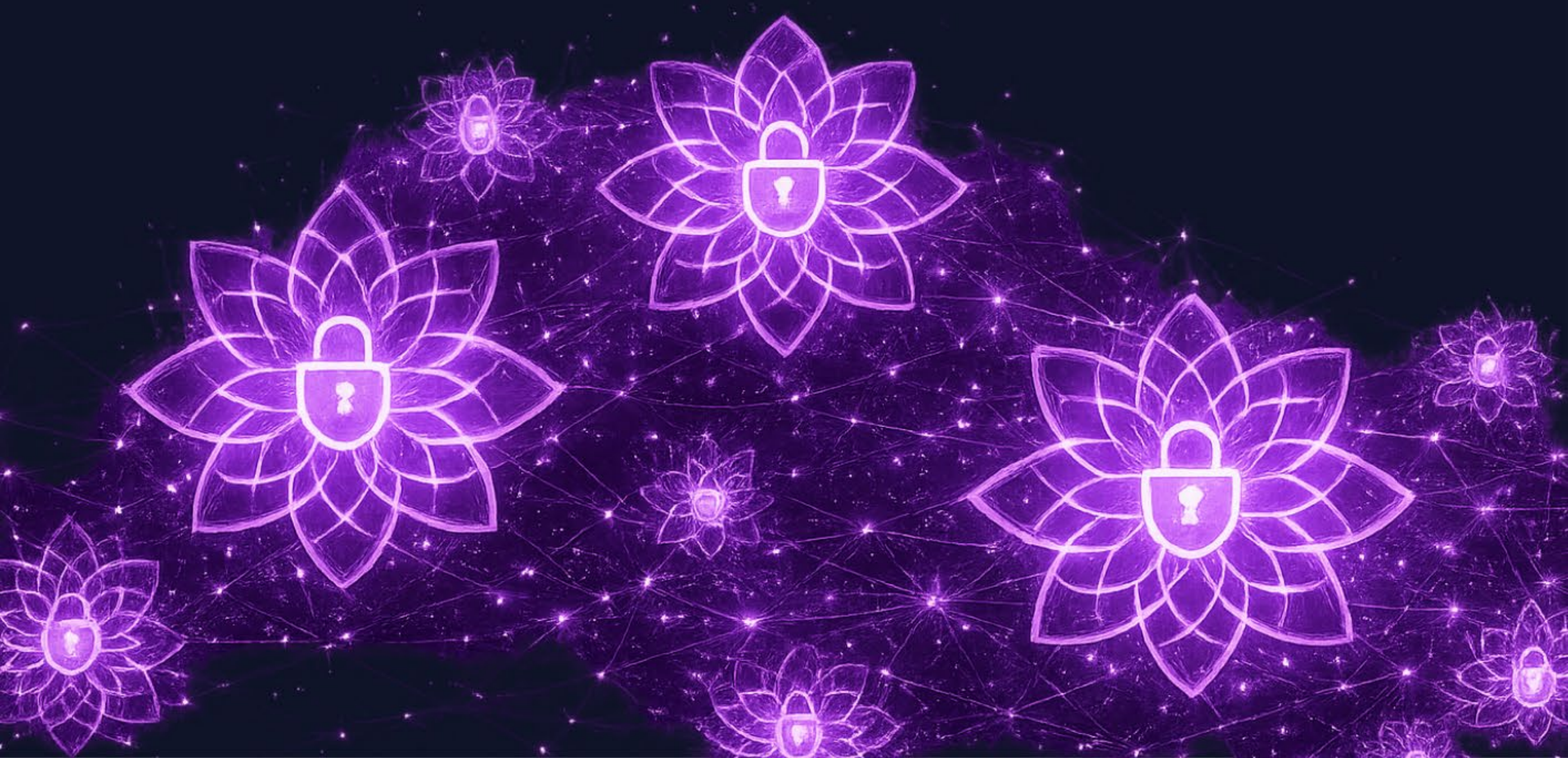
Authentication for the Intelligent Age

Controlling AI Token Cost

Author: FIOR Group

Date published: May 2026
Version: 1.0



Escalating Enterprise AI Token Costs and Governance Challenges

A major theme in enterprise AI discussions in recent weeks has been the rapid escalation of “token spend” — the cost enterprises incur every time employees, applications or AI agents interact with large language models such as OpenAI or Anthropic.

In practical terms, tokens are the billing units of AI systems. Every prompt, document upload, reasoning step and generated response consumes tokens. As enterprises move from occasional chatbot use to large-scale deployment of coding agents, autonomous workflows and AI-assisted operations, token consumption has started to resemble cloud-compute spending: variable, difficult to forecast and increasingly strategic.

Recent Industry Discussion

- Boston Consulting Group executives have discussed internal tensions between CIOs encouraging aggressive AI adoption and CFOs concerned by rapidly growing token bills.
- Engineering teams have reported AI usage bills reaching six figures for very small developer groups. One example cited a four-person AI engineering team generating a $113,000 monthly bill.
- Analysts increasingly describe AI as becoming a major line item in enterprise software budgets, displacing traditional SaaS expenditure patterns.
- Anthropic revised expected enterprise usage costs for Claude Code sharply upwards, estimating many enterprise developers may consume $150–250 per month each in AI usage alone, with some users far exceeding that.

Why Token Costs Are Difficult to Control

- Usage can grow non-linearly as autonomous agents and reasoning models consume very large volumes of tokens.
- Many enterprises lack central visibility into AI consumption across departments and applications.
- Different business units frequently connect independently to multiple AI providers with little oversight.

- AI costs combine per-user fees, API metering, inference costs and autonomous agent activity.
- Small changes in prompts or workflows can materially change token usage and cost.

Selected Statistics

Statistic	Reference
Enterprise AI/software spending reportedly increased nearly 58% year-on-year, largely driven by AI adoption.	Industry reporting
Anthropic reportedly has over 1,000 enterprise customers spending at least $1m annually.	Market analysis
Many enterprise developers are estimated to consume $150–250/month in token costs.	Anthropic estimates
One four-person AI engineering team reportedly generated a $113,000 monthly AI bill.	Industry reporting
OpenAI projected annualised revenue reportedly exceeded $12.7bn in Q1 2026.	Market analysis

Conclusion

The strategic concern emerging inside large enterprises is that AI tokens are becoming analogous to cloud compute in the early AWS era: initially decentralised, rapidly adopted, poorly governed and capable of producing major uncontrolled cost expansion before governance catches up.
As a result, enterprises are increasingly recognising that managing AI consumption may become as important as deploying AI itself.

FIOR solves this problem

The FIOR AI Gateway is a low cost easily deployed software appliance designed to give the enterprise total control over what agents enter the network or are spun up by employees. Its primary purpose is cyber security – it is effective in preventing agentic attack by creating a cryptographically secure unique identity for every single agent either at creation (for internal agents) or at arrival at the network boundary (external). Then policies selected by the security team are applied to give it permissions, with bad actors breaching policy revoked in under 3 milliseconds. This triggers a revocation alert sent to all of FIOR's customers thus preventing large scale industry wide attacks.

But it also allows the enterprise to create a governance framework for internal AI use, controlling what employees are able to do with agents. Finally, cost control becomes real time. The system can send the finance team an email every day to reveal what agents have been used, of what category and at what cost.