uptime INTELLIGENCE

INTELLIGENCE UPDATE

A culture of token abundance collides with financial reality

Dr. Owen Rogers 23 Jun 2026

Over the past year, many of us — operators, end-users and analysts — have struggled to distinguish widespread AI hype from AI reality. AI vendors, that have a vested interest in amplifying this hype, optimistically talk about adoption, use-cases and successes. But understanding how enterprises are deploying AI in practice, and how successful these initiatives are, has been more tricky.

A recent FinOps X event in San Diego (US) provided an opportunity to hear from those directly responsible for understanding the costs and value of AI. The overall theme from the attendees was that AI is being used in production applications, but that the cost of AI — and the value of investments — is hard to quantify.

At the event, many enterprises across a gamut of industries talked about how generative AI is being increasingly used in production applications and integrated into everyday business processes.

Such developments are not restricted to technology giants. Companies presenting their AI projects included:

- Global financial institutions, such as BlackRock, HSBC, Capital One and Prudential.
- Major retailers, including Macy's and Grainger.
- Consumer brands, such as Estée Lauder and NRG Energy.
- Media and technology firms, including Pinterest, Shutterstock and Squarespace.
- Industrial, healthcare and nonprofit organizations, such as Bayer and Communities Foundation of Texas.

Case studies presented at the event described using AI to assist with software development and code generation, deploying generative AI tools to improve employee productivity and applying AI to customer-facing digital experiences. Other cases studies explored agentic systems capable of taking actions on behalf of users and building data-driven assistants that interact with proprietary business information.

AI has firmly taken hold in enterprises. But many organizations using AI shared a common challenge: understanding, allocating and controlling rapidly growing AI expenditure.

The token challenge

Many attendees — FinOps professionals whose job it is to prescribe procedures, best practices and tools to balance technology expenditure against business objectives — were upfront that they were now under pressure to control and optimize AI costs. Several attendees described a period of "token abundance" in which costs were ignored to drive adoption and discover new use cases. Tokens are fragments of words that represent the basic units of data manipulated by most large language models (LLMs); see *Where to deploy AI inference: a guide to the economics*.

In early 2026, some employees were encouraged to "tokenmax" — using AI as frequently as possible to write code, draft documents and automate tasks in the hope that widespread adoption would translate into productivity gains. A few organizations even had leaderboards tracking token consumption, with prizes for the most active users. Major model providers, such as OpenAI and Anthropic, were bundling tokens into extremely generous packages, giving enterprises little financial incentive to restrict usage.

Unrestrained usage, however, is not sustainable. In effect, major model developers were subsidizing adoption and absorbing the loss on some of their tokens, once the costs of R&D, infrastructure and service delivery are factored in. The aim of this "loss leader" strategy was to demonstrate value to enterprises, in the hope they use more tokens in the future, at which point the model developers will recoup their costs.

Recently, more capable models require even more expensive infrastructure for model training and inference, pushing up the cost of tokens. In addition, models are being asked to process larger inputs, generate larger outputs and perform increasingly complex tasks — all of which require more tokens.

An example demonstrates the issue: at the top of a public leaderboard of the most active users of Claude Code is a developer who has consumed 11 billion tokens (valued at $35,000 to Anthropic) on a plan that costs just $200 per month.

The rate of consumption of such "inference whales" is unsustainable for model developers, and several "frontier" AI labs have now made changes to their contracts, increasing bundle pricing and shifting more users toward per-token, consumption-based billing.

AI overspends have recently made headlines. Uber reportedly exhausted its entire annual budget for AI coding tools by April as developers rapidly adopted AI-assisted software development. Meta has reportedly warned their 6,000 employees that internal AI usage could cost the company billions of dollars in 2026, and is now developing its own coding tools, processes and spending caps.

Reuters recently reported that 71% of companies experienced AI cost overruns in 2025. The challenge, therefore, appears broader than a handful of high-profile examples suggest.

This combination of factors — more expensive models, a move towards per-token pricing and rapid growth in usage — has put organizations, specifically FinOps practitioners, under pressure to bring this period of token abundance under control.

A major issue for these practitioners is that token usage has not been adequately tracked or allocated during this period of rapid growth. The focus was on growth and development, not on

governance and cost management. Many organizations remain unclear on how AI expenditure maps to projects, teams or outcomes. Determining the return on investment for AI initiatives is difficult without understanding where the money is being spent, but such is the appetite for AI experimentation, few organizations planned how a return would be calculated.

Many attendees reported that their organizations introduced spending caps — as an emergency measure rather than a solution. Capping expenditure may prevent costs from spiralling, but it can also slow down experimentation and innovation. High token consumption is not inherently bad. If an organization is developing products, improving services or generating new revenue, substantial AI expenditure may be entirely justified. However, the value delivered by much of this AI usage remains unclear.

Early-day jitters

Similarly to the early adoption of cloud computing, AI adoption has been driven by a period of relatively unconstrained experimentation. Engineers were encouraged to use the tools, push their limits and discover value through usage. In some cases, this meant maximizing consumption rather than optimizing it.

As with cloud computing, costs increased as usage grew. AI and cloud both often rely on consumption-based pricing. It is reasonable for providers to charge more when customers consume more resources. It is the customer's responsibility to determine whether that additional consumption delivers sufficient value.

FinOps emerged to help organizations consume cloud services where appropriate without allowing costs to grow unnecessarily. Most enterprises use public cloud today, but relatively few allow employees to consume cloud resources without oversight.

In many organizations, AI governance has not yet reached the maturity that cloud governance enjoys today. Basic questions (who is using AI, for what purpose, at what cost, and with what outcome?) are often difficult to answer.

Organizations do not necessarily need to ration their AI usage, but they should be able to justify it.

Outlook

Imposing controls after a period of unrestricted experimentation is always difficult, particularly when AI tools have become embedded in workflows and are perceived as productivity multipliers. Any attempt to reduce usage or introduce friction is likely to encounter resistance.

But controls do not signal the end of AI adoption. In fact, they represent a key step in the maturity of AI as an enterprise technology. Goldman Sachs predicts global token consumption will rise from 6 quadrillion today to 120 quadrillion in 2030. Adoption will continue, but CEOs and CFOs will increasingly demand greater clarity on how investments in AI deliver business value.

Some have started looking for solutions: in June 2026, the Linux Foundation announced its intention to launch the Tokenomics Foundation, with organizations including Booking.com, JPMorgan Chase, Google Cloud, Microsoft, Oracle, IBM and Salesforce expressing an interest.

The initiative aims to establish standards, benchmarks and best practices for measuring AI costs and token use efficiency.

The next phase of enterprise AI adoption is likely to be defined not by the overall number of tokens, but by a better understanding of which tokens create value. Unless organizations can reliably connect AI expenditure to business outcomes, FinOps practitioners will struggle to bring greater visibility, accountability and discipline to AI spending.

The challenge extends beyond spending. Organizations operating their own models on private infrastructure are under pressure to manage usage, because GPU capacity remains expensive and finite. Every token consumed by a low-value task is capacity that cannot be used for a more important application, user or development project. As AI infrastructure becomes a strategic resource, organizations will need mechanisms to ensure that capacity is allocated where it delivers the greatest return.

To get there, enterprises need to shift from a culture of token abundance to one of token accountability.

Uptime Intelligence View

The shift from token abundance to token accountability marks a necessary maturation of enterprise AI. Rapid experimentation has helped organizations identify use cases and build familiarity with generative AI, but unchecked consumption risks turning AI from a productivity tool into an opaque and poorly governed cost centre. As pricing becomes more consumption based and internal AI infrastructure becomes more constrained, enterprises will need to treat tokens as measurable business resources rather than incidental by-products of adoption. The organizations that benefit most from AI will not necessarily be those that consume the most tokens, but those that can connect usage to outcomes, allocate capacity deliberately and manage AI spending with the same discipline now expected of cloud services.

ABOUT THE AUTHOR

Dr. Owen Rogers

23 Jun 2026

Dr. Owen Rogers is Uptime Institute's Senior Research Director of Cloud Computing. Dr. Rogers has been analyzing the economics of cloud for over a decade as a chartered engineer, product manager and industry analyst. Rogers covers all areas of cloud, including AI, FinOps, sustainability, hybrid infrastructure and quantum computing.

orogers@uptimeinstitute.com

About Uptime Institute

Uptime Institute is the Global Digital Infrastructure Authority. With over 4,000 awards issued in over 122 countries around the globe, and over 1,100 currently active projects in 80+ countries, Uptime has helped tens of thousands of companies optimize critical IT assets while managing costs, resources, and efficiency. For over 30 years, the company has established industry-leading benchmarks for data center performance, resilience, sustainability, and efficiency, which provide customers assurance that their digital infrastructure can perform across a wide array of operating conditions at a level consistent with their individual business needs. Uptime's Tier Standard is the IT industry's most trusted and adopted global standard for the design, construction, and operation of data centers.

Offerings include the organization's Tier Standard and Certifications, Management & Operations reviews and assessments including SCIRA-FSI financial sector risk assessment, the Sustainability Assessment, and a broad range of additional risk management, performance, availability, and related offerings. Uptime Education training programs have been successfully completed by over 100,000 data center professionals, such as the much-valued ATD (Accredited Tier Designer) and AOS (Accredited Operations Specialist). The Uptime Education curriculum has been expanded by the acquisition of CNet Training Ltd. In 2023.

Uptime Institute is headquartered in New York, NY, with offices in London, Sao Paulo, Dubai, Riyadh, and Singapore, and full-time Uptime professionals based in over thirty-four countries around the world.

For more information, visit www.uptimeinstitute.com