J. C. Bonbright, A. L. Danielsen, and D. R. Kamerschen, Principles of public utility rates, New York: Columbia University Press, 1961, p. 33.
2. Q. Wang, C. Zhang, Y. Ding, G. Xydis, J. Wang, and J. Østergaard, "Review of real-time electricity markets for integrating distributed energy resources and demand response," Applied Energy, vol. 138, pp. 695-706, 2015.
3. L. G. Roberts, "The evolution of packet switching," Proceedings of the IEEE, vol. 66, no. 11, pp. 1307-1313, 2005.
4. P. L. Joskow, "Regulation of natural monopoly," Handbook of law and economics, vol. 2, pp. 1227-1348, 2007.
5. J. Calzada and F. Martínez-Santos, "Pricing strategies and competition in the mobile broadband market," Journal of Regulatory Economics, vol. 50, no. 1, pp. 70-98, 2016.
6. M. Medjaoui, E. Wilde, R. Mitra, and M. Amundsen, Continuous API management, "O'Reilly Media, Inc.", 2021.

7. P. Baran, "On distributed communications networks," IEEE Transactions on Communications Systems, vol. 12, no. 1, pp. 1-9, 1964.
8. L. Han, Market acceptance of cloud computing: An analysis of market structure, price models and service requirements, Bayreuther Arbeitspapiere zur Wirtschaftsinformatik, no. 42, 2009.
9. S. Yao et al., "React: Synergizing reasoning and acting in language models," in The eleventh international conference on learning representations, Oct. 2022.
10. V. Cerf and R. Kahn, "A protocol for packet network intercommunication," IEEE Transactions on Communications, vol. 22, no. 5, pp. 637-648, 1974.
11. M. Bhola and S. Bajeja, "Enhancing Cloud-Native Relational Database Systems: Proposed Design Patterns for AWS RDS Application," SN Computer Science, vol. 6, no. 5, p. 556, 2025.
12. A. Odlyzko, The history of communications and its implications for the Internet, 2000.
13. M. Armbrust et al., "A view of cloud computing," Communications of the ACM, vol. 53, no. 4, pp. 50-58, 2010.
14. M. Sabbir Rahman and M. Nusrate Aziz, "Service quality and behavioural intentions in broadband services selection," Marketing Intelligence & Planning, vol. 32, no. 4, pp. 455-474, 2014.
15. A. P. Sanghvi, "Economic costs of electricity supply interruptions: US and foreign experience," Energy Economics, vol. 4, no. 3, pp. 180-198, 1982.
16. K. Abdesselam et al., "The development of respondent-driven sampling (RDS) inference: a systematic review of the population mean and variance estimates," Drug and Alcohol Dependence, vol. 206, p. 107702, 2020.
17. D. H. Maister, Managing the professional service firm, Simon and Schuster, 2007.
18. J. Barr, AWS Storage Update–S3 & Glacier Price Reductions+ Additional Retrieval Options for Glacier, Amazon Web Services, 2016.
19. F. Casolari, C. Buttaboni, and L. Floridi, "The EU Data Act in context: a legal assessment," International Journal of Law and Information Technology, vol. 31, no. 4, pp. 399-412, 2023.
20. S. Russell, P. Norvig, and A. Intelligence, A modern approach, Artificial Intelligence, Prentice-Hall, Egnlewood Cliffs, vol. 25, no. 27, pp. 79-80, 1995.
21. K. K. Gelli, Improving Security and Transparency in Data Sharing with Web3 Integration and Blockchain Smart Contracts for Amazon S3 Access, Doctoral dissertation, Dublin, National College of Ireland, 2025.
22. M. Wooldridge, An introduction to multiagent systems, John Wiley & Sons, 2009.
23. V. Persico, P. Marchetta, A. Botta, and A. Pescapè, "Measuring network throughput in the cloud: The case of Amazon EC2," Computer Networks, vol. 93, pp. 408-422, 2015.
24. E. P. Lazear and M. Gibbs, Personnel economics in practice, John Wiley & Sons, 2014.