Solution Brief

Securing AI with Netskope One

Netskope One AI Security provides a single platform to govern your AI ecosystem and protect your data. It secures users and agents across public SaaS, private AI tools, and agentic workflows. Combining high-performance access with context-aware zero trust controls, we enable organizations to move from AI experimentation to AI advantage.

Quick Glance

- Gain unified visibility and control across all AI traffic: human users, applications, and autonomous agents.
- Protect sensitive data across the entire AI lifecycle, from model training to real-time prompts and responses.
- Defend against AI-specific threats including prompt injection, jailbreaking, and data exfiltration in real time.
- Harden private AI models before deployment with automated adversarial testing across 18,000+ scenarios.
- Secure agentic AI workflows by decoding and governing model context protocol (MCP) traffic.

AI spending will surpass $867B by 2029—a proxy for the scale of emergent data security challenges.

Data reference: IDC, Worldwide Artificial Intelligence IT Spending Forecast, 2025–2029

The challenge

AI has moved from sporadic experimentation to the business front line. Every department, workflow, and customer interaction is becoming AI-enabled, with organizations deploying customer service chatbots, automating critical decisions, and using autonomous agents. The opportunity is huge, but so is the risk:

- Each use case introduces data into AI systems that were not designed with security as a foundation.
- Attackers are exploiting LLMs through prompt injection, jailbreaking, and data extraction.
- Meanwhile, employees using unmanaged AI tools create blind spots that are difficult to address with traditional security measures.

Organizations need unified protection to secure data, users, and agents while maintaining the speed of innovation AI promises.

Netskope for AI security

Netskope One AI Security secures users, agents, applications, and data across the entire AI ecosystem within a single unified platform. Whether your workforce is accessing public generative AI (genAI) SaaS tools, your developers are building private AI-powered applications, or autonomous agents are interacting via APIs and MCP, Netskope provides real-time visibility and context-aware protection. By unifying security across all AI traffic, we enable organizations to adopt AI at scale without expanding risk.

Secure every AI interaction across your ecosystem

Enterprise AI involves complex interactions, with AI-powered applications and autonomous agents communicating at machine speed, often beyond the reach of traditional security tools.

Netskope One AI Security provides a zero trust access layer to monitor and secure these unique traffic flows, with flexible deployment methods to see and protect who (or what) is communicating, and where the traffic is going. For user-to-AI application traffic, the Netskope One Next Gen Secure Web Gateway (NG-SWG) manages access and provides granular visibility and protection for public genAI SaaS tools and embedded AI within SaaS apps, including identifying shadow AI instances of those applications. Netskope One Private Access (NPA) then controls appropriate user access to private AI applications.

Two purpose-built products close the gap created when agentic communications bypass traditional security perimeters:

- The **Netskope One AI Gateway** secures east-west traffic (app-to-LLM) as a virtual appliance deployed directly within your private environment. By centralizing authentication and traffic management across the major LLM providers, it ensures autonomous data flows remain governed.
- The **Netskope One Agentic Broker** decodes and secures MCP traffic between AI agents and external data sources and tools. It evaluates public MCP servers with dedicated MCP risk scoring, delivering supply chain risk management for AI at scale.

Enforcing granular access controls allows security posture to keep pace as agents gain greater autonomy.

> Combining high-performance with context-aware zero trust controls, Netskope enables organizations to move from AI experimentation and unlock AI advantage.

Real-time defense against AI threats

Once traffic channels are established, Netskope applies unified inspection engines to every interaction in real time—regardless of whether it comes from a human or an autonomous agent.

Netskope One AI Guardrails provides a dedicated runtime defense layer, purpose-built for AI environments, with support for 29 languages. It mitigates sophisticated attacks, including prompt injections and jailbreak attempts, by analyzing every request and response in depth and identifying the multi-stage intent behind linguistic exploits. Content moderation is also a focus, detecting inappropriate content to ensure AI usage stays within your organization's risk tolerance, protecting your corporate reputation.

Powered by the Netskope One platform's AI functionality (SkopeAI), every prompt and response passes through three Netskope engines simultaneously:

- **Netskope One Data Loss Prevention (DLP):** to identify sensitive data in motion.
- **Netskope One Threat Protection:** for malware and malicious links.
- **Netskope One AI Guardrails:** for AI-specific threats and content moderation.

When a single interaction involves both a jailbreak attempt and a data leak, Netskope correlates these under a single incident ID, eliminating the fragmented alerts typical of multi-vendor approaches.

Searchable conversation logs map directly to MITRE ATLAS and the OWASP Top 10 for LLMs, giving investigators a full context of user, prompt, intent, and application in a single unified view.

> Netskope One AI Security enables our customers to say yes to safe AI innovation.

Proactive model hardening before deployment

Private AI models represent a growing frontier for organizations. An organization must take full responsibility for the security of a private model, which can create risk when pushed to production without robust safety checks. Once live, hidden vulnerabilities can be exploited by sophisticated multi-turn attacks that trick LLMs into bypassing their own safety guardrails. A model that seems perfectly safe could be tricked through layered conversational prompts into leaking sensitive training records or intellectual property.

Netskope One AI Red Teaming closes this critical gap by automating adversarial simulations that uncover vulnerabilities before attackers can strike.

- Drawing from a library of more than 18,000 adversarial scenarios and seed prompts, it systematically stress-tests models against prompt injection, jailbreaking, data leakage, and malicious use scenarios.
- Includes coverage for complex skeleton key and crescendo attacks that layer multi-stage conversations to circumvent standard defenses.
- Manual testing alone cannot keep up with the speed of modern AI development. Netskope One AI Red Teaming identifies model vulnerabilities before attackers find them.
- Integrating directly into CI/CD pipelines via APIs, every code change or model update is automatically screened for new security risks before production release.

Scheduled simulations track how risk profiles change over time, shifting model security from passive observation to active defense. Altogether, it means organizations can confidently transition from experimentation to production-ready AI, knowing their foundation is secure.

Unified AI data governance at scale

Data risk in the AI era is multi-faceted. It requires total oversight of where data travels, from the ingestion phase for model training to real-time prompts and responses. On top of this, AI models require a massive amount of data to ensure the accuracy and context needed to deliver on efficiency goals.

This makes AI governance paramount. With controls built for modern data, Netskope can discover, classify, and label all structured and unstructured data for model training.

Netskope secures data across the full AI lifecycle. We provide deep visibility into shadow AI use and data flows, ensuring your organization only uses the data it needs, with comprehensive dashboard metrics that share visibility into usage and interaction.

- Netskope One CASB API inspects data at rest for data governance and threat protection.
- For private AI applications, Netskope One DLP On Demand enables organizations to build DLP directly within their private AI applications.
- Netskope One DSPM implements a robust data-centric security posture management framework to ensure proper data posture.
- Netskope One SSPM provides application-centric security posture management for data in SaaS environments.
- To ensure auditability, Netskope One Agentic Broker logs comprehensive MCP events, including initializations, tool requests, and responses, underpinning the capability needed for proper AI governance and retrospective investigations.

BENEFITS	DESCRIPTION
Secure every AI interaction	Gain total visibility and control across all AI traffic, including humans, internal applications, and agents, through a unified, high-performance access layer.
Smart defense for AI	Protect AI interactions with real-time guardrails that block AI-specific threats and enforce content moderation, ensuring safe and compliant use across the entire enterprise.
AI data governance at scale	Secure the entire data lifecycle through automated discovery, classification, and proactive pre-deployment hardening, to ensure your intellectual property remains protected and compliant.
Harden private models before deployment	Automate adversarial testing with 18,000+ scenarios to stress-test private LLMs for vulnerabilities before and after deployment to production environments, with CI/CD pipeline integration via APIs.
Secure agentic AI and MCP traffic	Decode MCP traffic to identify active agents and remote servers. Evaluate public MCP servers using risk scoring, enforce access controls, and prevent data loss through integrated DLP and AI guardrails.
Enforce responsible AI use	Automatically detect inappropriate content including violence, discrimination, weapons, and copyrighted material. Protect intellectual property rights by blocking patented or copyrighted data in AI responses.
Accelerate secure AI adoption	Move your organization from AI experimentation to AI advantage with a unified security platform delivering high-performance access for public AI traffic through the NewEdge Network.
Enhanced SecOps and compliance efficiency	Map AI policy violations including content moderation, DLP, and threats to MITRE ATLAS and OWASP Top 10 for LLMs to provide a unified view of incidents, reducing investigation time.