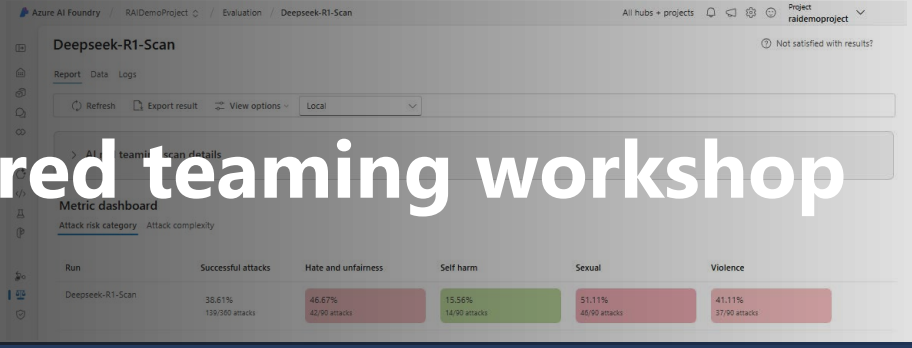

AI red teaming workshop

Workshop and Lab Program

Duration: 90 minutes Session [On-Site or Remote]
Difficulty Level: **300** - Moderate

Description

The workshop introduces participants to Microsoft's AI Red Teaming Agent, a tool that automates adversarial testing of generative AI systems using PyRIT strategies and Azure AI Foundry's Risk & Safety Evaluators to uncover vulnerabilities like jailbreaks, prompt injections, and harmful content.

Through hands-on labs and guided evaluations, attendees learn to simulate attacks, assess model resilience via Attack Success Rate (ASR) metrics, and integrate responsible AI practices into their development lifecycle

Objectives

- Learn the evolving threat landscape of attacks on applications using Generative AI LLM Models
- The practice of Azure AI red teaming to help reduce content safety risks.
- Understand the integration of Azure AI Foundry and Defender CSPM and Defender for AI services.

Outcomes

- Hands on practice using and interpreting Azure AI Foundry red teaming tools and results.
- A greater understanding of defense in depth using Defender CSPM and Defender for AI Services for investigations

Methodology

Knowledge Transfer

- The Microsoft Engineer will help you understand "How Azure AI Foundry red teaming agent works".
- This delivery is designed to help you understand "AI content safety risks and jail break attacks" against AI and ways to improve your security posture and defenses.
- Introduce on Topics AI Security Posture (Start Secure) Management and AI Workload Protection (Stay Secure)

Hands on Experience

- The Microsoft Engineer will help you understand how to setup and run the AI red reaming agent.
- Interpret AI red teaming results and context
- Setup some best practices to get the red teaming into operation and integrate into your security workflow.

Delivery

• (90 Minutes) Session: AI red teaming discussion and labs

Microsoft