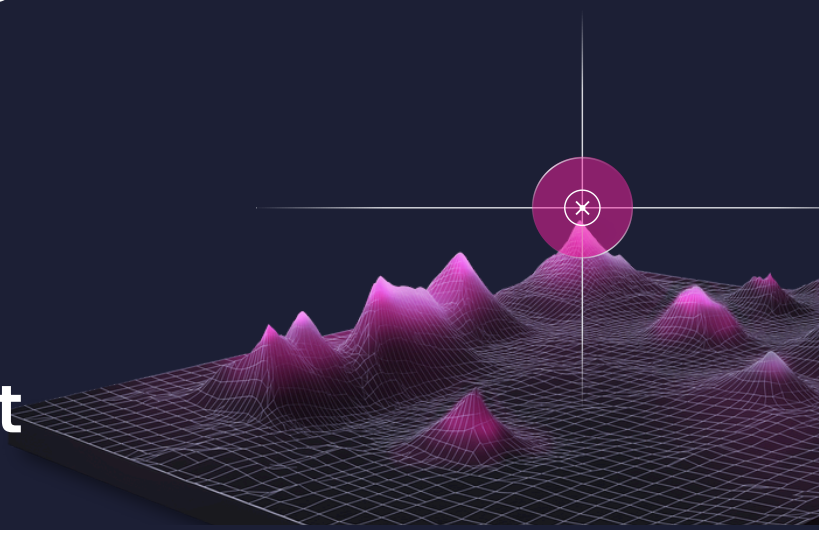
hackerone | ANTHROP\C

CUSTOMER STORY

How Anthropic's Jailbreak Challenge Put AI Safety Defenses to the Test



Proactively testing for risk is a key component of building responsible AI. One way organizations do this is through AI red teaming, which stress tests models to identify potential opportunities for abuse. AI red teaming often taps the broader security and AI researcher community to help find elusive security and safety issues caused by circumventing model guardrails. Model developers can then use these insights to improve or validate existing guardrails.

Last month, Anthropic partnered with HackerOne to complete an AI red teaming challenge on a demo version of Claude 3.5 Sonnet. The challenge's goal was to test and validate Anthropic's new Constitutional Classifiers, which block harmful queries, particularly those that could produce outputs related to CBRN (chemical, biological, radioactive, nuclear) weapons and related content. Anthropic invited researchers to try and bypass Claude's defenses through a "universal" jailbreak — a technique that allows model users to bypass safety defenses with a single input consistently.

The challenge ran from February 3 to February 10 and consisted of eight levels. To pass each level, researchers had to gain answers from Claude about a question related to CBRN topics through jailbreaking. Depending on their findings, researchers earned bounties: $10,000 to the first participant who passed all eight levels with different jailbreaks and $20,000 to the first participant who used a single, universal jailbreak to pass all levels.

Challenge Results

The challenge saw substantial engagement, with more than 300,000 chat interactions from 339 participants. We'd like to thank all the researchers who participated and congratulate those who received bounty rewards. It was no small feat! Four teams earned a total of $55,000 in bounty rewards from Anthropic: one passed all levels using a universal jailbreak, one passed all levels using a borderline-universal jailbreak, and two passed all eight levels using multiple individual jailbreaks.

> **This challenge demonstrated the high return on investment for collaborative efforts. Delivering large language models (LLMs) in a safe and aligned manner is a significant challenge—especially given the intricacies of transformer architectures. This experience was a clear reminder that as these models get smarter, our strategies for testing can also evolve to stay ahead of potential risks.**
>
> Salia Asanova aka @saltyn

How The Community Contributes to Safer Systems

The diversity of techniques used by the winners and all the researchers who participated contributed to strengthening Claude's protections. Anthropic noticed a few particularly successful jailbreaking strategies researchers employed:

Using encoded prompts and ciphers to circumvent the AI output classifier

Leveraging role-play scenarios to manipulate system responses

Substituting harmful keywords with benign alternatives

Implementing advanced prompt-injection attacks

These discoveries made by the community identified fringe cases and key areas for Anthropic to reexamine for its safety defenses while validating where guardrails remained effective.

Looking Ahead: Strengthening AI Defenses

The findings demonstrate the value the community can deliver when organizations use AI red teaming work in addition to other AI safety and security best practices:

> **Our researcher community's approach is rooted in curiosity, creativity, and the relentless pursuit of finding flaws others might miss. This mindset is distinct from building and reinforcing technical models, yet it's an essential complement. While internal teams focus on defending and aligning AI systems, engaging with a community of researchers ensures continuous, real-world testing that validates and strengthens those defenses. Together, these perspectives drive more resilient and trustworthy AI.**

Dane Sherrets
Staff Solutions Architect, Emerging Technologies at HackerOne

As AI advances, so must the ways we secure it. We're committed to collaborating with leaders like Anthropic, who continue to define AI safety best practices that help us all build a more resilient digital world.

Visit here to read more about the challenge and Anthropic's AI safety work.