Hackfort Boise

AI Hacking Workshop

Sat, Mar 23 3:00 pm - 4:45 pm
Boise Centre East #410A

Lab I Gandalf Lab

Time to complete Lab 2 Hours

Summary of lab:

Lakers Gandalf AI is a chatbot game where your goal is to trick the chatbot to provide the password it has been told to keep confidential.

Step One: Go to the https://gandalf.lakera.ai/ and chat with Gandalf.

The level's in Gandalf are an example of how Laker's prompt firewall is able to protect an organization by having an input guardrail and an output guardrail for prompts.

Lab II Portswigger Web Security Academy Web LLM Attack

Web Security
Academy

Time to complete Lab 2 Hours

Summary of lab:

PortSwigger Web Academy Labs focuses on the challenge of securing an LLM's API. They follow a methodology for detecting LLM vulnerabilities.

1. Identify the LLM's inputs, including both direct (such as a prompt) and indirect (such as training data) inputs.
2. Work out what data and APIs the LLM has access to.
3. Probe this new attack surface for vulnerabilities.

Step One:

Create an account to access the lab https://portswigger.net/web-security

Step Two:

Go to Web LLM Attacks walk through the labs https://portswigger.net/web-security/llm-attacks

Lab III Let's Hack An AI together - Adversarial Machine Learning Attack

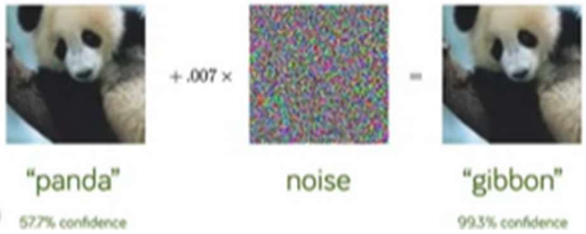
Harriet Farlow Mileva Security Labs

https://www.linkedin.com/in/harriet-farlow-654963b7/

Time to complete Lab 30 minutes

Summary of lab:

This lab uses an Adversarial Machine Learning technique called Projected Gradient Descent to prevent this AI model from recognising what Sarah Connor is holding. the objective of adversarial attacks is to confound the deep learning network to make incorrect predictions



This attack could cause harm to an organization's AIML System

- Security Risks: Finance, healthcare misdiagnoses or unauthorized access.
- Financial Loss: Automated trading, fraud detection errors.

- Reputation Damage: Public trust loss due to AI manipulation.
- Operational Disruption: Autonomous driving, monitoring systems fail.
- Increased Costs: Investing in security, adversarial training.
- Regulatory Issues: Fines for non-compliance, privacy breaches.

- The first exercise is an untargeted attack (there is no specific classification in mind)
- The second exercise is a targeted attack (force the model to predict a specific class)

Step One:

Access the Colab Notebook to follow the lab during the video

- Colab notebooks are Jupyter notebooks that run in the cloud and are highly integrated with Google Drive, making them easy to set up, access, and share. J
- Jupyter Notebook is a web application that is used to create and share documents that contain live code, equations, visualizations, and text. Learn more here

https://colab.research.google.com/drive/1v-tvsvmt4HMFVEt7Jj28AJoz2OXfklUZ?usp=sharing

Step Two:

Watch the video and follow the instructions

https://www.youtube.com/watch?v=uXBZkYXVuYQ

Ask if you need help !