TECH mahindra

Scale at Speed™

READ TIME: 18 minutes

REPORT

Rethinking How Work Is Produced: Agentic Teams and the New Outcome Economics

AUTHOR

Tushar Srivastava

Head of AI and Quantum Computing,
UK & Europe, Tech Mahindra

FEATURING RESEARCH FROM FORRESTER

FORRESTER®

AI Agent Pricing: Innovation, Confusion, and Caution Ahead

IN THIS DOCUMENT

Why AI Pricing Fails Without Reinventing the Operating Model

The enterprise AI market is accelerating, but the foundations for pricing, measuring, and governing AI systems remain unstable. Most organizations attempt to price AI at the level of individual agents, mirroring the way software components have historically been bought and deployed. Yet as the industry experiments with subscription models, consumption models, digital worker constructs, hybrid approaches, and outcome-based contracts, one pattern is increasingly obvious: **pricing models fail when the delivery model remains unchanged.**

The critical assumption that needs to be challenged is not how AI agents are priced, but how enterprise work is produced. Enterprise outcomes—whether in IT services, IT operations, finance, HR, customer service, procurement, or sales—are not the output of individuals or isolated automations. They are the product of teams: orchestrated processes, separation of duties, coordinated roles, cross-checks, governance layers, and multi-system interactions.

Trying to layer AI agents into these environments without rethinking the operating model leads to the same problems Forrester observes across today's pricing landscape:

- outcome-based models become difficult to measure,
- consumption-based models become disconnected from value,
- digital worker models misrepresent the nature of enterprise work,
- hybrid models become confusing and inconsistent.

The missing layer is the delivery architecture—the structure that determines how AI, automation, tools, and humans work together to produce outcomes.

Tech Mahindra introduces three integrated constructs that reshape this foundation:

1

Vector Squads: multi-role, human-agent teams designed to mirror existing enterprise workflows while eliminating volume-driven variability.

2

Service Tokens: productized outcome units that make contractible, measurable outcomes possible.

3

Pricing Model Suitability Quadrant: a new analytical lens showing where outcome-based pricing becomes both high-value and operationally measurable.

These constructs represent more than a refinement of pricing. They represent **a new operating model** for AI-driven enterprises: a shift from deploying AI agents to **deploying hybrid human+AI agent teams that consistently produce outcomes**.

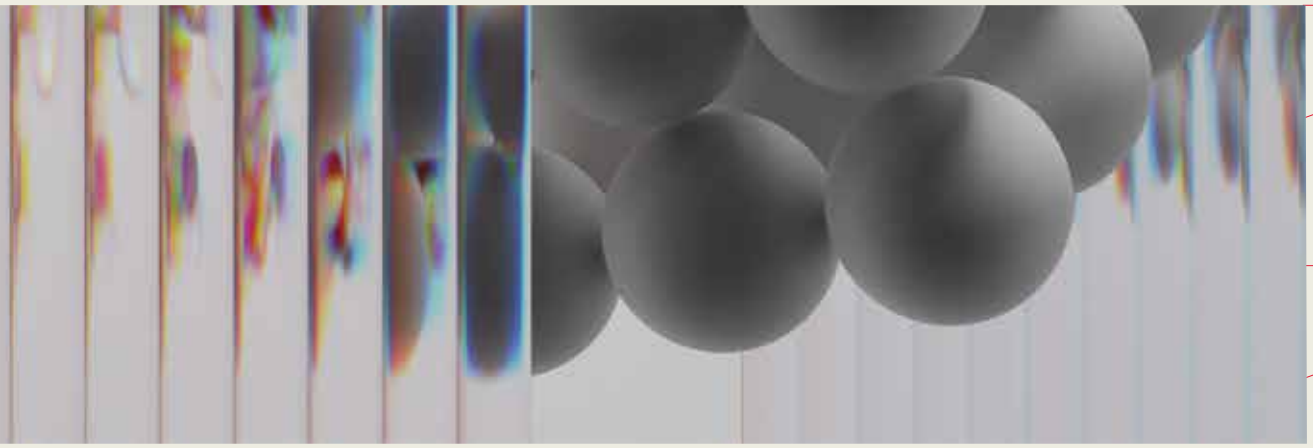
Outcomes are Made by Teams, not by Individuals or AI Agents

In enterprises, meaningful outcomes, across IT services, operations, customer support, finance, HR, procurement, and sales, are delivered through teams, not individuals. These teams follow well-understood operating patterns: different people play different roles, own different parts of a process, and outcomes emerge from coordination, checks, and approvals.

This is why outcomes cannot be delivered by a single AI agent or a chained sequence of agents. Enterprise work requires boundaries of responsibilities, approvals, checks, systems and skill sets. Simply chaining agents do not recreate these boundaries. It yields speed but removes structure: no separation of duties, no independent checks, no role-based accountability.

TechM's Vector Squads mirror existing enterprise teams, replacing large human groups with teams of humans and AI agents. For example: Regression Testing Squad, Data Migration Squad, Environment Provisioning Squad, Employee Onboarding Squad, etc.

Each squad delivers its familiar outcome, with humans handling judgment and exceptions, and agents handling high-volume cognitive work. This concept applies across the entire enterprise. This shift, from individual agents to structured human+agent teams, underpins why outcome-based pricing becomes feasible.



The New Economics of AI: Outcome-Based Pricing Finds its Sweet Spot

Pricing models across the industry face a core tension between outcome orientation and **measurement feasibility**. Some models are simple to count (API calls, consumption) but disconnected from value. Others are outcome-centric but historically impossible to measure.

TechM's Pricing Model Suitability Quadrant positions pricing constructs across X-axis: Outcome Orientation and Y-axis: Measurement Feasibility Quadrant Summary (see Figure 1).

- **Top-Left (Easy to Measure, Low Outcome):** Subscription, Consumption
- **Bottom-Left (Hard to Measure, Low Outcome):** Digital Worker/FTE
- **Bottom-Right (High Outcome, Hard to Measure):** Traditional Outcome-Based, Shared-Value, Hybrids
- **Top-Right (High Outcome, High Measurement Feasibility):** Vector Squads + Service Tokens

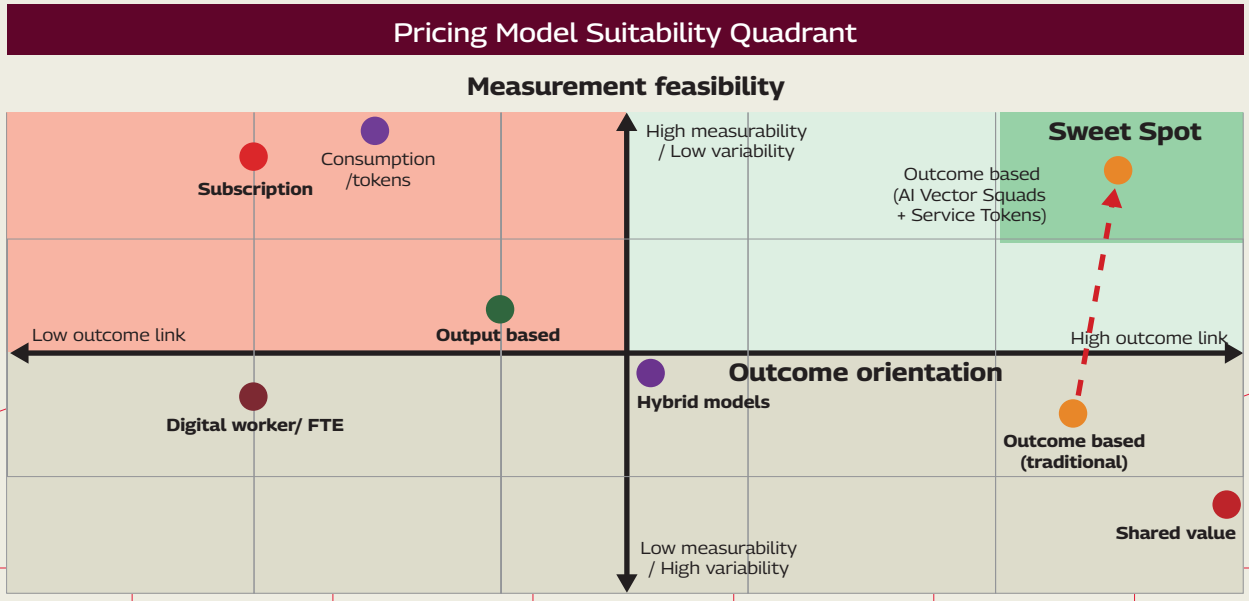


Figure 1: TechM's Pricing Model Suitability Quadrant

For years, the industry treated pricing models as points on a linear spectrum, with consumption models at one end and outcome-based models at the other. Outcome-based pricing was seen as the 'highest value' option, but too unstable to implement because human-led delivery introduced too much variability.

However, agentic AI teams like **Vector Squads can shift the outcome-based pricing model in the true sweet spot**, where pricing can be strongly linked to outcomes and reliably measured. That sweet spot did not exist in a human-only delivery model but becomes achievable with agentic teams.

Human+Agent Teams Create Stable and Scalable Outcomes

Vector Squads modernize existing service delivery teams into human+agent hybrids. Humans handle judgment, approvals, and ambiguity. Agents handle high-volume cognitive tasks, logs, data analysis, and multi-system operations. Vector Squads fundamentally reshape this economic landscape by creating a stable, repeatable delivery engine:

- Agents absorb workload spikes, preventing cost blowouts.
- Humans handle judgements, not volume.
- Separation of duties and role clarity make contributions auditable.
- Consistent patterns of execution make outcomes measurable across cycles.

This convergence of predictable delivery and high-value outcomes is what shifts outcome-based pricing into the Top-Right quadrant. Outcome-based pricing is no longer a theoretical ideal.

Service Tokens: Turning Outcomes into Products

Service Tokens convert Vector Squad outcomes into measurable units. Each Service Token represents a complete, end-to-end outcome, such as regression validation service token, readiness assessment service token, resolution service token, proposal assembly service token, reconciliation service token etc. Service Tokens include scope, criteria, governance, measurement rules, and boundaries.

With Vector Squads providing stability and governance, TechM's **Service Tokens** give outcome-based pricing a productized form. Traditional outcome pricing failed due to variability, e.g. number of test scripts to be run and verified can change significantly. Vector Squads break this link because:

- Agents do not scale linearly with volume
- Variability becomes a technical challenge, not staffing challenge
- Humans scale with complexity, not quantity

This produces **stable and predictable outcomes.**

In summary, Service Tokens is a catalogue-based approach where customers purchase well-defined outcomes instead of variable effort or hourly consumption.

- Service Tokens can be offered as a structured outcome catalogue, like how cloud providers offer SKU-based services. Each Service Token becomes a repeatable, productized unit with a fixed price, precise scope, standardized governance and predictable delivery patterns. This catalogue model eliminates the ambiguity and negotiation cycles typical of effort-based services.

- Historically, outcome-based pricing was disrupted by variations in workload - e.g., doubling test scripts, unpredictable spikes in incidents, fluctuating document volumes, or inconsistent data quality. These fluctuations drove up human effort and made outcome pricing risky for providers.

- With Vector Squads, these variations are absorbed by agents whose throughput does not scale linearly with volume. As a result:

    - volume spikes do not usually change Service Token pricing.
    - complexity, not quantity, determines human load; and
    - Service Token delivery remains stable even when underlying work expands

Enhanced Value Delivery, not just Cost Efficiencies

Service Tokens also **enable value-based outcomes** because Vector Squads dramatically compress cycle times:

- activities that once took one month (e.g., proposal creation, reconciliation cycles) can now be performed in one week
- processes that took one week (e.g., regression testing, readiness cycles) can be completed in one day
- near-real-time operations (incident resolution, monitoring, compliance checks) become possible without increasing human headcount

These reductions in cycle time introduce tangible business value: faster releases, quicker revenue cycles, reduced operational risk and improved customer experience. Service Tokens make this value measurable and contractible

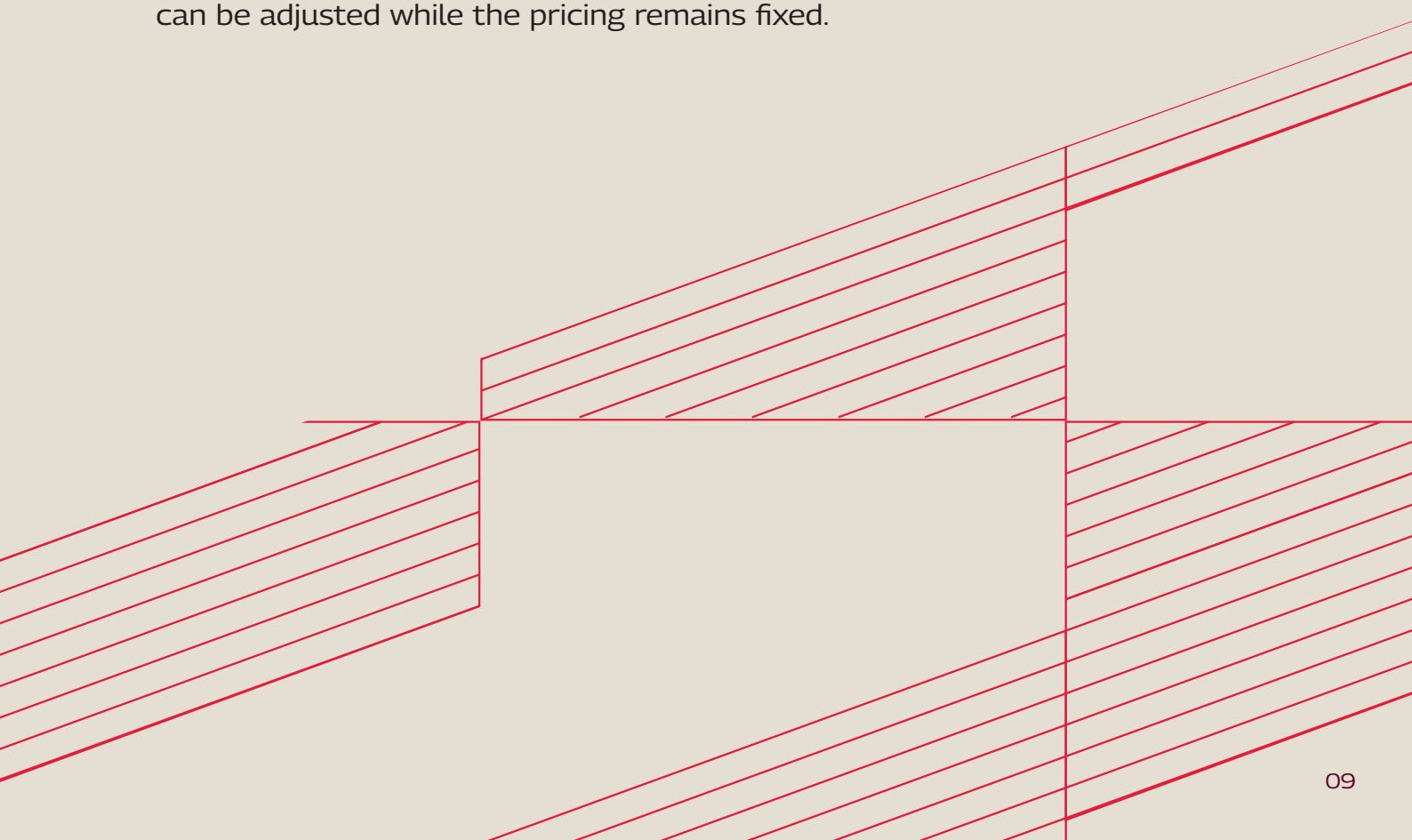
Governance in Human+AI Teams

Vector Squads deal with the important question around error-handling and overall governance. There are in-built checkpoints in Vector Squads where every agent action has defined validation points for human reviews or specific flagged exceptions. Human operators need to have well-defined roles as the approver/exception handler with clear accountability.

Service Tokens should be considered delivered only when:

- All steps as defined (within set parameters) are complete
- Human approval is provided at every checkpoint
- Final output meets all the quality checks as defined in the token specifications

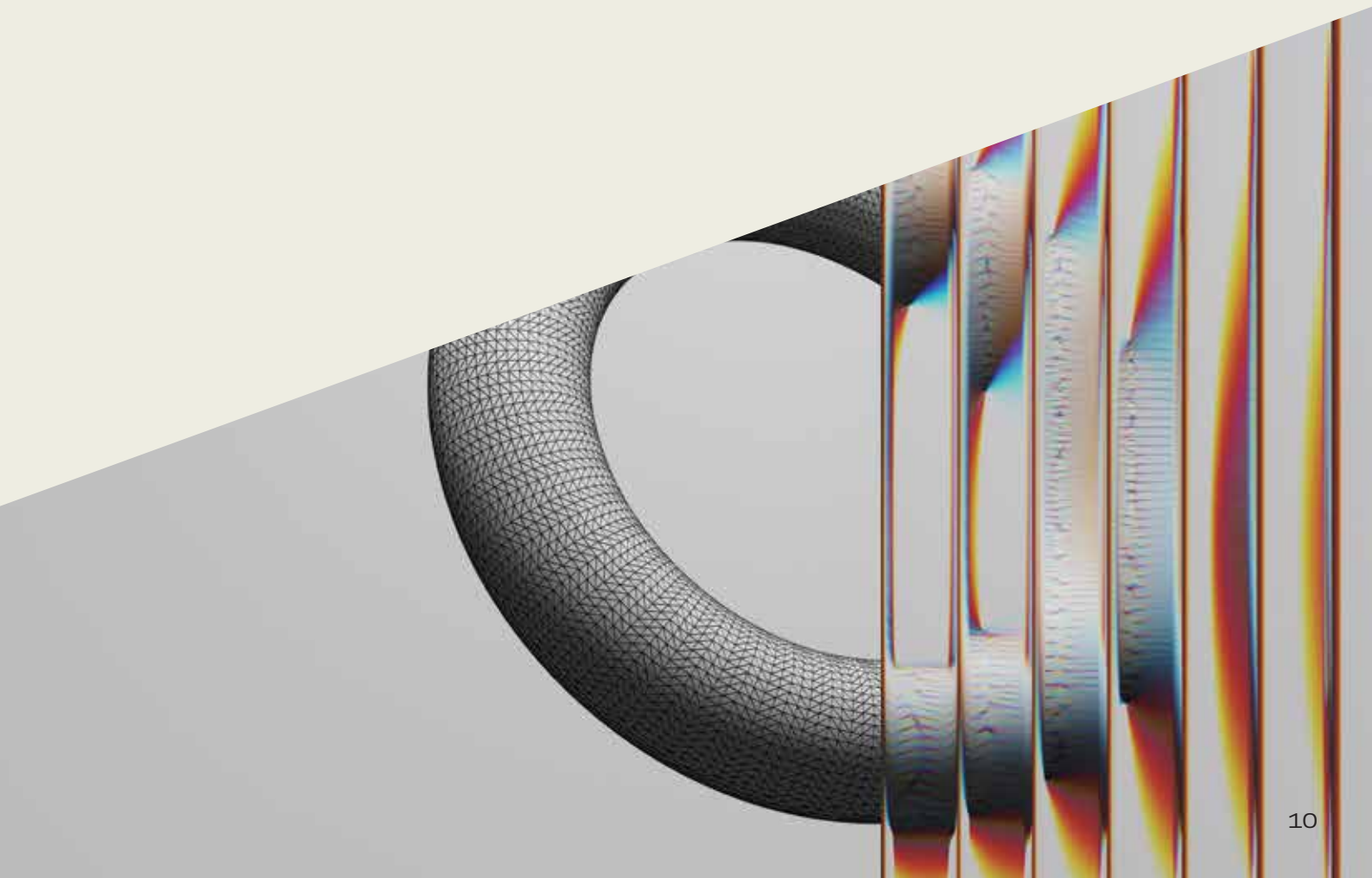
In case of human errors, the process must be paused at the following approval checkpoint. The exception handler will need to ascertain and correct. A root cause analysis must be conducted to determine if this is a recurrence or a one-time error. The timelines for Service Token delivery can be adjusted while the pricing remains fixed.



The Final Verdict

Outcome-based pricing has historically failed due to workload variability, attribution complexity, and governance overhead. Vector squads eliminate these constraints, enabling predictable delivery, measurable units, and stable pricing. TechM Service Tokens shift enterprise services from effort-based to outcome-based, creating predictable cost, stable margins, faster throughput, and a unified model across IT, Operations, and business functions.

Service Tokens and Vector Squads represent the kind of executional framework that can help organizations transition from theoretical value to measurable, meaningful outcomes.



About the Author

Tushar Srivastava

Head of AI and Quantum Computing, UK & Europe, Tech Mahindra

Tushar heads the AI and Quantum Computing practice of Tech Mahindra, UK & Europe. He is currently advising clients across sectors in their AI transformation journey with deep though leadership in the area. Tushar holds an MBA from the University of Oxford (Dean's List) and a B.Tech in Chemical Engineering from IIT Varanasi, along with qualifications in Quantum Computing from the University of Bristol. He actively contributed to shaping AI policies in the UK as a member of the All-Party Parliamentary Group on Artificial Intelligence.

CHART YOUR COURSE REPORT

AI Agent Pricing: Innovation, Confusion, And Caution Ahead

Critical Observations on AI Agent Pricing

September 12, 2025

By Craig Le Clair with Chris Gardner, Stephanie Liu, Lisa Singer, Renee Taylor-Huot, and Meg Bellavance

FORRESTER®

Summary

AI agents will change how businesses operate. This prompts a critical question for both vendors and buyers: How should they be priced and consumed? Unlike traditional software, an AI agent's value — and therefore its price — is dynamic, shifting with function, proximity to core business value, and increasing autonomy. This report delves into pricing models across different AI agent platforms and recommends how providers can maximize revenue and buyers can manage costs as agent capabilities evolve.

Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

AI Agent Platforms Currently Price Innovation In Contrasting Ways

The rise in AI is driving the need to better reflect value and cost. Understanding the pros and cons of pricing strategies has never been more important. Four platform types build most AI agents (see Figure 1). To effectively price and procure AI agents, buyers and vendors must understand the underlying platform's business model, strategy, and potential advantages and disadvantages. Recognize that:

- **Application-embedded agent platforms price for commitment.** Agent pricing for portfolio companies like Salesforce, ServiceNow, Oracle, and SAP is designed to drive commitment to their ecosystems, bundling agent-building capabilities into existing user license subscriptions to encourage full suite adoption. To better tie agent value to investment, they increasingly favor linking costs directly to business outcomes. Salesforce asks customers to buy "flex units" that convert to "actions," which then update systems of record (such as an appointment scheduled); it still employs a per-user license model for employee-facing agents. Platforms like Zendesk and Chargeflow charge only when AI agents successfully resolve inquiries or recover chargebacks, directly aligning costs with financial results. SaaS firms' outcome-based approach leverages domain expertise that adaptive process orchestration (APO) platforms, custom frameworks, and hyperscalers and data-centric platforms lack.

- **APO providers like digital worker pricing but haven't nailed it yet.** APO providers build on deterministic automation and hope their deep automation backgrounds will make them a viable agent platform. These platforms, particularly those originating in robotic process automation, are eyeing a "digital worker" or "digital employee" "agent hourly rate," or even an "agentic FTE" pricing model that captures fixed monthly or annual fees for deployed agents. However, most APOs wrap agent pricing into broader enterprise licenses and may add usage or output metrics. Digital worker pricing for agents has the potential to act as a proxy for

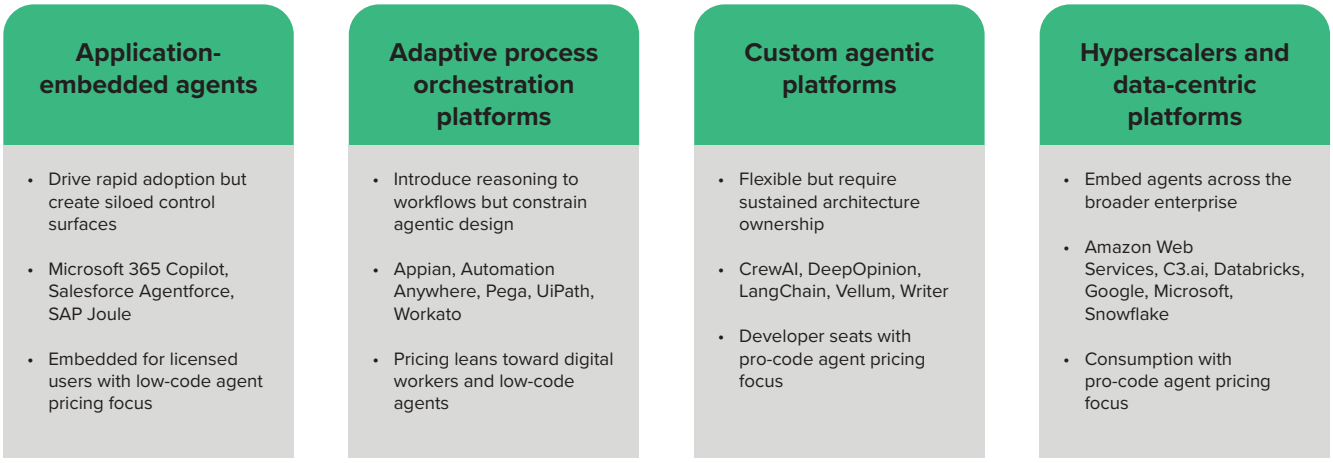
Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

employee support value but still lacks acceptance, as it's difficult to calculate and draws attention to the fact that AI is replacing human workers.

- **Custom agentic platforms are adopting seat-based pricing.**
  Custom agentic platforms sell developer seats to create and manage agents, augmenting platform licenses with consumption-based fees for foundational model costs. C3.ai primarily charges per licensed user but also incorporates consumption metrics like queries and data processed. DeepOpinion combines consumption (units) with output (documents or pages). LangChain charges a platform license but also tracks agent invocations and traces to offset model costs. Most custom agentic platforms offer pricing tiers.

- **Hyperscalers are adopting consumption pricing.**
  Like utilities, hyperscalers sell raw AI power and infrastructure at the most granular rate; developers spend time optimizing token and other consumption charges. They track raw compute resources through API calls, tokens, virtual machines, and specialized hardware. Consumption fails to capture the business value delivered and leads to unpredictable costs for customers. We asked an insurer to estimate the number of tokens for his claims process; he said that was impossible. Many customers monitor and optimize costs via FinOps and cost transparency approaches. Consumption-based pricing often works well when the agent's value is directly tied to usage or the action is system to system. Shortened procurement cycles, lower cost of experimentation, and less upfront commitment are positive benefits.

Figure 1: Platforms Currently Price Agents In A Variety Of Ways



© Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

Uncertainty Dominates The Outlook Of Both Buyers And Sellers

A core challenge lies in aligning pricing models with value as AI agents become more autonomous. And AI agents have yet to prove their value across all use cases; it's not always evident that an LLM-infused agent is cost-effective for document automation when a simpler machine learning model could suffice. The industry is grappling with how to quantify and monetize AI agents in a market still defining their true utility. Recognize that:

- **Agent progression supports outcome-based pricing.**
  The price of an AI agent can correlate to its degree of action and autonomy in fulfilling an output goal. As agents progress from simple "solvers" to more sophisticated "workers" to "executive" agentic systems, providers can tie prices to their level of accomplishment. The highest price tag will apply to AI agents that can collaborate with other agents, interact with numerous systems and automation endpoints, form decision loops for optimization, and take responsibility for broader goals — effectively replacing entire employee departments. Retrieval-augmented generation (RAG) models and standalone large language models (LLMs), which are far less independent, will command lower prices.

- **Hybrid pricing will be common and make platform comparisons difficult.**
  With hundreds of custom agent platforms and major platforms introducing agent-building capabilities, expect price to be a point of differentiation and innovation. For some use cases, hybrid just makes sense; an AI agent optimizing inventory workflow (efficiency) may also improve product sales (output). Pricing should combine subscription charges, consumption, and outcome-based approaches. The variety of hybrid pricing models add complexity in vendor and customer communication, sales compensation, and vendor billing; this is exacerbated by the agent deployment decisions companies face, such as whether to pursue traditional development with offshore resources or opt for a low-code strategy with citizen developers for lower costs, enhanced control, and agility. All of these affect agent platform selection and the resulting pricing model.

Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

- **New approaches to gathering feedback from customers will be needed.**
  Providers too often base pricing on the effort to build a feature or its perception as a differentiator, not what it means to the customer. Pricing decisions must start with understanding the value generated by the agent's performance. This should be the starting point for any pricing strategy. This intelligence helps connect pricing metrics to perceived value.

- **AI monetization will benefit from a shift to outcome-based pricing.**
  Enterprises struggle to prove AI's ROI, particularly for standalone LLMs and RAG agents with limited integration with business workflows; outcome-based pricing will help. As AI agents become more goal-oriented and autonomous and collaborate with other agents, their value will better align with broader business outcomes. Expect a shift from traditional subscription licensing to output, outcome, or digital worker engagements, where the returns on AI investment become clearer. Despite their granular, utility-based pricing, even hyperscalers will face pressure to connect costs directly to business value as customers seek clear impact.

- **Shared-value models will accelerate.**
  As the AI agent market matures and competition rises, we expect providers to offer shared-value models based on a percentage of increased revenue or cost savings generated. In an ideal scenario, outcome-based pricing would align agent performance to a client goal, backed by a revenue-sharing approach with the deploying services firm, such as a percentage of output or outcome generated. Outcome-oriented models work best when a benefit — tangible results like cost savings or increased revenue — can be shared.

- **Prices will rise and adjustments will be common.**
  AI agents will become more autonomous and burrow more deeply into core business processes. As a result, agreed-upon performance metrics and low initial prices to drive adoption will no longer seem fair to vendors (e.g., "You paid $500 an hour for that agent, but it's much better now.") Vendors will also gain leverage as user comfort and agent "stickiness" grow; they will adjust initial prices to reflect escalating utility and deeper integration into business operations; it's unlikely that application-embedded agents will be free with existing seat licenses. While prices may seem low now, they will rise as we move from a "try-it-out" phase to a mature, value-driven market.

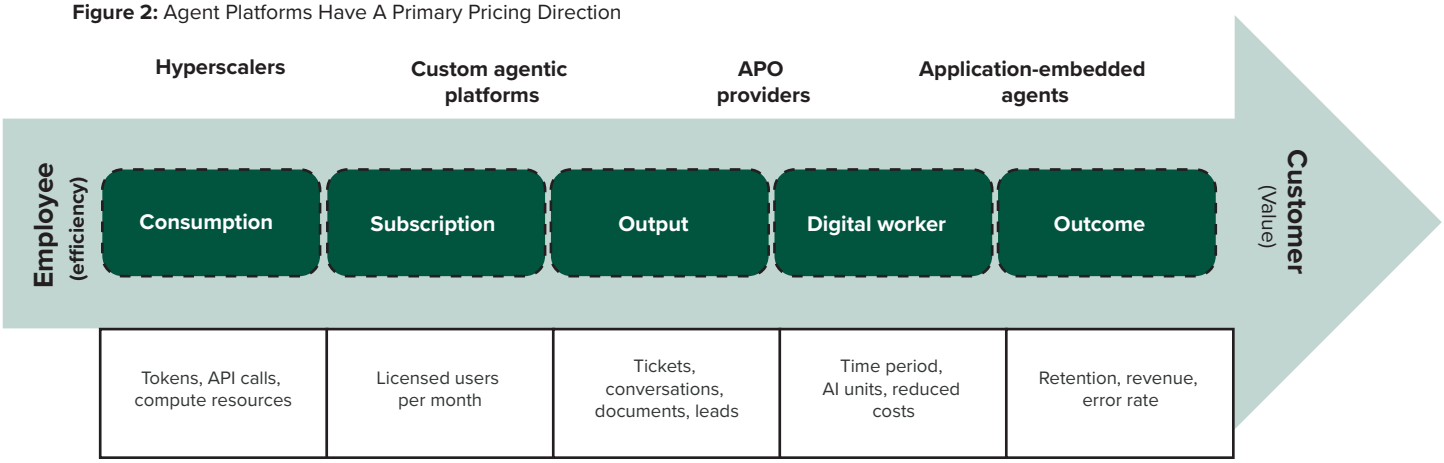
Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

- **Cost management for agentic systems will emerge.**
  While definitions of agentic systems differ, one certainty is their potential expense. Direct costs associated with AI development, models, infrastructure, data preparation, and model training will demand careful oversight. Operating costs for business transformation, governance, training, and ongoing operations will easily dwarf initial license fees. Deployment costs alone will see service-to-license ratios ranging from 5:1 to 10:1.

AI Agent Use Cases Must Influence Pricing

The closer an AI agent operates to a business's core customer base and direct revenue streams, the greater its potential for outcome- or output-based pricing (see Figure 2). Conversely, employee support agents designed to enhance efficiency are often better aligned with subscription, consumption, and digital worker pricing models. You need to price — and buy — agents by use case.

Figure 2: Agent Platforms Have A Primary Pricing Direction



© Forrester Research, Inc. Unauthorized reproduction, citation, or distribution prohibited.

When an AI agent primarily supports internal employees and operational functions, the value is less about direct value creation and more about higher productivity, fewer errors, and lower costs. In these scenarios, agent pricing is best as a subscription, consumption, or worker proxy where:

- **Value is an indirect efficiency gain.**
  Similar to traditional software, when an AI agent helps employees with operational tasks, it's best seen as a productivity tool. Buyers will estimate their ROI based on efficiency gains

Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

such as how much an agent save their compliance team (e.g., by reducing data entry and document review by 10 hours per week). While subscription pricing is a simple and predictable pricing approach, vendors and buyers may struggle to identify users who will realize tangible efficiency gains.

- **Agent work incorporates multiple systems, stakeholders, and processes.** It can be difficult to apply output- or outcome-based pricing to employee-facing AI agents; their contributions are often intertwined with multiple workers, systems, and workflows, making specific contributions hard to isolate. An AI agent designed to automate scheduling for heavy construction equipment will be difficult to charge per delivery. A per-user license for logistics staff using the agent would be better, even though the flat-rate pricing does not align with value. Heavy users pay the same as light users and a greater upfront commitment is required.

- **The agent can be priced like an employee.** Like all automation before it, the goal is to reduce human effort. Whether we admit it or not, AI agents will become coworkers, assuming some tasks currently performed by humans. Pricing as a percentage of a job done aligns better with value delivered than subscription and consumption pricing and may work when output- or outcome-based pricing cannot be clearly identified. Ask what percentage of a worker's job the agent will take. Solver agents today focus only on one component of a job, like Salesloft's deal summary agent, but worker and executive agents will take most of it. This approach aligns with the future direction of AI agents.

Price And Buy Revenue-Generating Agents

Many agents will directly interact with or support customers or align with tangible operational value. They will influence purchasing decisions or directly contribute to revenue generation. Their value is often tangible, measurable, and directly tied to business growth. Outcome pricing reduces the risk of paying for little value, simplifies the internal business

Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

case, and creates a shared goal with the provider. Use output or outcome pricing when:

- **An agent's value can be aligned with tangible results.**
  For outcome-based AI agent pricing, customers pay directly for measurable results the agent achieves; the cost is intrinsically tied to the benefit. Instead of a flat monthly fee, a customer service AI agent might charge per resolved inquiry. This model directly aligns the vendor's revenue with the AI agent's effectiveness, incentivizing strong performance and greater value generation for the customer.

- **Volumes are more predictable.**
  This model is often preferred by enterprise buyers deploying agents but can make budgeting costs more difficult than recurring monthly fixed fees. If the history of events or transactions that an agent will produce or handle are consistent, then outcome pricing is best. Volume spikes can be mitigated by tiered pricing that aggregates output volumes into pricing bands.

- **Outputs or outcomes are easily defined, measured, and attributable to the agent.**
  Defining and measuring the precise output of an AI agent can be difficult, hindering its use as a primary price anchor. Vendors are burdened with tracking often complex customer metrics and risk turning pricing into a science project with complicated negotiations. Buyers accustomed to straightforward license tracking are reluctant to adopt new, outcome-tracking procedures that incur administrative costs and have the potential for disputes and compliance risks.

Supplemental Material

Companies We Interviewed For This Report

We would like to thank the individuals from the following companies who generously gave their time during the research for this report.

- **Appian**
- **Automation Anywhere**
- **DeepOpinion**
- **Pega**
- **Salesforce**
- **UiPath**

Not Licensed For Distribution.
© 2025 Forrester Research, Inc. All trademarks are property of their respective owners.
For more information, see the Citation Policy, contact citations@forrester.com, or call +1 866-367-7378.

About Tech Mahindra

Tech Mahindra (NSE: TECHM) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 152,000+ professionals across 90+ countries helping 1100+ clients, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, customer experience & design, AI & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies. For more information on how TechM can partner with you to meet your Scale at Speed™ imperatives, please visit https://www.techmahindra.com/

TECH mahindra

www.techmahindra.com
www.linkedin.com/company/tech-mahindra
www.twitter.com/tech_mahindra

Copyright © Tech Mahindra Ltd 2026. All Rights Reserved.

Disclaimer: Brand names, logos, taglines, service marks, tradenames and trademarks used herein remain the property of their respective owners. Any unauthorized use or distribution of this content is strictly prohibited. The information in this document is provided on "as is" basis and Tech Mahindra Ltd. makes no representations or warranties, express or implied, as to the accuracy, completeness or reliability of the information provided in this document. This document is for general informational purposes only and is not intended to be a substitute for detailed research or professional advice and does not constitute an offer, solicitation, or recommendation to buy or sell any product, service or solution. Tech Mahindra Ltd. shall not be responsible for any loss whatsoever sustained by any person or entity by reason of access to, use of or reliance on, this material. Information in this document is subject to change without notice.