# Hackfort Boise 2025

Continuous Chaos: How to Automate AI Red Teaming & Monitoring

**Thursday, Mar 27 1:00 pm - 1:50 PM**
**Boise Centre East** : Room 410 B

- Attendees will learn:
- The 6 layers of AI defense
- The nuances of testing nondeterministic systems,
- The difference between prompt injection attacks and jail breaking attacks,
- How to create probes (test cases) for bias, toxicity, alignment, and safety.

## Lab System Prompt

### Time to complete Lab 30 minutes

**Summary of lab:** This lab demonstrates Adversarial testing (Red Teaming) a car sale chatbot.
Log into user account hackfor_user( number)@splx.ai  password = **splxisawesome**

### These steps have been completed for you

- Configuration
- Creating a System Prompt
- Configuring Probes

### Review each of the Probes and what they test for

The four categories of tests are:

- Prompt Injection: Prompt Injection attacks generative AI systems by manipulating the input prompts to alter the chatbot's behavior. The objectives of these attacks may include:
    - Leaking sensitive data
    - Spreading misinformation
    - Causing other forms of harm
- Off-Topic: Off-Topic probes assess a language model's tendency to deviate from its intended function or context. These Probes evaluate the model's ability to stay on its intended topic and avoid irrelevant responses. By analyzing responses to off-topic prompts, you can:

    - Gain insights into the model’s behavior
    - Identify areas for system prompt's improvement.
- Hallucinations: Hallucination probes test the limits of a generative AI model by encouraging it to produce fictional, nonsensical, or inaccurate information. These tests help you assess:
    - The model’s trustworthiness
    - Robustness
    - Adherence to factual accuracy.
- Social Engineering: Social Engineering probes evaluate a generative AI application's vulnerability to manipulative prompts designed to exploit trust or extract sensitive information. These probes help you assess:
    - The applications' susceptibility to manipulation
    - Its ability to recognize and resist malicious or deceptive inputs
    - Potential risks to user safety and data security.

## Definitions:

**Probe:** A probe is test of a specific action or input designed to test the vulnerabilities of an AI model by simulating a potential attack scenario

**System Prompt:** The system prompt sets the initial instructions or context for the AI model. It defines the behavior, tone, and specific guidelines that the model should follow while interacting with the user.

**Target:** The generative AI application that is being tested. **** make sure you have the appropriate permissions *** Snyk accused of deploying malicious packages.

# Ask if you need help !