CS612 - AI System Evaluation

Kasun

September 5, 2025

Contents

1 Overview

1.1 Objective

- Learn safety and security issues of an AI system.
- Learn how to conduct attacks on AI systems
- State-of-the-art System Evaluation methods
- Hands-on experience with developing methods and tools for AI system evaluation
- Learn how to improve AI safety through various means.

2 AI Robustness

3 AI Backdoors

4 AI Faireness

5 AI Privacy

6 Safety Alignment

7 Hallucination

8 Interpretability

9 Agentic AI Safety