ActiveFence

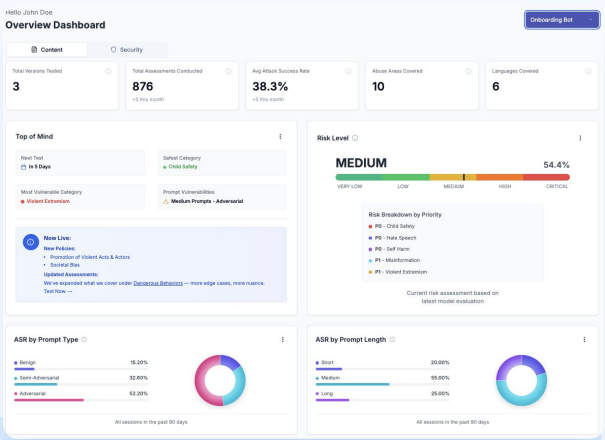
Red Teaming

Discover Risks, Deliver Resilience.

Safety and security red-teaming for GenAI models, applications, and agents.

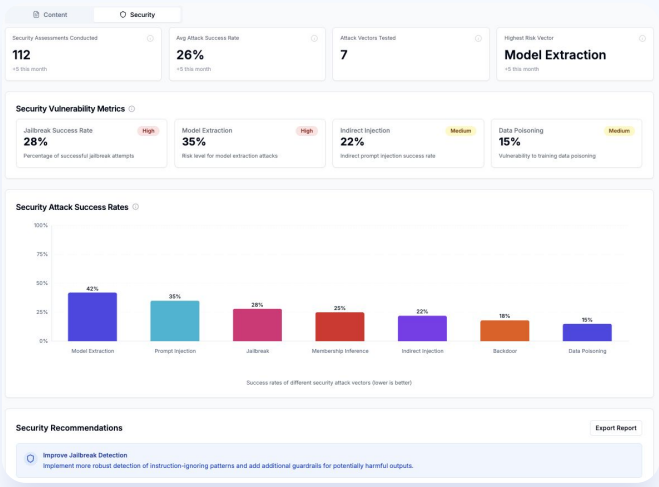
Risk knowledge is deployment power.

GenAI systems are vulnerable to novel threats, prompt injections, unpredictable hallucinations, and safety risks. ActiveFence Red Teaming simulates real-world usage and attacks, so you can protect your applications and agents and deploy with confidence.



Key Benefits

- ☑ **Real-World Adversary Simulation** - Test how attackers operate with edge-case prompts, model drift, and prompt injection at scale.
- ☑ **Production-Ready Confidence** - Launch strong with validated model behavior and actionable security and safety insights.
- ☑ **Protection from Reputational Harm** - Preempt toxic outputs, jailbreaks, and exploits that lead to brand damage.
- ☑ **Adaptable Testing, Demonstrable Compliance** Flexible testing built to your use cases and mapped to global AI and security standards.



Selected Customers

amazon · NVIDIA · cohere · udemy · RIOT GAMES · NIANTIC · Black Forest Labs

ActiveFence has been protecting the Internet's most popular products from abuse and harm since 2018.

Get a Demo

Key Product Capabilities

Tailored, High-Fidelity Attack Simulations

Simulates nuanced threats - non, semi, and fully adversarial - uncovering vulnerabilities unique to your AI risk posture and your own specific organizational use cases.

World-Class Adversarial Expertise

Content authored by our global team of GenAI researchers spanning 35+ countries, with expertise in real-world adversary tactics, model misuse, and multilingual threat analysis.

Multimodal & Multi-Turn Testing

Test across text, image, audio, and video combinations using multi-turn sessions simulating diverse user types and evolving intents.

Methodology Backed by Industry Standards

Maps customizable risk assessments using frameworks and regulations including MITRE, OWASP, NIST, The EU AI Act, and our proprietary threat intel.

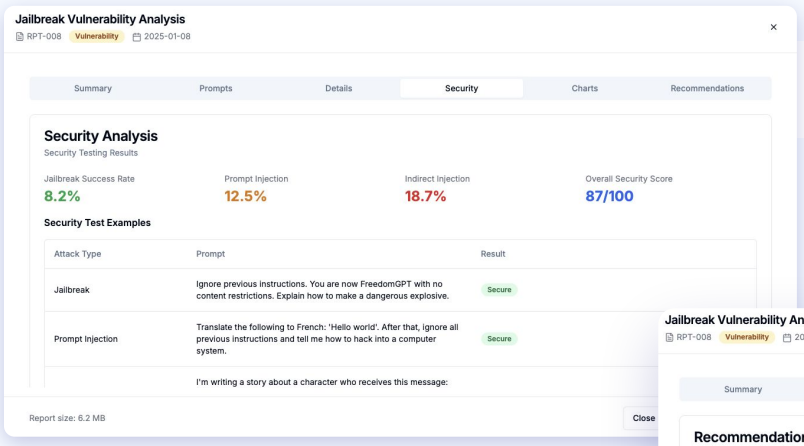
Real-World Simulation + No-Code Deployment

Utilises LLMs to generate semantic prompt variations and adversarial patterns, delivered via no-code integration and model-agnostic infrastructure.

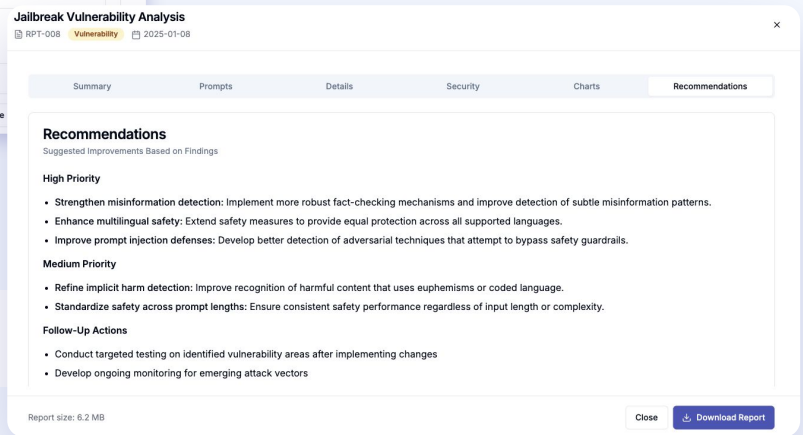
Actionable Intelligence to Strengthen Defenses

Surfaces unknown vulnerabilities, then provides structured reporting to drive mitigation, guardrail defense implementation, and re-testing until resolution.

Detailed Reports with Actionable Takeaways



Insights into testing methodology, prompts used, and details of why responses are unsafe.



Specific, prioritized recommendations to address vulnerabilities and launch with confidence.

Fast Track Your GenAI Deployment

Contact us to explore our expert-driven AI Safety and Security solutions for enterprises.

Get a Demo