

AI token costs: Why enterprises need to stretch AI budgets, not just cap them

Token prices have been falling, yet enterprise AI bills keep rising. Most organisations are responding with caps and quotas, which make spend predictable but quietly ration the value AI can deliver. As agentic AI scales, the winners will be those that get more from every token, not those that simply buy fewer of them.

Gabija Cepurnaite, Senior Consultant

November 2026

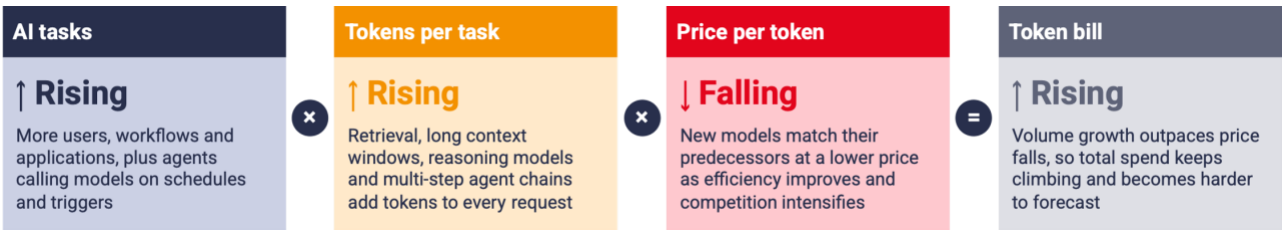
© STL Partners

Volume, not price, is driving the bill

The unit cost of AI has been falling. Model providers have routinely released models that match or exceed their predecessors at a lower price. In July 2026, for example, Google released Gemini 3.6 Flash, which it says costs less per token than the previous version and uses up to 17% fewer output tokens. Yet for many enterprises, the AI bill is moving in the opposite direction.

The reason is simple: the token bill is a product of price and volume, and volume is growing faster than price is falling. Three forces are driving it. Adoption is broadening, as AI moves from a handful of pilots into everyday workflows across customer service, software development, sales and operations. Workloads are getting heavier, as retrieval-augmented generation, long context windows and reasoning models increase the tokens consumed by each request. And, most significantly, agentic AI is beginning to scale.

Figure 1: Enterprise token bills rise because volume growth outpaces falling prices



Source: STL Partners

AI software vendors are already responding by shifting the risk onto customers. When GitHub moved Copilot to usage-based billing in June 2026, it said its previous model was "no longer sustainable" because "a quick chat question and a multi-hour autonomous coding session can cost the user the same amount". We expect other vendors facing the same economics to follow, turning AI tools that many enterprises budgeted as fixed licence costs into variable costs.

Agents change the shape of the problem

Agents do not just consume more tokens – they consume them differently. Human usage is naturally bounded by working hours and typing speed. Agents call models on schedules and event triggers, often calling other agents in turn. A single task can chain dozens of model calls, with no one in the loop to notice when it is consuming far more than it should.

The number of agents is also set to grow sharply. Agentic capabilities are rapidly being built into mainstream enterprise software, and a growing share of day-to-day work decisions is expected to be made autonomously. Because these capabilities arrive as features rather than as projects enterprises consciously choose to fund, token consumption can scale well ahead of any budget decision.

As a result, even well-resourced organisations are being caught out. Uber's CTO has said the company has already exhausted its 2026 AI budget by April 2026, just four months into the year, driven by heavy use of AI coding tools. According to a company spokesperson, Uber then limited employees to $1,500 a month in token spend per AI coding tool. Accenture has also introduced limits: since October 2026, it has centrally funded just $29 a month per employee for Claude, with any usage above that charged to projects.

Most enterprises cannot see where their tokens go

Uber at least knew when its budget had run out. For many enterprises, the more fundamental obstacle is visibility rather than price. Large organisations rarely consume AI through a single provider: models are accessed through several hyperscaler platforms, embedded in SaaS tools, called from in-house applications

© STL Partners

and, increasingly, invoked by agents. Each provider console shows its own slice of this estate, but none shows the whole picture.

The situation echoes the early days of cloud. Finance teams receive an aggregate invoice from each provider but cannot tell which team, application or agent drove the spend, let alone whether it delivered any value. Responsibility for AI budgets is increasingly landing with the FinOps teams that spent the last decade bringing cloud costs under control, yet AI spend behaves differently from cloud infrastructure. It is driven by individual prompts and model choices, often made by people and systems far removed from the budget holder, which makes it far harder to forecast and manage centrally.

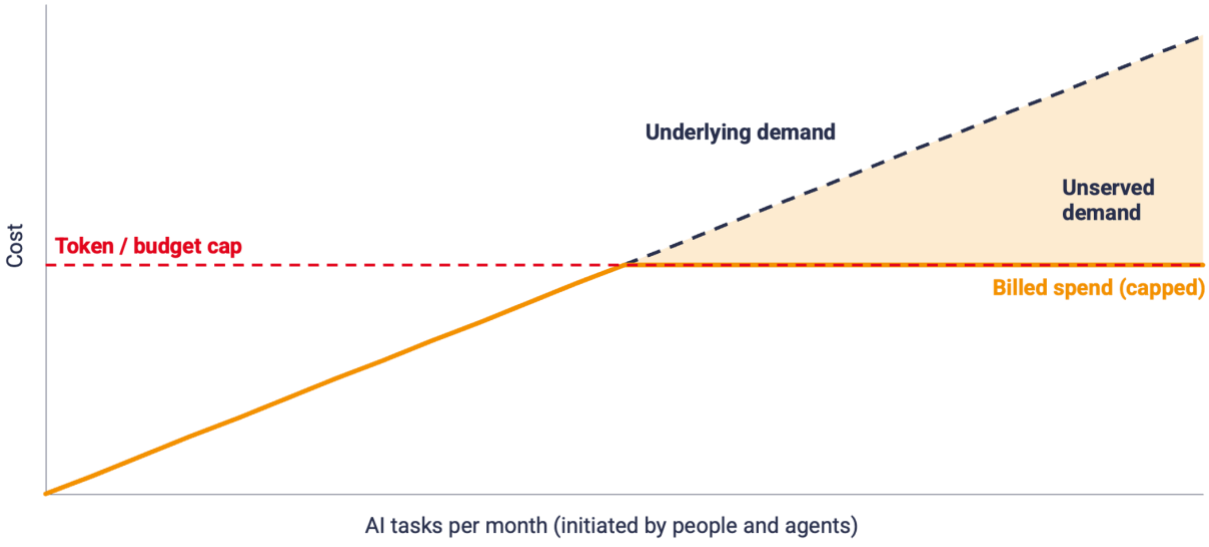
These blind spots create governance exposure as well as cost risk. An organisation that cannot attribute its token spend is unlikely to know which sensitive data is flowing to which models and providers. As such, cost control and AI governance are increasingly closely linked, and enterprises are likely to benefit from addressing them together rather than in isolation.

Caps keep spend predictable, but ration value

Faced with escalating bills, the instinctive response is to cap. A common pattern is emerging: organisations that began by giving employees free choice of models found that users defaulted to premium models for routine tasks. They responded with role-based access, model quotas and hard budget limits. Users who exhaust their allocation are downgraded to cheaper models or blocked until the quota resets.

This approach achieves its immediate objective: reported spend becomes predictable. However, caps constrain consumption rather than underlying demand. Once quotas are reached, use cases are deferred, requests are routed to less capable models and access is curtailed. None of this appears on the AI invoice, so the true cost of the cap remains hidden: a growing backlog of unserved demand as more employees, applications and agents compete for a fixed budget.

Figure 2: Spend caps make AI spend predictable, but constrain consumption rather than demand



Source: STL Partners

The shift towards agentic AI intensifies this trade-off. While employees can typically absorb a delay until a quota resets, an automated workflow that stalls mid-process introduces operational risk. Blanket restrictions

© STL Partners

that were acceptable during the pilot phase therefore risk constraining the very productivity gains that underpinned the original investment case.

The goal is more value per token, not fewer tokens

The more productive question is not "how do we spend less on tokens?" but "how much more can we do within the same budget?" Answering it requires enterprises to manage AI consumption as a portfolio, matching each workload to the right model, cost and level of control. We see four levers that, applied together, can expand what a fixed budget delivers.

- **Right-fit model routing.** Most enterprise tasks do not need frontier capability. AT&T, which processes around 45 billion tokens a day, says its AI gateway routes each prompt to the most cost-effective model capable of completing the task, reducing its AI costs by as much as 90%. The underlying principle is that model choice should be a dynamic, per-request decision rather than a fixed procurement choice, with premium models reserved for the minority of tasks that genuinely require them.
- **Cost attribution.** Metering spend per request, team, application and agent turns an opaque invoice into a manageable budget, and separates legitimate heavy workloads from inefficient habits. It also enables chargeback to business units, giving budget holders a direct incentive to optimise usage, and provides the foundation for the governance controls discussed above.
- **Caching and dynamic budgets.** Reusing responses to repeated requests and shifting budget towards workflows that demonstrably deliver value can free headroom without blanket restrictions. Many enterprise queries, from customer service requests to routine coding tasks, follow recurring patterns, so caching can avoid a meaningful share of repeat inference. Dynamic budgets, in turn, allow allocations to follow demonstrated value rather than historical spend.
- **Measuring the backlog.** Tracking deferred workflows, forced downgrades and quota-reset requests exposes the demand that caps are hiding and can help build the business case for optimisation. Without this data, the opportunity cost of caps remains invisible to decision makers and budget discussions continue to focus on reported spend rather than on the value left unrealised.

These capabilities increasingly sit in an AI gateway: a control layer between users, applications and agents and the models they call. Many large enterprises are building their own, but the maintenance burden is easy to underestimate. New models, provider APIs and security requirements arrive faster than most internal teams can absorb them. As a result, a gateway built for today's model landscape can quickly become a constraint in its own right. Enterprises should therefore weigh whether building and maintaining this layer is a source of competitive advantage or whether it is better sourced from a specialist partner that can absorb the ongoing integration and compliance burden.

Conclusion

Falling token prices have not reined in enterprise AI bills on their own, and spend caps only hide the problem. As agentic AI moves into production, the question for enterprises will shift from how much they spend on AI to what each token buys. Enterprises that establish visibility, attribution and routing across their AI estate now will be better placed to scale agentic workloads without either losing control of spend or rationing value. Early evidence from agentic coding suggests that usage is growing faster than token prices are falling. If that continues, the organisations that extract the most value from each token will pull ahead, while others find their AI ambitions limited by budget.

© STL Partners

Frequently asked questions

What is token cost in AI?

For an enterprise, the token bill is a product of price and volume: the price per token charged by model providers and cloud platforms, and the number of tokens consumed. Price per token has been falling, as each new generation of models has delivered comparable capability at a lower price, but volume is growing faster than price is falling.

Why are enterprise AI costs rising when token prices are falling?

Because volume is growing faster than price is falling. Three forces drive it: adoption is broadening as AI moves from pilots into everyday workflows; workloads are getting heavier as retrieval-augmented generation, long context windows and reasoning models increase the tokens consumed by each request; and, most significantly, agents are arriving.

How do AI agents affect token costs?

Agents do not just consume more tokens; they consume them differently. They call models on schedules and event triggers, often calling other agents in turn, so a single task can chain dozens of model calls with no one in the loop to notice when it is consuming far more than it should.

Do AI spending caps and quotas work?

Caps make reported spend predictable, but they constrain consumption, not demand. When quotas bite, use cases are deferred, requests fall back to less capable models and access lapses. None of this appears on the AI invoice, so the cost of the cap is a growing backlog of unserved demand.

How can enterprises stretch their AI budget instead of capping it?

We see four levers: right-fit model routing, since most enterprise tasks do not need frontier capability; cost attribution per request, team, application and agent; caching and dynamic budgets that shift spend towards workflows that demonstrably deliver value; and measuring the backlog of deferred workflows, forced downgrades and quota-reset requests.

What is an AI gateway?

An AI gateway is a control layer between users, applications and agents and the models they call, and it is increasingly where these levers sit. AT&T, which processes around 45 billion tokens a day, says its AI gateway routes each prompt to the most cost-effective model capable of completing the task, reducing its AI costs by as much as 90%.

How is FinOps for AI different from cloud FinOps?

FinOps teams that spent the last decade bringing cloud costs under control are now being handed AI budgets too. But AI spend behaves differently: it is driven by individual prompts and model choices, often made by people and systems far removed from the budget holder, and no single provider console shows the whole estate.

© STL Partners

Gabija Cepurnaite is a Senior Consultant at STL Partners, specialising in AI monetisation and edge computing.

Get in touch with the author to learn more
gabija.cepurnaite@stlpartners.com

Or visit STL Partners' Edge Hub
www.stlpartners.com/edge-computing

© STL Partners