Teneo®

Agentic Pricing: Back to the Future?

What Telecom Pricing History Suggests About Monetizing AI

Teneo Insights | April 2026

Teneo

Summary

- **AI pricing is replaying part of the early telecom story**: markets start with visible usage meters but mature toward simpler recurring commitments with pooled usage and clear guardrails.
- **Raw tokens are unlikely to be the long-term buying unit**: they help vendors manage cost but are often a poor proxy for customer value delivered and a weak basis for predictable buying.
- **The key break from telecom is that AI can price completed work**, not just access: internal assistant use may become effectively bundled while external or mission-critical agents can be priced on conversations, resolutions, workflows or other auditable events.
- **The winning model is likely hybrid**: a recurring subscription or platform fee, a built-in usage allowance and straightforward expansion paths when customers scale.
- **Over time, value will shift away from model cost and toward owning the workflow:** the strongest monetization will come from accuracy, autonomy, governance, integration depth, and proven business outcomes.

1. Predicting the Future Using the Past

The current debate over AI agent pricing often looks like a false choice: seat pricing vs credits vs tokens, fixed ARR versus pure consumption, outcomes versus cost pass-through. Teneo's recent analysis suggests that many of these same challenges have been seen before by telecom operators as they launched and scaled the first generation of cellular services across the U.S.

The telecom of the 1990s looks very similar to the AI-first companies of today; large, fixed expenses, rapidly growing usage and little visibility into what the future steady state is going to look like.

Typically, new markets start by charging for the unit the system exposes most cleanly; mature markets migrate toward the unit customers can forecast, approve and renew.

- Wireless carriers began with minute buckets, peak and off-peak windows and overage economics because voice capacity was scarce and usage patterns were uncertain.[1]

- Over time, as adoption broadened and traffic exploded, the winning commercial design moved toward pooled plans, bigger bundles, and 'unlimited' packaging with hidden controls.[2]

AI agents are at the same messy stage today, but only up to a point.

Vendors do not yet know the true steady-state usage curve or the durable value created by each task. Customers, meanwhile, struggle to budget for variable token-driven bills while still being asked to believe in large labor-replacement stories. The deeper issue is that tokens are not just hard to budget for; they are often a poor proxy for value delivered.

---

[1] AT&T. Various years. "Nights and Weekend Minutes" and "Rollover Minutes." Accessed April 3, 2026. https://www.att.com/support/; Verizon. Various years. "Share Everything Plan" and "Unlimited Plans." Accessed April 3, 2026. https://www.verizon.com/plans/

[2] Verizon. Various years. "Share Everything Plan" and "Unlimited Plans."; T-Mobile. 2016. "T-Mobile Says 'R.I.P. Data Plans.'" Accessed April 3, 2026. https://www.t-mobile.com/news/

In telecom, a minute or megabyte was at least loosely correlated with customer value. In AI, the same volume of tokens can sit behind something trivial or something mission-critical. That weakens pure consumption pricing as a long-run anchor, not just as a customer-facing construct but as a value metric.

Our viewpoint is therefore simple: tokens are likely to be an important internal accounting unit but a weak mature buying unit.

The end state is more likely to be hybrid, similar to telecoms of the 2000s: a recurring platform commitment, an included pool of agent capacity and, where the use case allows, a business-facing meter tied to conversations, resolutions, workflows or other auditable events.

Exhibit A: The Three Objectives Every AI Pricing Model Must Balance

Buyer predictability Enterprise buyers will accept variable pricing only up to the point where budgeting breaks.	**Margin protection** Vendors still need a way to absorb volatile compute cost and protect gross margin.	**Adoption** Low-friction entry and generous inclusion matter when the usage curve is still being learned.

2. History of Telecom Pricing and the Drivers Behind It

Telecom pricing did not move in a straight line from metered to unlimited but instead moved through a sequence of commercial compromises designed to reduce customer anxiety without giving up network control.

In the 1990s and early 2000s, carriers sold fixed plans with limited anytime minutes and large nights-and-weekends allowances.[3] That structure matched the engineering reality: peak capacity mattered (to manage network bandwidth), minutes were easy to meter and voice was the core product.

As adoption widened, strict metering started to create bill shock and buyer friction. Carriers responded with rollover minutes[4] and larger bundles, effectively letting customers smooth uneven usage across months.

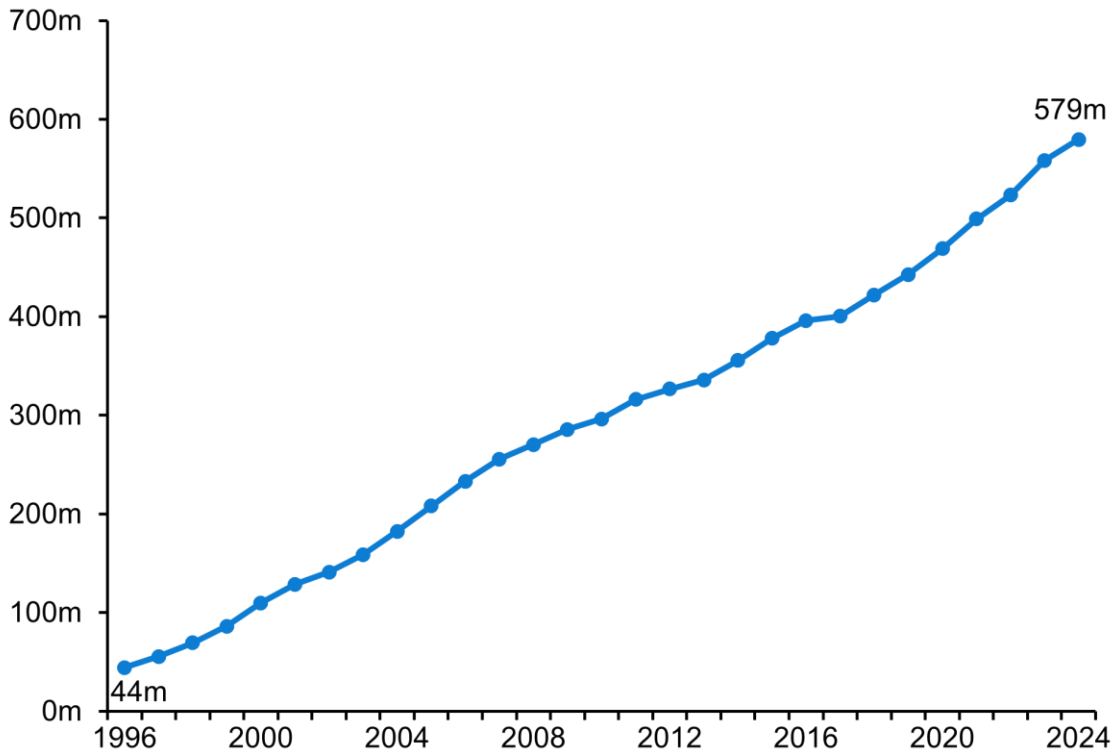
The next step was pooling. As households accumulated multiple devices and voice and text became less differentiated, plans such as Mobile Share and Share Everything[5] pooled one data allowance across several lines while making talk and text unlimited. How the phone was used mattered less than whether you were using it.

[3] AT&T. Various years. “Nights and Weekend Minutes.” Accessed April 3, 2026. https://www.att.com/support/
[4] AT&T. Various years. “Rollover Minutes.” Accessed April 3, 2026. https://www.att.com/support/
[5] Verizon. Various years. “Share Everything Plan.” Accessed April 3, 2026. https://www.verizon.com/plans/

Eventually, even data itself was simplified for the buyer. T-Mobile's 2016 'R.I.P. data plans' announcement[6] captured the market direction: sell predictability first, then use throttling thresholds, hotspot caps, priority levels and other terms to manage economics in the background.

Figure 1: Wireless Adoption and Usage Exploded Even as the Customer Buying Unit Became Easier to Understand

U.S. Wireless Connections, 1996-2024
Number of Connections

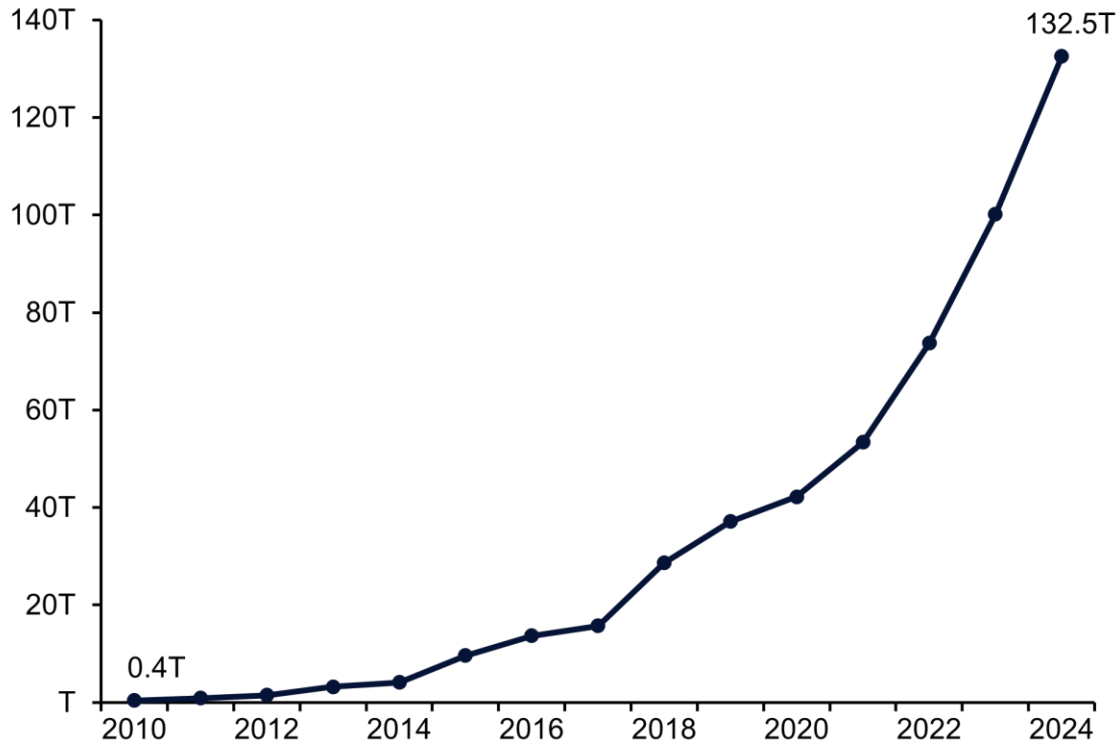


Source: CTIA 2025 Annual Survey Highlights; CTIA 2016 Year-End Top-Line Survey Results.

[6] T-Mobile. 2016. "T-Mobile Says 'R.I.P. Data Plans.'" Accessed April 3, 2026. https://www.t-mobile.com/news/

U.S. Wireless Data Usage, 2010-2024

MB

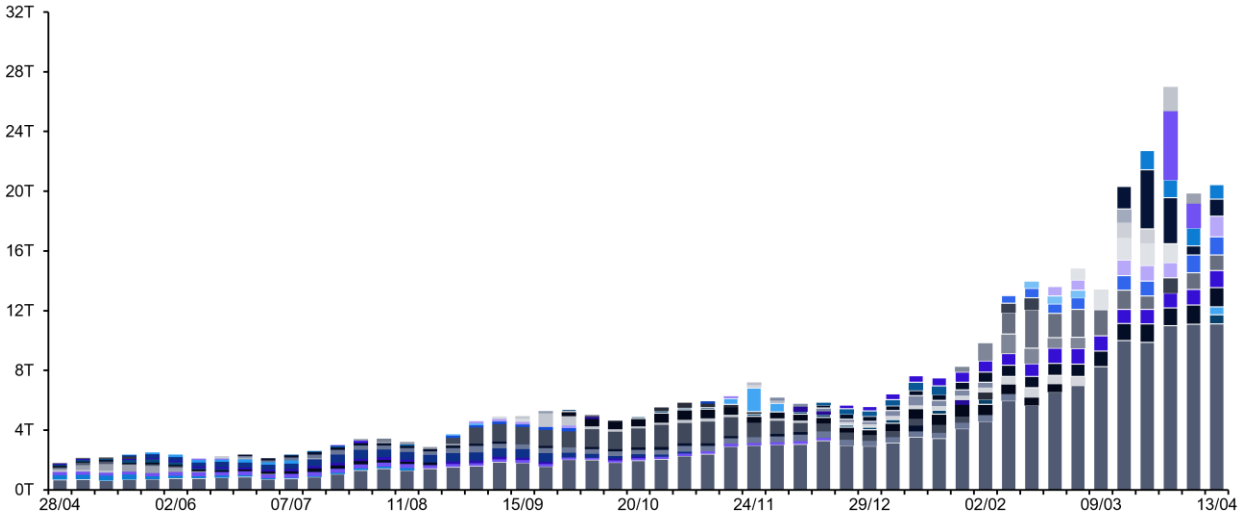


Source: CTIA 2025 Annual Survey Highlights; CTIA 2016 Year-End Top-Line Survey Results.

Figure 2: AI Token Usage Expansion Exhibits Similar Exponential Characteristics as the U.S. Wireless Data Usage Historical Dataset

Weekly Usage of Models Across OpenRouter, 2025-2026

MB



Model	Usage
MiniMax M2.7	961MB
MiMo -V2- Pro	1.15T
Claude Sonnet 4.6	1.38T
Claude Opus 4.6	1.22T
MiniMax M2.5	1.05T
Gemini 3 Flash Preview	1.14T
DeepSeek V3.2	1.28T
Grok 4.1 Fast	537MB
Gemini 2.5 Flash Lite	595MB
Others	11.2T
Total:	20.6T

Telecom regulation has now made the same point explicitly. The UK's Ofcom requires providers to disclose mid-contract price rises upfront in pounds and pence, reflecting a simple lesson: uncertainty itself can become a barrier to trust and scale.[7]

[7] Ofcom. Various years. "Price Transparency Rules and Guidance." Accessed April 3, 2026. https://www.ofcom.org.uk/phones-and-broadband/bills-and-charges

Teneo

Exhibit B: Telecom's Pricing Evolution Was a Series of Simplifications, Not a Single Jump to 'Unlimited'

Era	What telecom sold	Why it won	AI analogue
Scarcity and minute buckets	Limited anytime minutes, plus nights and weekends or other peak/off-peak constructs.	Carriers priced the cleanest engineering unit they could meter.	Tokens and raw actions are the AI equivalent of the early meter.
Bill smoothing	Rollover minutes and larger bundles softened month-to-month volatility.	Customers wanted fewer surprises without losing flexibility.	Included credits and starter allowances reduce fear of trying agents.
Shared pools	Family plans and shared data buckets spread uneven usage across devices and people.	Pooling matched real household behavior better than rigid per-line caps.	Workspace-level pools smooth uneven usage across teams and workflows.
Unlimited with guardrails	Unlimited plans hid the meter behind speed tiers, hotspot caps, and deprioritization.	Buyers wanted predictability, but carriers still needed controls.	Unmetered internal agents can work if model, channel, or concurrency guardrails stay in place.

3. Where We Are With AI Agent Pricing

AI agent pricing today resembles telecom before the category settled.

Software vendors are considering a wide range of pricing and monetization levers[8] with little consistency in approach across similar software sectors. Tokens and compute protect margins when model cost is variable. Credits create a vendor-controlled currency that can normalize many different agent behaviors. Conversations, resolutions, workflows and outcomes are easier for buyers to budget because they map more directly to work completed. Seats still matter for internal productivity and deployment breadth. The important distinction is that these units do not do the same job: tokens explain cost while outcomes explain value.

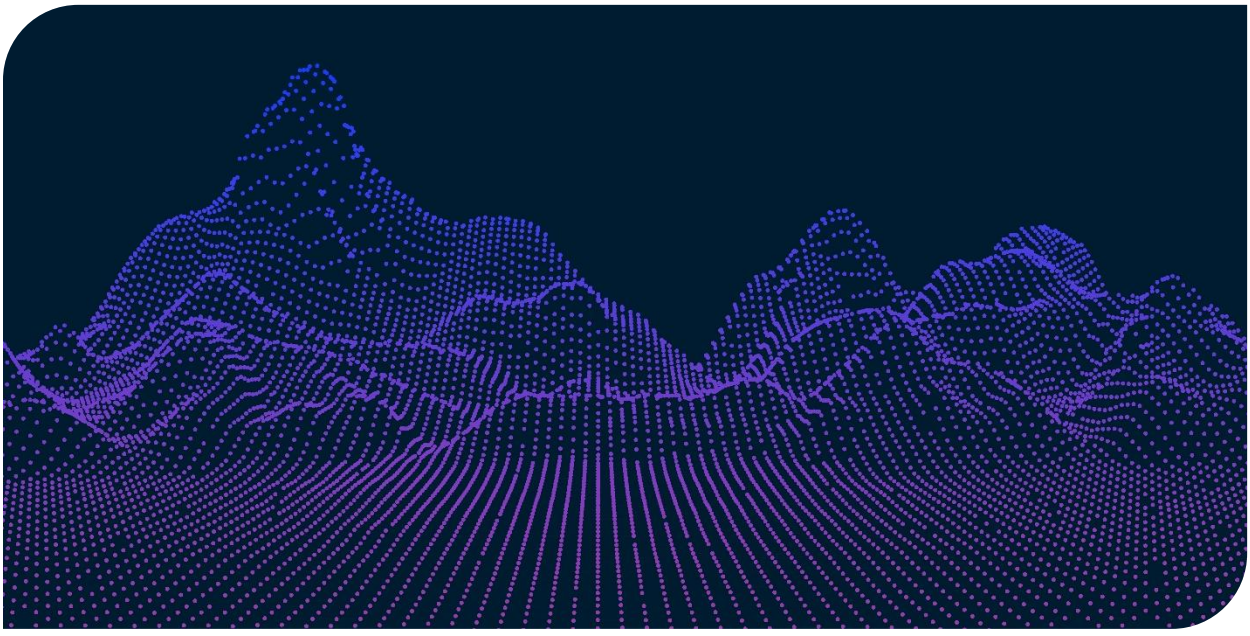
[8] Microsoft. 2025–2026. "Copilot Pricing and Copilot Studio Documentation."; Zendesk. 2025–2026. "Automated Resolutions and Pricing." Accessed April 3, 2026.

Exhibit C: A Plain-English Guide to the Units Buyers Encounter in AI Pricing Token

Token A low-level model input/output unit. Useful for internal cost accounting but often a poor proxy for customer value.	**Credit** A vendor-defined currency that bundles many different AI actions into one commercial pool.	**Outcome** A completed business result such as a conversation, resolution or workflow event. Strongest when value attribution is clear and auditable.

This means that for many of our clients, the answer is a hybrid model taking the best and the worst from each approach. But over time, where value attribution is clean, the center of gravity is likely to move toward workflow- or outcome-based pricing rather than remain anchored to pooled consumption alone.

The demand side is moving quickly enough to make this experimentation unavoidable. OpenAI says more than 1 million business customers now use its tools, more than 7 million workplace seats are live, enterprise message volume is up 8x year over year and reasoning-token consumption per organization is up about 320x.[9] At the same time, frontier model pricing has been deflating rapidly.

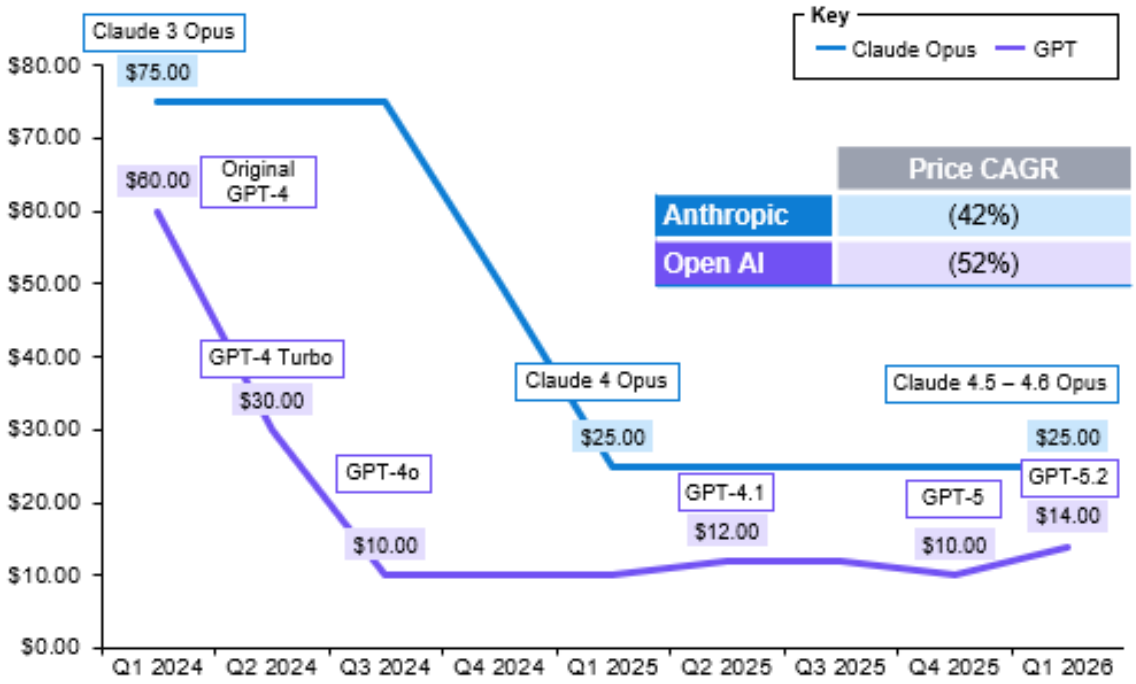


[9] OpenAI. 2025. The State of Enterprise AI. Accessed April 3, 2026. https://openai.com/

Figure 3: Published Frontier-Model Prices Are Already Falling Faster Than Most Enterprise Pricing Can Adjust

Anthropic and OpenAI Flagship Model Costs

AI Model List Price Per 1M Tokens, 2024 - Present[10]



OpenAI's published API pricing moved from GPT-4 Turbo at $10/$30 per million input/output tokens in late 2023 to GPT-4o at roughly half that price in 2024 and then to GPT-4.1 at $2/$8 in 2025. When the infrastructure cost curve falls that quickly, pure token pass-through is unlikely to support sustained growth within developers of agents and models as customers will invariably ask to share in the savings.[11]

Source(s): OpenAI pricing, Anthropic Pricing, Teneo research and analysis
Note(s): 1. Represents the cost for output tokens only
[11] OpenAI. 2023. "GPT-4 Turbo."; OpenAI. 2024. "GPT-4o."; OpenAI. 2025. "GPT-4.1."; OpenAI. Various years. "API Pricing." Accessed April 3, 2026. https://platform.openai.com/pricing

Current agent vendor packaging already hints at the direction of travel:

- OpenAI combines seat-based access with shared credits.
- Salesforce offers conversation pricing, flex credits and per-user licensing inside the same family.
- Intercom prices Fin at $0.99 per outcome.
- HubSpot sells seats, integrations, credits and, as of April 2026, sells outcomes at $1 per completed event.
- Figma offers seats with an entitlement of credits that can be drawn down.[12]

The pattern is clear: vendors are trying to preserve ARR-like predictability while creating a route for usage and value to scale.

4. What Are the Parallels?

The strongest parallel is not literal. Tokens are not the same thing as minutes and that matters. In telecom, the meter was imperfect but still broadly aligned to customer value. In AI, two tasks with identical token counts can have radically different economic significance. The real parallel is commercial. Both markets first charged for what the system naturally measured, then gradually moved toward what customers could plan around.

1. First, the raw engineering meter is better for finance teams than for buyers. Minutes, megabytes and tokens all work as internal units of record but are weak buying units for most non-specialists. In AI, that weakness is sharper because the meter often tracks activity more cleanly than impact.

2. Second, pooled usage is an elegant response to uneven demand. Family plans solved variation across people and devices; workspace pools and shared credits solve variation across teams, use cases and agent classes.

3. Third, ‘unlimited’ is never truly unlimited. Telecom hid the meter behind fair-use policies, hotspot caps and deprioritization. AI will do something similar via model routing, concurrency limits, channel restrictions, task-class rules, approval steps or post-threshold behavior.

[12] OpenAI. Various years. “API Pricing.”; Salesforce. 2025–2026. “Agentforce Pricing.”; Intercom. 2025–2026. “Fin Pricing.”; HubSpot. 2026. “Pricing.”; Figma. 2025–2026. “Pricing.”; Microsoft. 2025–2026. “Copilot Pricing and Copilot Studio Documentation.”; Zendesk. 2025–2026. “Automated Resolutions and Pricing.” Accessed April 3, 2026.

There is also a parallel in where long-run monetization migrates. As voice and text commoditized, telecom increasingly sold service class: speed, coverage, priority, bundles and guarantees. AI is likely to follow the same path. As model costs fall and more vendors access similar foundation models, the premium layer moves upward into workflow ownership, autonomy, accuracy, latency, governance, compliance and integration depth.

The main difference is that AI can sometimes charge for completed work, not just access. That means the end state should be more segmented than telecom and more outcome-led: internal assistant use may be effectively unmetered while external or mission-critical agents may still be tied to conversations, resolutions, workflows or specific business events.

5. What We Can Learn From This

The lesson for software vendors is not to copy telecom's old meters. It is to copy the way telecom matured away from them while recognizing that AI has one option telecom never really had: pricing the work itself.

1. First, keep raw tokens as an internal operating unit wherever possible. They are useful for cost accounting, model routing and margin analysis but rarely the best customer-facing price metric. If a buyer cannot explain the meter to a budget owner in one sentence, it is probably too low-level to anchor the commercial model.

2. Second, use a two-part architecture by default: a platform or seat commitment that preserves revenue quality, plus an included pool of agent capacity, plus a clear way to buy more. That can be overage, tier upgrades, additional committed pools or new packages for specific agent families. The objective is to smooth variability and channel it into a structure buyers can live with.

3. Third, pick the most business-facing meter you can defend. For narrow use cases, that may be documents summarized, invoices processed, cases resolved or workflows completed. For broader suites, credits can be a useful translation layer that simplifies buying and selling. Tokens should generally sit underneath the hood unless action-level cost volatility is so extreme that abstraction becomes risky.

4. Fourth, be honest about what each meter is doing. Tokens are useful for cost recovery and margin management. They are not, on their own, a compelling story of value. Where attribution is clean, outcome- or workflow-based pricing should take precedence over pooled or subscription models rather than sit at the edge of the portfolio.

5. Fifth, price adoption before extraction. In early markets, generous included usage is often strategically rational even if gross margins compress. The point is to build habit, reveal the usage curve and create evidence of ROI. Over-monetizing early agent usage can suppress the very behavior vendors need to understand and expand.

Over time, the most defensible premiums will sit less in raw compute consumption and more in the reliability and scope of the work done: better automation rates, fewer escalations, safer actioning, richer integrations, stronger auditability and faster time to value. That is the AI equivalent of telecom's move from selling minutes to selling plan quality.

So: is agentic pricing 'back to the future'? Yes, but only in one sense. The category is replaying telecom's commercial evolution in compressed time.

The winners will not be the companies that cling to tokens forever or the companies that declare every agent outcome-priced from day one. They will be the ones that make AI easy to buy on the surface, precise to meter underneath and increasingly tied to the value of the work completed as the market matures.

Exhibit 4. A Practical Playbook for Pricing Agents in the Next Two to Three Years

Design principle	Practical implication
Hide the raw meter	Use tokens to manage cost internally but present the simplest external unit that still protects margin.
Pool before you punish	Default to workspace or account-level included usage rather than rigid per-seat caps, especially early in adoption.
Price the job when possible	Prefer conversations, resolutions, workflows or outcomes where attribution is clear and disputes are manageable.
Make guardrails explicit	Use alerts, admin controls, pause modes, allowable overage and true-up rules to make variable usage feel safe, not scary.
Expect margin to migrate upward	Over time, differentiate on workflow coverage, accuracy, autonomy, governance and service levels, not raw model cost.
Bundle the assistant, meter the agent	Include everyday internal AI assistance in the core bundle; reserve separate pricing for external-facing, high-cost or high-stakes agent workflows.
Bundle in extras	Add useful adjacent features at a discount so customers feel they are getting more for the money. In AI, that could mean bundling analytics, governance, security, connectors or premium support into higher tiers instead of charging separately for every little thing.

Author

Sam Little
Senior Director

About Teneo's Private Equity Growth Practice

Teneo's Private Equity Growth Practice partners with PE sponsors and PE-backed management teams to unlock growth where it matters most: sharper market focus, stronger pricing and packaging, more effective go-to-market execution, and AI-ready commercial models. With deep experience in software, technology, and tech-enabled services, the practice supports clients from diligence to exit, helping them build revenue, improve profitability, and grow enterprise value.

Sign up to our quarterly Private Equity Growth Roundup newsletter to receive early access to our research.

Teneo
The Global CEO Advisory Firm

Teneo is the global CEO advisory firm.

We partner with our clients globally to do great things for a better future.

Drawing upon our global team and expansive network of senior advisors, we provide advisory services across our five business segments on a stand-alone or fully integrated basis to help our clients solve complex business challenges. Our clients include a significant number of the Fortune 100 and FTSE 100, as well as other corporations, financial institutions and organizations.

Our full range of advisory services includes strategic communications, investor relations, financial transactions and restructuring, management consulting, physical and cyber risk, organizational design, board and executive search, geopolitics and government affairs, corporate governance and ESG.

The firm has more than 1,800 employees located in 50+ offices around the world.

teneo.com