News AI and machine learning

7 min read

# Announcing Microsoft’s open automation framework to red team generative AI Systems

By Ram Shankar Siva Kumar, Microsoft AI Red Team Lead

February 22, 2024

Security management

Security operations

Threat trends

Today we are releasing an open automation framework, PyRIT (Python Risk Identification Toolkit for generative AI), to empower security professionals and machine learning engineers to proactively find risks in their generative AI systems.

At Microsoft, we believe that security practices and generative AI responsibilities need to be a collaborative effort. We are deeply committed to developing tools and resources that enable every organization across the globe to innovate responsibly with the latest artificial intelligence advances. This tool, and the previous investments we have made in red teaming AI since 2019, represents our ongoing commitment to democratize securing AI for our customers, partners, and peers.

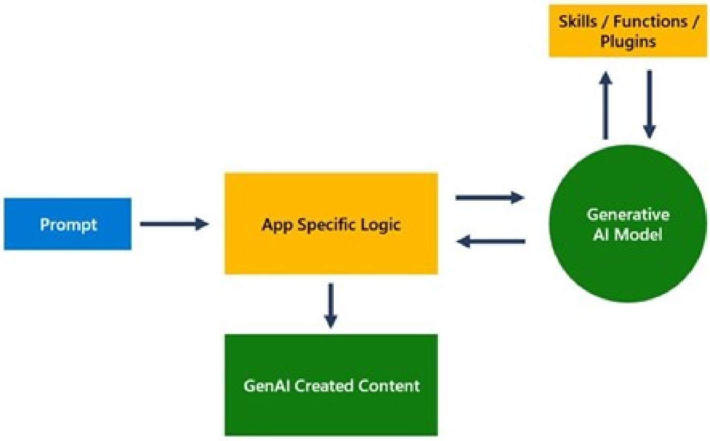
## The need for automation in AI Red Teaming

Red teaming AI systems is a complex, multistep process. Microsoft’s AI Red Team leverages a dedicated interdisciplinary group of security, adversarial machine learning, and responsible AI experts. The Red Team also leverages resources from the entire Microsoft ecosystem, including the Fairness center in Microsoft Research; AETHER, Microsoft’s cross-company initiative on AI Ethics and Effects in Engineering and Research; and the Office of Responsible AI. Our red teaming is part of our larger strategy to map AI risks, measure the identified risks, and then build scoped mitigations to minimize them.

Over the past year, we have proactively red teamed several high-value generative AI systems and models before they were released to customers. Through this journey, we found that red teaming generative AI systems is markedly different from red teaming classical AI systems or traditional software in three prominent ways.

### 1. Probing both security and responsible AI risks simultaneously

We first learned that while red teaming traditional software or classical AI systems mainly focuses on identifying security failures, red teaming generative AI systems includes identifying both security risk as well as responsible AI risks. Responsible AI risks, like security risks, can vary widely, ranging from generating content that includes fairness issues to producing ungrounded or inaccurate content. AI red teaming needs to explore the potential risk space of security and responsible AI failures simultaneously.



## 2. Generative AI is more probabilistic than traditional red teaming

Secondly, we found that red teaming generative AI systems is more probabilistic than traditional red teaming. Put differently, executing the same attack path multiple times on traditional software systems would likely yield similar results. However, generative AI systems have multiple layers of non-determinism; in other words, the same input can provide different outputs. This could be because of the app-specific logic; the generative AI model itself; the orchestrator that controls the output of the system can engage different extensibility or plugins; and even the input (which tends to be language), with small variations can provide different outputs. Unlike traditional software systems with well-defined APIs and parameters that can be examined using tools during red teaming, we learned that generative AI systems require a strategy that considers the probabilistic nature of their underlying elements.

## 3. Generative AI systems architecture varies widely

Finally, the architecture of these generative AI systems varies widely: from standalone applications to integrations in existing applications to the input and output modalities, such as text, audio, images, and videos.

These three differences make a triple threat for manual red team probing. To surface just one type of risk (say, generating violent content) in one modality of the application (say, a chat interface on browser), red teams need to try different strategies multiple times to gather evidence of potential failures. Doing this manually for all types of harms, across all modalities across different strategies, can be exceedingly tedious and slow.

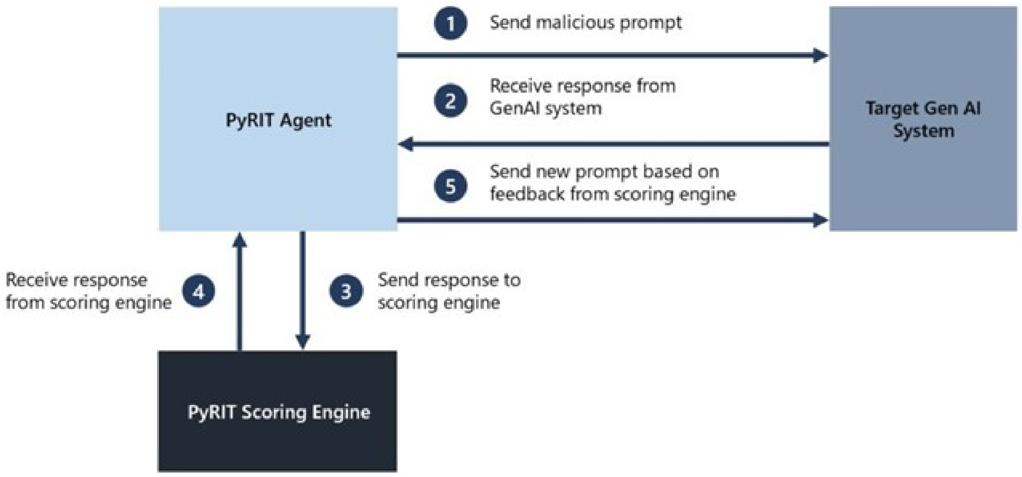
This does not mean automation is always the solution. Manual probing, though time-consuming, is often needed for identifying potential blind spots. Automation is needed for scaling but is not a replacement for manual probing. We use automation in two ways to help the AI red team: automating our routine tasks and identifying potentially risky areas that require more attention.

In 2021, Microsoft developed and released a red team automation framework for classical machine learning systems. Although Counterfit still delivers value for traditional machine learning systems, we found that for generative AI applications, Counterfit did not meet our needs, as the underlying principles and the threat surface had changed. Because of this, we re-imagined how to help security professionals to red team AI systems in the generative AI paradigm and our new toolkit was born.

We like to acknowledge out that there have been work in the academic space to automate red teaming such as PAIR and open source projects including garak.

# PyRIT for generative AI Red teaming

PyRIT is battle-tested by the Microsoft AI Red Team. It started off as a set of one-off scripts as we began red teaming generative AI systems in 2022. As we red teamed different varieties of generative AI systems and probed for different risks, we added features that we found useful. Today, PyRIT is a reliable tool in the Microsoft AI Red Team's arsenal.


PyRIT Agent
1 Send malicious prompt
2 Receive response from GenAI system
5 Send new prompt based on feedback from scoring engine
Target Gen AI System
4 Receive response from scoring engine
3 Send response to scoring engine
PyRIT Scoring Engine

The biggest advantage we have found so far using PyRIT is our efficiency gain. For instance, in one of our red teaming exercises on a Copilot system, we were able to pick a harm category, generate several thousand malicious prompts, and use PyRIT's scoring engine to evaluate the output from the Copilot system all in the matter of hours instead of weeks.

PyRIT is *not* a replacement for manual red teaming of generative AI systems. Instead, it augments an AI red teamer's existing domain expertise and automates the tedious tasks for them. PyRIT shines light on the hot spots of where the risk could be, which the security professional than can incisively explore. The security professional is always in control of the strategy and execution of the AI red team operation, and PyRIT provides the automation code to take the initial dataset of harmful prompts provided by the security professional, then uses the LLM endpoint to generate more harmful prompts.

However, PyRIT is more than a prompt generation tool; it changes its tactics based on the response from the generative AI system and generates the next input to the generative AI system. This automation continues until the security professional's intended goal is achieved.

## PyRIT components

Abstraction and Extensibility is built into PyRIT. That's because we always want to be able to extend and adapt PyRIT's capabilities to new capabilities that generative AI models engender. We achieve this by five interfaces: target, datasets, scoring engine, the ability to support multiple attack strategies and providing the system with memory.

- **Targets**: PyRIT supports a variety of generative AI target formulations—be it as a web service or embedded in application. PyRIT out of the box supports text-based input and can be extended for other modalities as well. PyRIT supports integrating with models from Microsoft Azure OpenAI Service, Hugging Face, and Azure Machine Learning Managed Online Endpoint, effectively acting as an adaptable bot for AI red team exercises on designated targets, supporting both single and multi-turn interactions.
- **Datasets**: This is where the security professional encodes what they want the system to be probed for. It could either be a static set of malicious prompts or a dynamic prompt template. Prompt templates allow the security professionals to automatically encode multiple harm categories—security and responsible AI failures—and leverage automation to pursue harm exploration in all categories simultaneously. To get users started, our initial release includes prompts that contain well-known, publicly available jailbreaks from popular sources.
- **Extensible scoring engine**: The scoring engine behind PyRIT offers two options for scoring the outputs from the target AI system: using a classical machine learning classifier or using an LLM endpoint and leveraging it for self-evaluation. Users can also use Azure AI Content filters as an API directly.
- **Extensible attack strategy**: PyRIT supports two styles of attack strategy. The first is single-turn; in other words, PyRIT sends a combination of jailbreak and harmful prompts to the AI system and scores the response. It also supports multiturn strategy, in which the system sends a combination of jailbreak and harmful prompts to the AI system, scores the response, and then responds to the AI system based on the score. While single-turn attack strategies are faster in computation time,

multiturn red teaming allows for more realistic adversarial behavior and more advanced attack strategies.

- **Memory**: PyRIT's tool enables the saving of intermediate input and output interactions providing users with the capability for in-depth analysis later on. The memory feature facilitates the ability to share the conversations explored by the PyRIT agent and increases the range explored by the agents to facilitate longer turn conversations.

## Get started with PyRIT

PyRIT was created in response to our belief that the sharing of AI red teaming resources across the industry raises all boats. We encourage our peers across the industry to spend time with the toolkit and see how it can be adopted for red teaming your own generative AI application.

1. Get started with the PyRIT project here. To get acquainted with the toolkit, our initial release has a list of demos including common scenarios notebooks, including how to use PyRIT to automatically jailbreak using Lakera's popular Gandalf game.
2. We are hosting a webinar on PyRIT to demonstrate how to use it in red teaming generative AI systems. If you would like to see PyRIT in action, please register for our webinar in partnership with the Cloud Security Alliance.
3. Learn more about what Microsoft's AI Red Team is doing and explore more resources on how you can better prepare your organization for securing AI.
4. Watch Microsoft Secure online to explore more product innovations to help you take advantage of AI safely, responsibly, and securely.

## Contributors

Project created by Gary Lopez; Engineering: Richard Lundeen, Roman Lutz, Raja Sekhar Rao Dheekonda, Dr. Amanda Minnich; Broader involvement from Shiven Chawla, Pete Bryan, Peter Greko, Tori Westerhoff, Martin Pouliot, Bolor-Erdene Jagdagdorj, Chang Kawaguchi, Charlotte Siska, Nina Chikanov, Steph Ballard, Andrew Berkley, Forough Poursabzi, Xavier Fernandes, Dean Carignan, Kyle Jackson, Federico Zarfati, Jiayuan Huang, Chad Atalla, Dan Vann, Emily Sheng, Blake Bullwinkel, Christiano Bianchet, Keegan Hines, eric douglas, Yonatan Zunger, Christian Seifert, Ram Shankar Siva Kumar. Grateful for comments from Jonathan Spring.

## Learn more

To learn more about Microsoft Security solutions, visit our website. Bookmark the Security blog to keep up with our expert coverage on security matters. Also, follow us on LinkedIn (Microsoft Security) and X (@MSFTSecurity) for the latest news and updates on cybersecurity.

Surface Laptop Go 3

Surface Pro 9

Surface Laptop 5

Microsoft Copilot

Copilot in Windows

Explore Microsoft products

Windows 11 apps

## Microsoft Store

Account profile

Download Center

Microsoft Store support

Returns

Order tracking

Certified Refurbished

Microsoft Store Promise

Flexible Payments

## Education

Microsoft in education

Devices for education

Microsoft Teams for Education

Microsoft 365 Education

How to buy for your school

Educator training and development

Deals for students and parents

Azure for students

## Business

Microsoft Cloud

Microsoft Security

Dynamics 365

Microsoft 365

Microsoft Power Platform

Microsoft Teams

Copilot for Microsoft 365

Small Business

## Developer & IT

Azure

Developer Center

Documentation

Microsoft Learn

Microsoft Tech Community

Azure Marketplace

AppSource

Visual Studio

## Company

Careers

About Microsoft

Company news

Privacy at Microsoft

Investors

Diversity and inclusion

Accessibility

Sustainability

English (United States)

Your Privacy Choices

Consumer Health Privacy

Sitemap Contact Microsoft Privacy Terms of use Trademarks Safety & eco Recycling About our ads