

Netskope Skylight AI Red Teaming

Data Sheet



Automated testing for production-ready AI

Moving from SaaS to private large language models (LLMs) creates a critical security gap. Netskope Skylight AI Red Teaming closes this by automating adversarial simulations to uncover vulnerabilities, ensuring private models are secure, compliant, and resilient.

Why is Netskope the best choice?

Proactively prevent and remove vulnerabilities in private AI deployments with Netskope Skylight AI Red Teaming. By exposing LLMs to thousands of simulated prompts, we can investigate how drift creates vulnerabilities. Integrating Netskope Skylight AI Guardrails enables policy creation and prompt hardening, protecting the development lifecycle from build to runtime.

Protect AI development across the lifecycle

- **Automated multi-technique adversarial testing**
Stress-test private and foundation models (OpenAI, Azure OpenAI, Bedrock) using 18,000+ scenarios across techniques like DAN, Refusal Suppression, and Deceptive Delight. This automated approach keeps pace with rapid development, replacing slow, manual testing.
- **Continuous security integration throughout the AI development lifecycle**
Use APIs to integrate into CI/CD pipelines, automatically screening for vulnerabilities or risks introduced by code changes before every production release.
- **Simulate sophisticated multi-turn attacks**
Identify where complex skeleton key and crescendo attacks that trick LLMs into bypassing safety guardrails could impact your models pre- and post-production.
- **Track attack success rate and model drift**
Monitor attack success rate trends, OWASP LLM risk breakdowns, and performance drift across model iterations to ensure continuous alignment and compliance.

Key benefits and capabilities

Bridge the AI security gap

Confidently transition from experimentation to production-ready AI with automated simulations that ensure your private models are resilient and secure.

Maintain robust adversarial defense

Ensure model updates never introduce new security vulnerabilities or increased risk by maintaining consistent, high-strength adversarial defense layers.

Ensure strict privacy compliance

Protect your brand by identifying model vulnerabilities that could accidentally reveal internal system prompts, training records, or sensitive intellectual property.

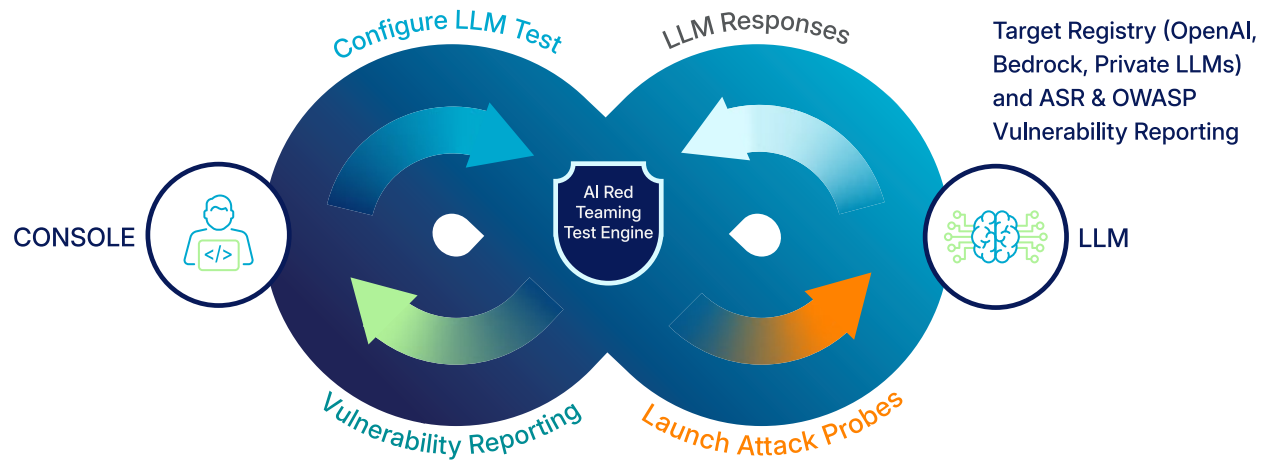
Accelerate secure AI innovation

Speed up development cycles by replacing manual security reviews with automated testing, allowing teams to deploy AI features faster without compromising safety.

Protect against evolving threats

Stay ahead of sophisticated attackers with continuous probing that identifies prompt injections, jailbreaks, and advanced logic-based conversational threats.

Netskope Skylight AI Red Teaming



The Netskope difference

Netskope One provides a unified inspection point for all user and agent traffic, delivering the real-time visibility, threat, and data protection required for modern AI security. While traditional tools struggle with the scale of new AI initiatives, Netskope Skylight AI Red Teaming automates the identification of vulnerabilities before and after they reach production. Our library of more than 18,000 adversarial scenarios and test cases stress-test models against prompt injection, jailbreaking, data leakages, and malicious use scenarios. Beyond simple filters, we simulate complex

multi-turn attack techniques including skeleton key and crescendo attacks that aim to bypass standard security guardrails. This approach ensures your AI journey is secure and scalable by design, fully integrated into a broader data security strategy. With Netskope, you can confidently move from experimentation to production, knowing your private models are resilient against the most sophisticated adversarial threats in the evolving AI landscape.

BENEFITS	DESCRIPTION
Automated CI/CD security integration	Integrate adversarial stress testing directly into your CI/CD pipelines via APIs. Automatically screen every code change or model update for new security risks and vulnerabilities before they reach production.
DLP profile alignment and leakage testing	Detect and evaluate models directly against your configured Netskope DLP profiles to prevent exfiltration of system prompts, training sets, and proprietary PII, ensuring compliance with strict privacy standards.
LLM security testing	Using more than 18,000 adversarial scenario test cases and seed prompts we enable you to systematically stress-test LLMs for vulnerabilities before and after they are deployed to production environments.
Multi-turn attack simulation	Simulates complex multi-turn attacks where adversaries attempt to trick LLMs or layer multi-stage conversations to bypass guardrails that lack full session context.
Model drift and quality continuity	Tracks alignment decay and vulnerability trends across scheduled test rounds to ensure model fine-tuning or software updates do not degrade safety posture.



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.](#)

©2026 Netskope, Inc. All rights reserved. Netskope, NewEdge, SkopeAI, and the stylized "N" logo are registered trademarks of Netskope, Inc. Netskope Active, Netskope Cloud XD, Netskope Discovery, Cloud Confidence Index, and SkopeSights are trademarks of Netskope, Inc. All other trademarks included are trademarks of their respective owners.

08/26 DS-966-2