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

Protect AI development across the lifecycle

- **Automated adversarial testing**
  Benefit from a library of more than 18,000 adversarial scenarios to systematically stress-test models. 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 changing risk assessments**
  Shift model testing from passive observation to active defense. By running scheduled red teaming simulations, we show the change in risks identified across all tests on the same model.

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.

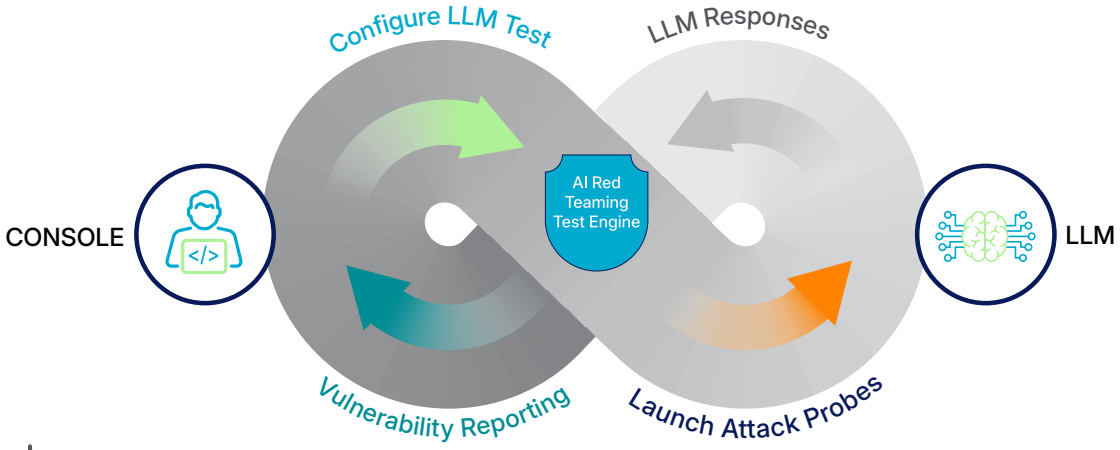
Accelerate secure AI innovation

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

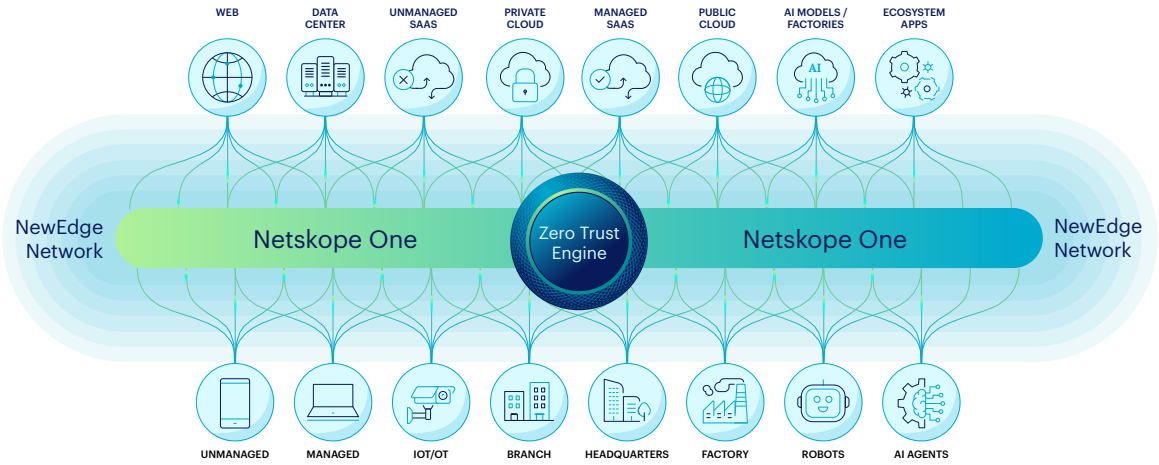
Protect against evolving threats

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

Netskope One AI Red Teaming



netskope



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 One AI Red Teaming automates the identification of vulnerabilities before and after they reach production. Our library of more tham 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 mutlti-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
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.
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.
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.
Data leakage prevention	Detect possible data leakage where model vulnerabilities could reveal internal system prompts or recall sensitive data including training records and internal knowledge, ensuring compliance with strict privacy standards.