LLM101 - DEFENDING AI

Course Learning Objectives

This course has been designed to provide you with an overview of the threat landscape regarding Artificial Intelligence. In module one, we'll take a look at the cybersecurity concerns around AI and how you can stay up-to-date with emerging standards. In module two, we'll examine security issues around deep learning models and how you can secure your training environment. In module three, we'll discuss how to deal safely with the output of an AI model, prevent misinformation and exposing sensitive data, and defend against prompt injection attacks. In module four, we'll look at defending against specialized attacks against AI systems, such as model inversion and membership inference attacks, excessive agency, and denial of service attacks. Finally, in module five, we'll take a closer look at the evolving regulatory landscape around creating an AI inventory, obtaining user consent, and mitigating AI bias.

Description

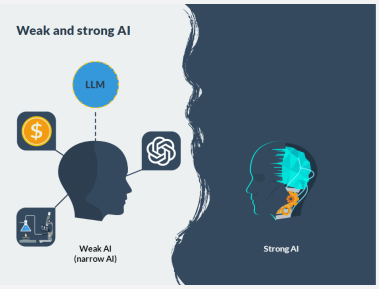
Defending AI is a course for anyone interested in learning more about the cybersecurity threats that affect AI systems, but is focused on Software Developers who use Large Language Models (LLMs), generative AI, and other AI tools. This course provides an overview of best security practices for using these tools and examines how the field is evolving rapidly. As more organizations begin to leverage AI in their day-to-day activities, it becomes increasingly important to practice AI security hygiene.

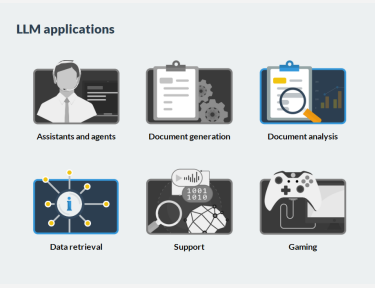
Audience

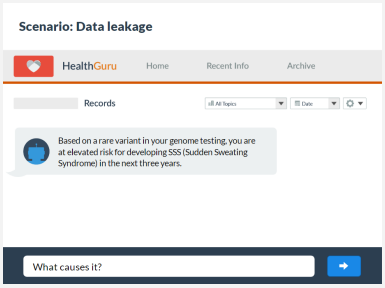
Software Developers

Time Required

Tailored learning - 70 minutes total







LLM101 - DEFENDING AI

Course Outline

1. The AI Cybersecurity Landscape

- The evolution of AI
- Deep learning
- Weak and strong AI
- AI through an API
- LLM applications
- Risk in AI
- AI security initiatives

2. Protecting the AI Model

- AI expands the attack surface
- Building AI models
- Securing the training environment
- Data leakage
- Data leakage in training
- Data poisoning
- Data poisoning examples
- Data poisoning defenses

3. Securing Model Interactions

- The opaque box problem
- Overreliance
- Hallucinations
- Mitigating overreliance
- Data exposure
- Prompt injection
- Defending against prompt injection
- Prompt engineering
- Prompt injection mitigations

4. Preventing AI Abuse

- Privacy attacks
- Membership inference
- Model inversion
- Defending against privacy attacks
- Excessive agency
- Defending against excessive agency
- Denial of service attacks
- Supply-chain vulnerabilities

5. AI Governance

- AI legislation
- AI inventory
- User consent
- Protecting user data
- AI bias
- Bias in AI legislation
- AI profiling
- The right to not be profiled

SecurityCompass