Strength in Numbers: NVIDIA and Generative Red Team Challenge Unleash Thousands to Vet Security at DEF CON

Getting hands-on with the latest technologies is a proud hacker tradition; this week in Las Vegas, thousands will dig into the latest AI safety tools.

Author: Daniel Rohrer

Thousands of hackers will tweak, twist and probe the latest generative AI platforms this week in Las Vegas as part of an effort to build more trustworthy and inclusive AI.

Collaborating with the hacker community to establish best practices for testing next-generation AI, NVIDIA is participating in a first-of-its-kind test of industry-leading LLM solutions, including NVIDIA NeMo and NeMo Guardrails .

The Generative Red Team Challenge, hosted by AI Village , SeedAI , and Humane Intelligence , will be among a series of workshops, training sessions and appearances by NVIDIA leaders at the Black Hat and DEF CON security conferences in Las Vegas.

The challenge — which gives hackers a number of vulnerabilities to exploit — promises to be the first of many opportunities to reality-check emerging AI technologies.

“AI empowers individuals to create and build previously impossible things,” said Austin Carson, founder of SeedAI and co-organizer of the Generative Red Team Challenge. “But without a large, diverse community to test and evaluate the technology, AI will just mirror its creators, leaving big portions of society behind.”

The collaboration with the hacker community comes amid a concerted push for AI safety making headlines across the world, with the Biden-Harris administration securing voluntary commitment from the leading AI companies working on cutting-edge generative models.

“AI Village draws the community concerned about the implications of AI systems – both malicious use and impact on society,” said Sven Cattell founder of AI Village and co-organizer of the Generative Red Team Challenge. “At DEFCON 29, we hosted the first Algorithmic Bias Bounty with Rumman Chowdhury’s former team at Twitter. This marked the first time a company had allowed public access to their model for scrutiny.”

This week’s challenge is a key step in the evolution of AI, thanks to the leading role played by the hacker community — with its ethos of skepticism, independence and transparency — in creating and field testing emerging security standards.

NVIDIA’s technologies are fundamental to AI, and NVIDIA was there at the beginning of the generative AI revolution. In 2016, NVIDIA founder and CEO Jensen Huang hand-delivered to OpenAI the first NVIDIA DGX AI supercomputer — the engine behind the large language model breakthrough powering ChatGPT.

NVIDIA DGX systems, originally used as an AI research instrument, are now running 24/7 at businesses across the world to refine data and process AI.

Management consultancy McKinsey estimates generative AI could add the equivalent of $2.6 trillion to $4.4 trillion annually to the global economy across 63 use cases.

This makes safety — and trust — an industry-wide concern.

That’s why NVIDIA employees are engaging with attendees at both last week’s Black Hat conference for security professionals and this week’s DEF CON gathering.

At Black Hat, NVIDIA hosted a two-day training session on using machine learning and a briefing on the risks of poisoning web-scale training datasets. It also participated in a panel discussion on the potential benefits of AI for security.

At DEF CON, NVIDIA is sponsoring a talk on the risks of breaking into baseboard management controllers. These specialized service processors monitor the physical state of a computer, network server or other hardware devices.

And through the Generative Red Team Challenge , part of the AI Village Prompt Detective workshop, thousands of DEF CON participants will be able to demonstrate prompt injection, attempt to elicit unethical behaviors and test other techniques to obtain inappropriate responses.

Models built by Anthropic, Cohere, Google, Hugging Face, Meta, NVIDIA, OpenAI and Stability, with participation from Microsoft, will be tested on an evaluation platform developed by Scale AI.

As a result, everyone gets smarter.

“We’re fostering the exchange of ideas and information while simultaneously addressing risks and opportunities,” said Rumman Chowdhury, a member of AI Village’s leadership team and co-founder of Humane Intelligence, the nonprofit designing the challenges. “The hacker community is exposed to different ideas, and community partners gain new skills that position them for the future.”

Released in April as open-source software, NeMo Guardrails can help developers guide generative AI applications to create impressive text responses that can stay on track — ensuring intelligent, LLM-powered applications are accurate, appropriate, on topic and secure.

To ensure transparency and the ability to put the technology to work across many environments, NeMo Guardrails — the product of several years of research — is open source, with much of the NeMo conversational AI framework already available as open-source code on GitHub , contributing to the developer community’s tremendous energy and work on AI safety.

Engaging with the DEF CON community builds on this, enabling NVIDIA to share what it has learned with NeMo Guardrails and to, in turn, learn from the community.

Organizers of the event — which include SeedAI, Humane Intelligence and AI Village — plan to analyze the data and publish their findings, including processes and learnings, to help other organizations conduct similar exercises.

Last week, organizers also issued a call for research proposals and received several proposals from leading researchers within the first 24 hours.

“Since this is the first instance of a live hacking event of a generative AI system at scale, we will be learning together,” Chowdhury said. “The ability to replicate this exercise and put AI testing into the hands of thousands is key to its success.”

The Generative Red Team Challenge will take place in the AI Village at DEF CON 31 from Aug. 10-13, at Caesar’s Forum in Las Vegas.

Original URL: https://blogs.nvidia.com/blog/2023/08/10/nvidia-generative-red-team-challenge/