

Nutanix Agent Gateway: The Control Plane for Agentic AI

A Visual Guide to Managing Token Economics and Control in the Enterprise Agentic AI World

As organizations race to operationalize autonomous agents, traditional infrastructure and governance models struggle to keep pace with escalating token costs, context proliferation, and security risks. Nutanix Agent Gateway serves as a unified control plane to help organizations govern AI traffic, optimize token spend, and secure enterprise workflows across hybrid environments.

Agentic AI Is Amplifying Cost and Governance Risk

Autonomous agents are transforming workflows and multiplying operational exposure.



5 to 30x

More Tokens Per Task:

“Advanced agentic models with reasoning will typically consume 5 to 30x more tokens than basic main-stream chatbots.”¹



93%

Exceed AI Budgets:

Despite falling per-token prices, enterprise LLM spend has tripled—leaving 93% of surveyed organizations exceeding their AI budgets due to unmonitored agent execution. [\(McKinsey\)](#)



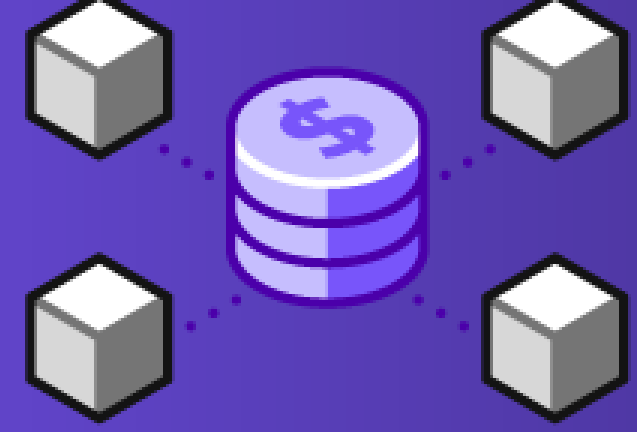
40%+

Project Cancellation Risk:

Gartner® predicts that “by 2027, over 40% of agentic AI projects will be canceled due to escalating costs, unclear business value or inadequate risk controls.”²

Bridging Control and Speed

Resolving the friction between AI Operators who demand control and AI Consumers who demand velocity.



Token Cost Control

Gain visibility into token consumption and optimize spend, preventing runaway costs and unexpected bill shock.



Observe

Track token telemetry by user, agent, team, and model.



Control

Enforce token budgets and rate limits per API.



Optimize

Apply token-based rate limiting, model fallback, and load balancing to manage cost.



Unified Governance

Govern agent access to enterprise tools and private data sources via MCP servers.

Shadow AI Mitigation



Role-Based Access

Centralized RBAC governs permissions across agents and systems.



Audit Trails

Track and review every MCP request.

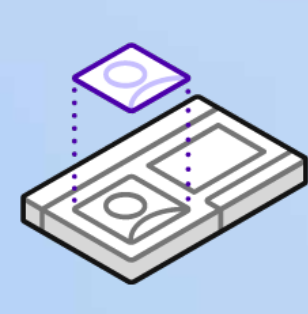


Tool Filtering

Control interactions with MCP servers and enterprise APIs.

Why Nutanix Agent Gateway? (Top Drivers)

Autonomous agents are transforming workflows and multiplying operational exposure.



One Platform:

Consistent identity management, Role-Based Access Control (RBAC), and API key configurations across Nutanix Agent Gateway and Nutanix Private Inference.



Predictable Economics:

Curb token sprawl by shifting appropriate workloads to self-hosted models on Nutanix Private Inference.



Unified Governance:

Secure tool calls and private data access using MCP servers and granular access controls.



Developer & Operator Simplicity:

An agent can move from a public frontier model to a private, self-hosted model running on Nutanix Private Inference with no new credentials and no code rewrite.



Deployment Flexibility:

Connect applications and agents to the models and environments that best fit your performance, cost, and regulatory requirements.

The Bottom Line?

Nutanix Agent Gateway gives enterprises the control plane required to master token economics, enforce security, and scale enterprise AI with total confidence. Run Agentic AI Anywhere—any app, any model, any environment.

[Learn More](#)

1 - Source: Gartner, Inc., AI Inference Tiering Matters More Than Foundational Model Selection, Will Sommer, 6 July 2026

2 - Source: Gartner, Top Actions to Drive Success in Building Agentic AI Solutions, Aaron Harrison, Haritha Khandabattu, 9 April 2026