Netskope One AI Guardrails

Data Sheet

Defence against AI threats and misuse

AI introduces entirely new threat vectors that traditional security tools cannot see. Attackers now use manipulative prompts to bypass safety controls, and users may inadvertently generate harmful, discriminatory, or copyrighted content that creates significant legal and reputational risk for the enterprise.

Why is Netskope the best choice?

Designed for the modern AI enterprise, Netskope One AI Guardrails provides a dedicated runtime defense layer for AI environments. It mitigates sophisticated attacks—including prompt injection and jailbreak attempts—through deep, real-time analysis of all traffic. While also serving as an automated moderator for both human and agentic interactions, it enforces continuous policy compliance and data integrity.

Runtime threat protection and content moderation

- **Stop threats and secure model integrity**
  Block adversarial attempts to override system rules or exfiltrate data. Inspect every request and response to identify the multi-stage intent behind sophisticated linguistic exploits.
- **Enforce responsible AI use and brand safety**
  Filter harmful or discriminatory content—including hate speech, violence, weapons, and crimes—automatically. This ensures AI usage stays within your organization's risk tolerance and protects your corporate reputation.
- **Mitigate legal and IP risk in generated material**
  Identify and block the delivery of patented or copyrighted data in AI responses. This proactively defends against emerging legal liabilities associated with generative model outputs.
- **Correlate detections with DLP and Advanced Threat Protection**
  AI Guardrails integrates seamlessly with Netskope One DLP and threat protection. Enabled through Netskope One platform's AI, SkopeAI, connected policy violation detections are unified in one cohesive view for greater context and faster investigation.

Key benefits and capabilities

Improved return on AI investment

Establish and enforce clear safety boundaries that allow you to deploy AI for high-stakes, high-value business use cases.

Enhanced SecOps and compliance efficiency

Map detections to MITRE ATLAS and OWASP Top 10 for LLMs. This unified view reduces investigation time and aligns teams with the latest TTPs.

Audit-ready traceability

Maintain searchable conversation logs matched to policy triggers with role-based access control to ensure only authorized investigators view prompt histories.

Deep user intent analysis

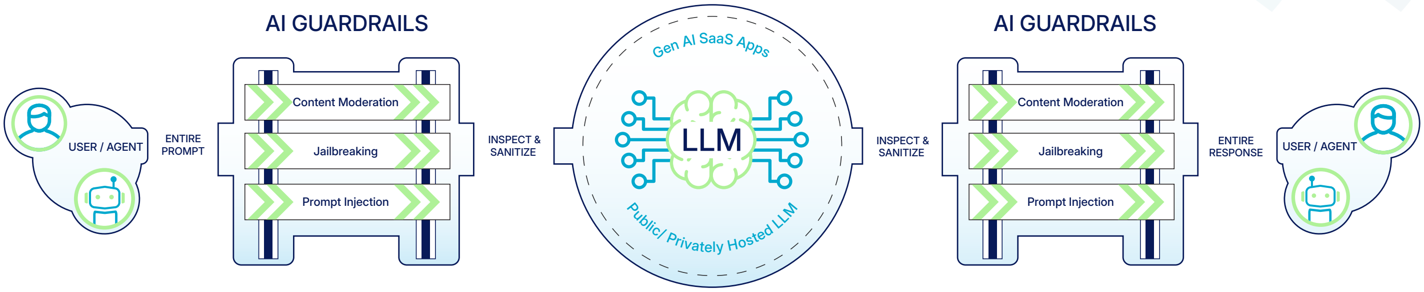
Use behavioral signals to distinguish between legitimate use and malicious activity, preventing data exposure before it actually occurs.

Uninterrupted user experience

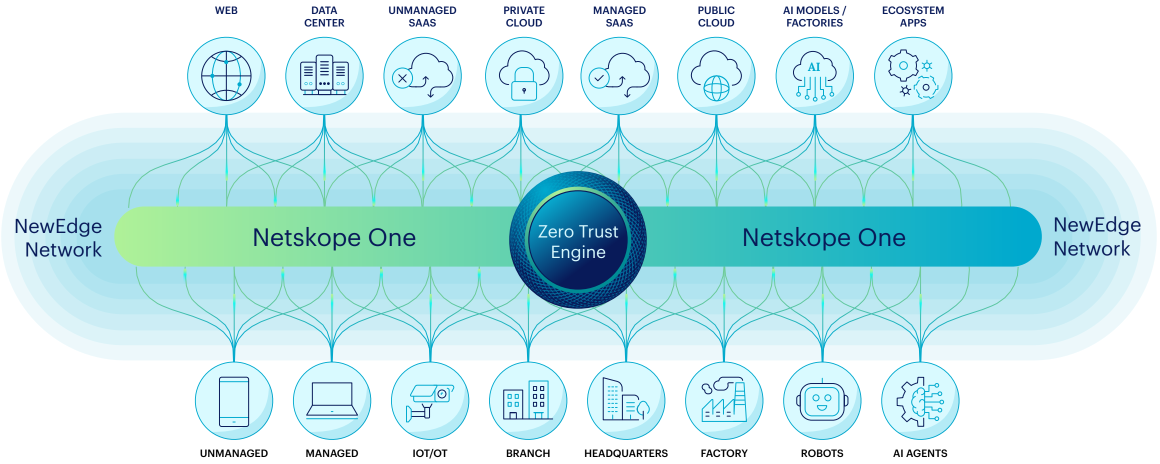
Maintain innovation speed with low latency guardrails designed to work at the scale of modern enterprise AI.

> "Driven by shadow AI and tool proliferation, 42% of GenAI policy violations involved source code, followed by regulated data (32%) and intellectual property (16%)."
>
> – Netskope Threat Labs, Cloud and Threat Report: Cloud and Threat Report 2026

Netskope One AI Guardrails



netskope



The Netskope difference

Netskope One AI Guardrails redefines data governance by moving beyond fragmented tools to a proactive, integrated defense system. The solution applies smart controls at the critical moment of dedication, uniquely combining specialized AI guardrails for content moderation and AI-specific threat protection with Netskope's industry-leading DLP and threat protection. This integration provides vital risk context, creating a single-incident view of all LLM threats to reduce alert fatigue and accelerate investigations.

This unified approach allows the platform to understand true user and agent intent by reviewing prompts and responses in context, distinguishing between normal work and malicious manipulation. Security teams benefit from searchable conversation logs of policy triggers that map directly to frameworks like MITRE ATLAS and the OWASP Top 10 for LLMs, ensuring they stay ahead of evolving adversary tactics. Whether managing public generative AI SaaS, enterprise plans, or private deployments on Amazon Bedrock, or autonomous agentic workflows, Netskope provides consistent content moderation and threat protection. By integrating these capabilities into the Netskope One platform, organizations can finally move from experimentation to full AI advantage securely.

BENEFITS	DESCRIPTION
Stop malicious AI attacks	Inspect prompts and responses in real-time to block sophisticated techniques like prompt injection and malicious jailbreaking attempts.
Enforce responsible AI usage	Our content moderation engine automatically detects inappropriate content categories such as violence, discrimination, weapons, sexual content, piracy and copyrighted material.
Multilingual support for the entire prompt and response	AI Guardrails has multilingual support for the entire prompt and response, including Chinese, English, French, Spanish, Portuguese, German, Italian, Russian, Japanese, Korean, Vietnamese, Thai, Arabic, and more.
Prevent sensitive data leaks	AI Guardrails integrates with Netskope One DLP to identify and block PII, source code, or proprietary secrets from entering AI models.
Unified view of incidents	Map AI policy violations including content moderation, DLP, and threats to MITRE ATLAS and OWASP Top 10 frameworks to provide a unified view of incidents.
Protect intellectual property rights	Netskope detects and blocks the retrieval or sharing of copyrighted and patented materials within AI-generated responses.
Audit and governance	Searchable logs with role-based access control ensure only authorized investigators can view sensitive chat histories.

netskope

Interested in learning more? Request a demo

Netskope, a leader in modern security and networking, addresses the needs of both security and networking teams by providing optimized access and real-time, context-based security for people, devices, and data anywhere they go. Thousands of customers, including more than 30 of the Fortune 100, trust the Netskope One platform, its Zero Trust Engine, and its powerful NewEdge network to reduce risk and gain full visibility and control over cloud, AI, SaaS, web, and private applications—providing security and accelerating performance without trade-offs. Learn more at netskope.com.

02/26 DS-968-1