citrix

What is an AI gateway?

The operational control layer for governed AI

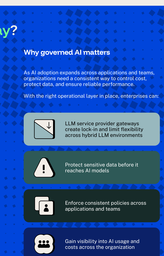


AI is scaling. Enterprise AI needs governance.

AI success depends on secure, reliable operations with visibility into usage, performance, and cost. But for many enterprise environments, challenges stand in the way of AI success.

- AI policies are defined per application, resulting in fragmented, inconsistent enforcement.
- Token usage and API spend lack visibility
- Sensitive data can flow into or out of LLMs without standard inspection
- AI agents and MCP servers introduce new types of threats

To scale AI successfully, organizations need centralized controls that manage how AI traffic is routed, secured, and monitored.



What is an AI gateway?

A centralized control layer for AI traffic

An AI gateway provides a centralized layer that governs how AI services are accessed, secured, and monitored across applications, teams, and environments.

It enables organizations to:

- Control how AI services are accessed and used across models and providers
- Protect sensitive data and enforce consistent security policies across every AI interaction
- Gain visibility into AI usage, performance, and cost drivers
- Optimize reliability and performance across AI services

Positioned between enterprise applications and AI services, the gateway gives organizations the control, visibility, and policy enforcement needed to scale AI without operational risk.

Why governed AI matters

As AI adoption expands across applications and teams, organizations need a consistent way to control cost, protect data, and ensure reliable performance.

With the right operational layer in place, enterprises can:

- LLM service provider gateways create lock-in and limit flexibility across hybrid LLM environments
- Protect sensitive data before it reaches AI models
- Enforce consistent policies across applications and teams
- Gain visibility into AI usage and costs across the organization
- Ensure consistent availability and performance across AI services

Without AI Gateway vs. With AI Gateway

Unmanaged AI	Governed AI
Policy enforced per app, inconsistently	Centralized routing and controls of model requests
Limited visibility into token usage	Rich visibility into token consumption
Basic policy enforcement	Policy-driven routing across models and providers
Sensitive data sent to models without inspection	Prompt inspection and LLM redaction measures
Fragmented logging and audit trails	Audit-ready observability of AI activity

How an AI gateway enables enterprise-scale AI operations



To scale AI successfully, organizations need five critical capabilities: risk management, cost control, performance optimization, data protection, and observability.

01

Risk & model governance

AI systems introduce new operational risks. An AI gateway helps manage risk and control model usage through:

- Model access controls to determine who can use which AI models
- Model governance to define which models are approved for enterprise use
- Prompt templating and policy enforcement
- Detection of prompt injection and misuse patterns
- Token quotas and usage limits to prevent uncontrolled consumption
- Model-aware routing based on performance, capability, or cost

Outcome: Reduce operational risk by controlling which models are approved for enterprise use, enforcing access policies, and eliminating shadow AI consumption.

02

Cost control & financial accountability

Organizations must understand where AI spend originates and how it grows across teams and applications. An AI gateway provides:

- Token-level usage visibility across teams and applications
- Full observability into requests, token consumption, and spend trends by team and application
- Centralized reporting on AI activity and spend
- Flexibility to route workloads across multiple LLM providers — optimizing cost and performance without being locked to a single model ecosystem

Outcome: Control AI spending, prevent unexpected costs, and improve financial accountability for AI consumption by gaining full visibility into token usage.

03

Performance & user experience

AI services must operate as reliable, high-performance infrastructure. An AI gateway ensures fast, consistent AI experiences through:

- Latency-aware routing to the best-performing models and providers
- Automatic failback and spillover routing when models fail or degrade
- Rate limiting and protection against denial-of-service attacks

Outcome: Deliver fast, reliable AI experiences, ensuring applications stay responsive for the best end-user experience.

04

Data protection & privacy governance

Sensitive data must be protected before it reaches AI models. An AI gateway enforces safeguards such as:

- Prompt templating to enforce consistent, policy-compliant prompt structures across AI requests
- PII/PHI and sensitive information redaction and anonymization
- Consistent content filtering policies
- Integration with specialized AI security platforms to for layered inspection and filtering

Outcome: Reduce regulatory risk, prevent unintended data exposure, and enforce consistent AI interactions before requests ever reach the model.

05

Compliance & audit readiness

AI activity must be auditable and transparent. An AI gateway helps support regulatory alignment through:

- Centralized routing of AI traffic
- Audit-ready logging of prompts, responses, and model interactions
- Consistent policy enforcement across applications
- Observability into AI requests and usage metrics
- Integration with existing SIEM, observability, and security platforms.

Outcome: Deliver fast, reliable AI experiences, ensuring applications stay responsive for the best end-user experience.

Why an independent, Enterprise AI Gateway?

Cloud AI platforms like AWS Bedrock, Google Vertex AI, and Databricks offer built-in AI gateways — but they're optimized for their own ecosystems. For enterprises that need to route across multiple providers, an Enterprise AI gateway delivers control that no single platform can match:

- Multi-provider routing — choose the right model for the right workload based on cost, performance, or compliance requirements, without being locked to one provider's ecosystem
- Consistent policy across every model — a single enforcement point for access controls, data protection, and compliance, regardless of which cloud AI platform is handling the request
- Enterprise-managed policy enforcement — governance defined and controlled by your organization, not inherited from a provider's defaults, with the ability to abstract complexity as new AI services are rolled out

Enterprise AI succeeds when AI traffic is governed at the infrastructure layer. An AI gateway provides the visibility, control, and optimization required to scale AI safely and efficiently.

Executive takeaway

Governing AI starts at the traffic layer

As AI capabilities expand across applications, developer tools, and automation platforms, governance must extend to the infrastructure layer where AI requests are executed.

An AI gateway enables organizations to:

- Reduce operational risk
- Enforce a consistent AI policy
- Protect sensitive data
- Manage token-driven costs
- Route across multiple LLM providers

See NetScaler AI Gateway in action.

citrix
Secure the Work.

© 2026. Citrix Systems, Inc. All rights reserved.