

Why Microsoft is repricing enterprise AI as cloud compute

The bet is clean for developers and far riskier for knowledge work, where value resists measurement.



Every so often, a product launch matters less for what it does than for how it is sold. Microsoft Copilot Cowork is one of those moments, and the signal is in its pricing.

Why the pricing model matters more than the product

- Microsoft is pricing Cowork like cloud compute, charging for AI work performed rather than seats licensed.
- The model fits developers cleanly, since AI-generated code converts directly into output leaders already track.
- It is riskier for knowledge work, where value surfaces as sharper judgment and less friction, not countable units.
- The deeper metric is intelligence per token, the useful reasoning an enterprise gains from each credit it spends.

Charging for AI work, not AI access

Our read is straightforward. Microsoft is testing an enterprise AI pricing model that looks more like cloud infrastructure than traditional software licensing. That model works cleanly for developers, because AI-generated code converts almost directly into measurable output. It is a far riskier bet for knowledge workers, whose gains, better decisions, fewer late nights, sharper judgment, rarely show up as a number anyone can point to.

So, the product isn't the real story. The real story is a commercial question. Can enterprise AI evolve from a licensed productivity tool into a metered business utility? Where organizations pay for the intelligence they consume rather than the software they own. And as the market matures, a sharper question follows. How much useful intelligence does an enterprise extract from every token or credit it spends?

What Cowork actually does

Microsoft introduced Copilot Cowork through its Frontier early-access program on March 30, 2026. The headlines went to the technology, understandably so. Cowork was not built to answer prompts. It was built to work across a company's entire Microsoft 365 environment. It reads emails, Teams conversations, SharePoint documents, and meeting notes, then acts on what it finds. It can compare thousands of policy documents to flag version differences. It can research a company's own data and assemble a board presentation from what it learns. It can coordinate multi-step workflows, draft documents, and follow up on open actions, largely without handholding.

We experienced Cowork firsthand through the Frontier program before general availability. What stood out was not the sophistication of the underlying AI. It was how seamlessly Cowork orchestrated research, document creation, analysis, and multi-step work across Microsoft 365. It felt less like using an assistant and more like collaborating with a capable coworker. That is a real shift, not from chatbot to assistant, but from assistant to something closer to a colleague.

The announcement that mattered more

On June 16, 2026, Cowork reached general availability. The product news was the headline. The pricing model, buried beneath it, may prove more consequential. Cowork now runs on Copilot Credits. Each credit costs a cent, and every task draws down credits based on the model used, the depth of reasoning required, the tools invoked, and the complexity of the workflow. Individual tasks can range from a few cents to several dollars. Instead of relying on a flat license alone, Microsoft has added a consumption-based layer—not a subscription tweak—but treating AI less like software and more like cloud compute. And pricing models do more than set revenue. They shape behavior, budgeting, and governance. Every major platform reaches a point where its commercial model matters as much as its tech. Cloud computing did. Enterprise AI may arrive at that same inflection point soon.

This wasn't a one-off decision

A few weeks earlier, GitHub Copilot made a similar move, and it's worth pausing on why. When GitHub Copilot launched in 2022, the pricing was refreshingly simple. A flat monthly fee, unlimited use. That made sense, because Copilot mostly generated code completions and answered programming questions. Then it stopped being simple. GitHub Copilot grew into something that could review pull requests, analyze full repositories, fix bugs, and plan implementations on its own. Running an autonomous agent for thirty minutes is not the same transaction as autocompleting a line of code, yet customers paid for both identically.

So, on April 27, 2026, GitHub Copilot announced AI Credits, with usage-based billing from June 1. Simple completions stayed unlimited. But chat, autonomous agent sessions, and code review now draw down credits based on actual token usage. Seen alone, that's a pricing change. Seen alongside Cowork, it looks like a principle. Stop charging for access to AI. Start charging for the amount of AI work performed.

Where the logic holds, and where could break

The engineering economics behind GitHub Copilot are easy to justify. An agent working across systems for half an hour consumes far more compute than a one-line autocomplete. A flat subscription covering both was never going to hold. The economics get interesting once you ask who is paying, and for what. Take a developer using GitHub Copilot. If AI ships a feature three days early, the payoff is legible. Organizations already track output through deployment frequency, cycle time, and features shipped. Faster development lands directly in numbers that leadership watches.

Now take a finance leader using Cowork. Say it assembles tomorrow's board deck in thirty minutes instead of three hours. The productivity gain is obvious. But then what? The meeting still happens at the same time. The deck still ships once. The executive doesn't produce five extra board decks that week. The value is real.

It just shows up somewhere harder to quantify. Perhaps, the deck is better informed. Maybe, the executive spends the saved time preparing for hard questions instead of formatting slides. Maybe, it is simply one less late night. None of it lands on a spreadsheet the way a shipped feature does.

The case for consumption pricing, and why it still cuts against Cowork

Buyers already accept consumption pricing for cloud, so they will accept it for AI. That's a fair assumption. But cloud consumption maps to measurable performance. Knowledge-work value doesn't, which is precisely the problem this model must solve. This is the crux. With developers, AI usage and business output move in near lockstep. With knowledge workers, that link loosens. Value shows up as reduced friction, and friction is notoriously hard to price. That is what makes consumption billing a bigger gamble on Cowork than it ever was on GitHub Copilot. It is also where intelligence per token becomes the number to watch. The unit economics may matter less for the cost of an answer and more for the value that answer produces. Ten dollars of AI that prevents a poor decision is extraordinarily cheap. Ten dollars spent on work an employee could have done just as well is waste. Telling the two apart is the real enterprise challenge.

Will this pricing model become the industry standard?

That's an uncertainty worth watching closely. Cowork has answered the easier question. Autonomous AI can do meaningful work inside a real organization. The harder question is still open. Will enterprises pay for that work the way they pay for cloud compute, by consumption, even when the value resists measurement? If they do, Microsoft will not simply have shipped another AI assistant. It will have helped set the commercial blueprint for how enterprise AI is bought, governed, and justified. And the winning metric may prove to be less about how many credits an enterprise consumes, and more about how much intelligence it extracts from each one.

What this means for how enterprises buy AI

- Budget for AI as a variable operating expense, not a fixed subscription, and expect spend to swing with demand.
- Govern consumption deliberately. Once every task carries a price, employees will weigh delegating against doing it themselves.
- Shift ROI from cost per task to value per credit, measuring which models and workflows produce the best outcomes.
- Treat intelligence per token as a first-class metric, not tokens consumed, but the business value each one returns.

About Brillio

Brillio is The Enterprise AI Accelerator helping Fortune 1000 companies move from AI ambition to scaled impact, faster. Powered by our AI accelerator platform – Agentic Data and Application Management (ADAM), Brillio is one of the fastest-growing digital technology service providers, delivering transformation across five core workstreams: business-led transformation, customer experience transformation, AI and data engineering, digital engineering, and infrastructure engineering.

With 14 delivery locations across North America, Europe, and Asia and a team of over 6,000 customer-obsessed professionals, Brillio combines deep industry expertise, modern engineering, and accelerators to deliver measurable outcomes.

Headquartered in Dallas, Texas, Brillio serves clients globally with a commitment to speed, scale, and measurable impact.



<https://www.brillio.com/>

Contact Us: info@brillio.com

