Solution Brief

Securing AI with Netskope One

AI adoption moves fast, and so do the risks. Netskope One AI Security delivers complete control over your AI ecosystem through comprehensive discovery and visibility, end-to-end pipeline protection for your data and models, and granular runtime enforcement over users, agents, and adversaries, enabling secure AI at speed.

Quick Glance

- Govern and control all user and agent traffic in one unified platform.
- Discover AI apps and MCP servers, with visibility of usage behavior, shadow AI, and personal AI app use.
- Protect sensitive data across the entire AI lifecycle, from model training to real-time interactions.
- Defend against AI-specific threats in real time, including prompt injection and inappropriate use.
- Protect the AI your developer builds before it ships and while it runs with automated adversarial LLM testing.

> "We rely on Netskope to secure operations as we future-proof the business through digitalization of professional services and the introduction of AI."

Stuart Walters, Partner and Chief Information Officer

BDO UK

The challenge

AI has moved from experimentation to the business front line. Every department, workflow, and customer interaction is becoming AI-enabled and the attack surface is expanding fast.

AI introduces sensitive data into systems not built with security at their core. Employees adopting tools in the cloud and on endpoints create shadow AI blind spots, with 94% of organizations reporting gaps in AI activity visibility[1]. Autonomous agents communicate across protocols that evade traditional network inspection. Meanwhile, adversaries are actively attempting to override model rules and exfiltrate data.

Organizations need a unified approach to discovering AI assets, securing the data fueling them, and governing user and agent behavior without sacrificing the speed AI makes possible.

Netskope for AI security

Netskope One AI Security secures users, agents, applications, and data across the entire AI ecosystem within a single unified platform: Netskope One. Whether your workforce is accessing generative AI 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.

netskope

[1] Netskope AI Risk and Readiness Report

Comprehensive AI discovery and usage visibility

Enterprise AI is beyond the reach of traditional security tools. The Netskope One AI Command Center ingests signals from across the Netskope One platform, continuously discovering and building an inventory of AI apps, embedded AI within SaaS apps, and the MCP servers enabling autonomous agents, with visibility of user and MCP activity to enforce real-time controls.

Netskope One AI Security provides a zero trust access layer to monitor and secure this traffic.

- **The Netskope One Next Gen Secure Web Gateway** (NG-SWG) manages user-to-AI app traffic with granular visibility and protection for SaaS AI tools, including shadow AI instances.
- **Netskope One Private Access (NPA)** controls appropriate user access to private AI applications.

Two purpose-built products secure agentic communications where traditional perimeters fail:

- **Netskope One AI Gateway** secures east-west traffic (app-to-LLM) as a virtual appliance for AI running in public cloud infrastructure, or in private cloud deployments, centralizing authentication and traffic management across LLM providers to ensure autonomous data flows remain governed.
- **Netskope One Agentic Broker** decodes and secures MCP traffic between AI agents and your data sources. It evaluates public MCP servers with dedicated MCP risk scoring, delivering supply chain risk management for AI at scale.

These signals encompass the full spectrum of AI risk, from ungoverned shadow AI and low-trust applications to high-risk user behavior and all associated DLP, and guardrail policy violations including AI threats, and content moderation.

> Complete AI asset visibility gives organizations the foundation they need to govern every interaction, and move from reactive to proactive AI security.

Real-time content moderation and threat protection

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

Netskope One AI Guardrails provides a dedicated runtime defense layer, purpose-built for AI environments, with multi-lingual support. It mitigates sophisticated attacks—including prompt injections and jailbreak attempts—by analyzing every request and response, identifying the intent behind linguistic exploits. Content moderation ensures AI usage stays within your organization's risk tolerance.

Every prompt and response passes through three Netskope engines simultaneously, with connected events correlated under a single incident ID:

- **Netskope One DLP:** identifying sensitive data in motion.
- **Netskope One Threat Protection:** detecting malware and malicious links.
- **Netskope One AI Guardrails:** blocking AI-specific threats and enforcing content moderation.

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

When a user takes a wrong turn, real-time coaching messages work alongside zero trust access controls to alert them to the risk and prevent the data from ever reaching the LLM. Where needed, Netskope goes further: blocking the app, re-authenticating the user, or isolating the browser session entirely.

Searchable conversation logs map directly to MITRE ATLAS and the OWASP Top 10 for LLMs, giving investigators full context—user, prompt, intent, and application—in a single unified view. With an average of 223 genAI data policy violations per month[2], correlated incident views ensure investigations are faster and findings immediately actionable.

[2] Netskope Cloud and Threat Report: 2026

Securing AI across every deployment frontier

Enterprise AI operates across three distinct frontiers, each with a different risk profile. Netskope One AI Security provides unified controls across all AI traffic: human users, applications, and autonomous agents.

For public SaaS AI, Netskope One NG-SWG delivers granular, instance-level zero trust controls and user coaching for more than 85,000 SaaS apps, including 1,800 genAI apps, distinguishing corporate from personal instances of tools such as ChatGPT and Microsoft Copilot. With 72% of genAI app usage identified as shadow AI, this instance-level precision makes the difference between a policy that exists and one that actually enforces secure and scalable AI adoption.

For agentic AI, Netskope One Agentic Broker sits in the path of all MCP-based communications, providing real-time policy enforcement to block unauthorized MCP communication, prevent sensitive data leaks through integration with Netskope One DLP, and enforce least-privilege access for AI agents. Every MCP interaction is logged, inspected, and governed, giving security teams the same level of oversight over agent behavior that they have for user behavior.

For privately built or self-hosted AI-powered apps, Netskope One AI Gateway provides visibility, access control, and runtime inspection for internal LLM and agent-to-agent traffic. Netskope One AI Red Teaming stress-tests privately hosted models before and after every production release, to detect model drift and simulate prompt injection, jailbreaking, and other LLM-specific attacks based on OWASP LLM Top 10 and MITRE ATLAS frameworks.

Unified AI data governance at scale

Data risk in the AI era spans the full lifecycle, from model training data ingestion to real-time prompts and responses. Netskope discovers, classifies, and labels all structured and unstructured data to ensure only the right data powers your AI systems. We provide deep visibility into shadow AI use and data flows, with metrics covering usage and interaction across your AI ecosystem. Netskope's governance tools include:

- **Netskope One CASB API** for inspecting data at rest in SaaS and AI apps.
- **Netskope Cloud Confidence Index (CCI)** to automate risk assessment of 85,000+ AI apps and MCP servers.
- **Netskope One DLP On Demand** to build DLP controls within private AI applications.
- **Netskope One DSPM** for discovering sensitive data stores used for RAG or training.
- **Netskope One SSPM** for security posture management of SaaS and genAI app misconfigurations.
- **Netskope One Agentic Broker** to log all MCP events.

With data moving at machine speed, your detection and response must stay ahead. Netskope One AgentSkope automates detection, investigation, and recommendations to accelerate your response. The Netskope One DLP AISecOps Agent deduplicates DLP incidents, identifies false positives, clusters cases, and automates workflows, enabling natural-language triage. This transforms AI data governance from reactive compliance to continuously improving operational posture, allowing analysts to focus on decisions.

BENEFITS	DESCRIPTION
Comprehensive AI discovery and risk visibility	Discover managed and shadow AI assets across SaaS apps, and MCP servers with visibility into user behavior and all associated DLP, threat, and guardrail policy violations.
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.
Transform AI data governance with agentic security operations	Ensure security operations can move at machine speed with automated agents continuously working in the background to detect, investigate, and recommend actions to accelerate your response to DLP incidents.
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.