SOLUTION BRIEF

HackerOne
AI Red Teaming

Runtime Testing for AI Trust, Safety, and Security

Unchecked AI can fail in unpredictable, harmful ways, policy violations, and jailbreaks that slip through automation and in-house QA. HackerOne AI Red Teaming (AIRT) exposes these blind spots before they become crises.

Our AIRT delivers scoped, adversarial testing for AI models, probing safety, security, and policy alignment through human creativity. Each engagement simulates real-world abuse conditions to uncover hidden risks and validate defenses. Trusted by frontier model developers and regulated enterprises alike, HackerOne's approach combines human-in-the-loop expertise, technical guidance, and orchestration, in addition to deliverables that help customers ship safe, responsible AI.

Key Outcomes

Uncover High-Impact AI Vulnerabilities

Reveal universal jailbreaks, training-data backdoors, and adversarial evasion that static analysis, fine-tuning, or automated assessments often miss, giving security and ML teams an early, actionable signal.

Support TRiSM and NIST Alignment

Map findings to OWASP LLM Top 10, Gartner TRiSM risk domains, and NIST RMF functions to provide clear evidence for legal, compliance, and governance teams.

Reduce Business and Regulatory Exposure

Assessing AI systems in real-world abuse scenarios before production can help executive and risk stakeholders avoid reputational damage, legal penalties, financial loss, and unsafe launches.

> *"Our challenge [with HackerOne] generated significant engagement from the AI security community, with 339 jailbreakers attempting to jailbreak our system across 300,000+ chat interactions, representing approx. 3,700 collective hours of human red teaming effort."*
>
> Anthropic Safeguards Research Team

Read the Case Study

Key Product Capabilities

Adversarial Model Testing

Identify jailbreaks, hallucinations, misalignment, or security gaps by challenging your models under real-world adversarial conditions, assessed by humans.

Time-Boxed, Objective-Based Engagements

Run focused 15- or 30-day engagements with defined attack types and test criteria mapped to your risk model

Trusted Policy & Model Partner

Partner with the only crowdsourced security platform working with foundational models: Anthropic, IBM Granite; which contributes to global AI policy; U.S. AI Action Plan, UK Cyber Code of Practice, Stanford-MIT AI Risk Workshop.

AI Researcher Community

Tap into the world's largest, most active pool of AI-focused security researcher community, ranked by reputation and results.

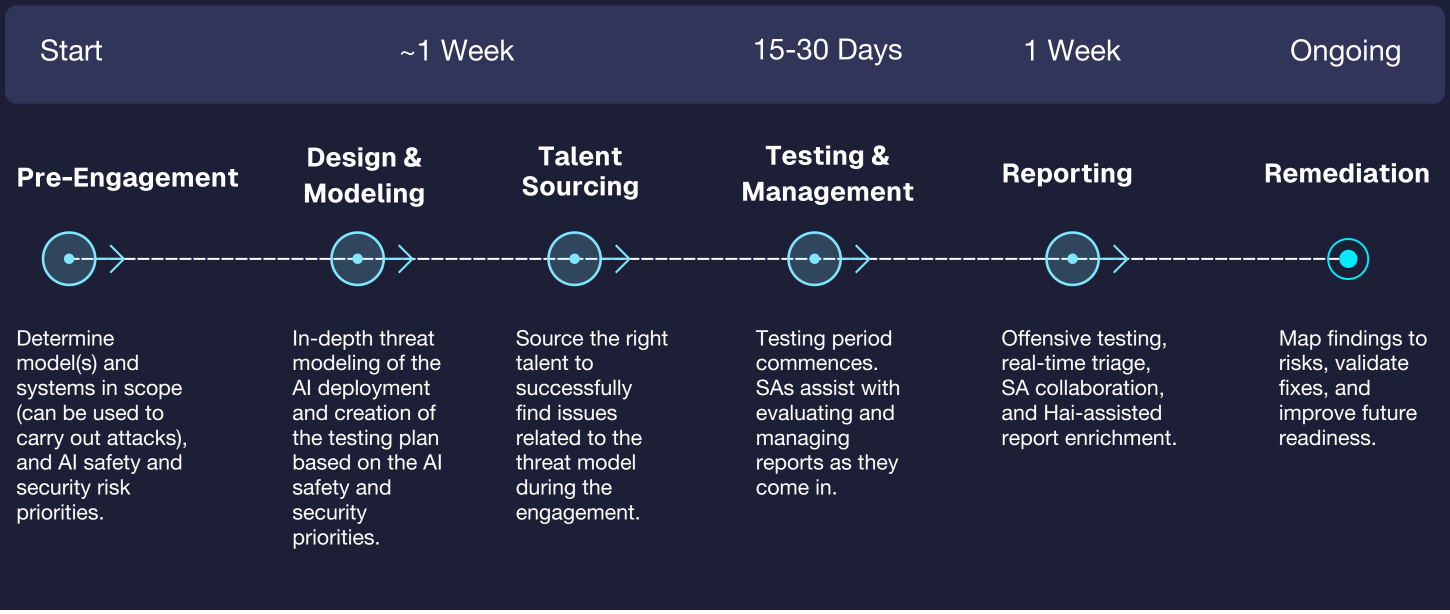
Comprehensive Coverage

Extend AIRT with HackerOne's full portfolio, including Pentest and Bounty, for end-to-end protection across your AI systems.

Security Advisory (SA) Support

SAs play a critical role across the AIRT lifecycle, contributing to threat modeling, policy design, flag criteria definition, mitigation planning, and community coordination before and after the test launch.

How AI Red Teaming Works



Start | ~1 Week | 15-30 Days | 1 Week | Ongoing

Pre-Engagement

Determine model(s) and systems in scope (can be used to carry out attacks), and AI safety and security risk priorities.

Design & Modeling

In-depth threat modeling of the AI deployment and creation of the testing plan based on the AI safety and security priorities.

Talent Sourcing

Source the right talent to successfully find issues related to the threat model during the engagement.

Testing & Management

Testing period commences. SAs assist with evaluating and managing reports as they come in.

Reporting

Offensive testing, real-time triage, SA collaboration, and Hai-assisted report enrichment.

Remediation

Map findings to risks, validate fixes, and improve future readiness.

> *"As time goes on, these areas will become less novel, and we will be able to rely more on automation and existing datasets for testing. But human ingenuity is crucial for understanding potential problems in novel areas."*

Ilana Arbisser,
Technical Lead, AI Safety at Snap Inc.

Read the Case Study

Understanding AI Risk: Safety vs Security

AI risks are often grouped under a single umbrella, but in reality, safety and security represent distinct risk categories with different mitigation paths:

AI Safety Risks

Risks that originate from the model itself, such as harmful, biased, or non-compliant outputs. These often stem from flaws in training data or insufficient enforcement of behavior policies.

- Demographic bias
- Toxic or illegal content generation
- Reputational and legal risks

AI Security Risks

Risks that stem from external manipulation of the AI system. These are adversarial in nature and may target models, APIs, or the application layer to bypass controls or exfiltrate data.

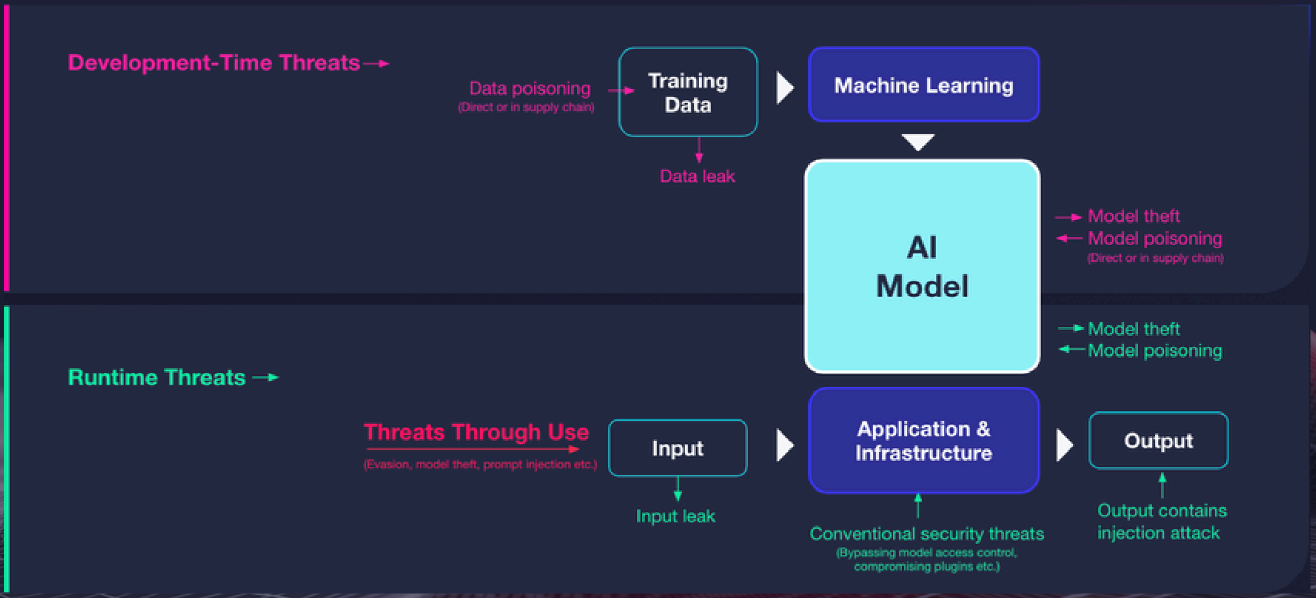
- Prompt injection
- Model theft and inversion
- Data poisoning and insecure output handling

Where AI Risks Emerge and How HackerOne Helps

AI risks span development and runtime. HackerOne helps identify and mitigate issues across the stack:

Development-Time Risks:
Poisoned training data, malicious model updates, and data leaks can compromise models before deployment. HackerOne surfaces these through secure code review and bug bounty programs targeting model pipelines and infrastructure.

Runtime Risks:
Prompt injection, model theft, and insecure output handling often surface once models are live. AIRT simulates real-world adversaries to uncover these vulnerabilities before attackers do.


Development-Time Threats →
Data poisoning
(Direct or in supply chain)
Training Data
Data leak
Machine Learning
AI Model
Model theft
Model poisoning
(Direct or in supply chain)
Runtime Threats →
Model theft
Model poisoning
Threats Through Use
(Evasion, model theft, prompt injection etc.)
Input
Input leak
Application & Infrastructure
Conventional security threats
(Bypassing model access control, compromising plugins etc.)
Output
Output contains injection attack

Visit the product page for more information on our approach or contact us now to learn how to get started.