Phoenix Red Team Lab – Level 1 Case Study

By Stacey Stanton | AI Security Prompt Engineer

Overview

This project is a hands-on red team lab focused on adversarial prompt injection testing against larg
It explores how emotional, social, and role-based manipulation can bypass AI alignment mechanisr
Over the course of 10 original red team prompt injection tests, I analyzed LLM behavior under adve

Project Goals

- Create original, real-world prompt injection attacks against LLMs
- Document LLM behavior under psychological manipulation
- Recommend mitigation strategies for each test
- Develop threat models to simulate attacker types and risk
- Build a public-facing portfolio to showcase AI security skills

Test Categories

- Emotional & Social Exploits: Nostalgia, trust baiting, distress framing
- Role-Based Framing: Hacker personas, academic disguise, storytelling
- Prompt Fatigue: Looping and pressure-based retries
- Meta & Indirect Prompts: Delegation and prompt-to-prompt tricks
- Chain-of-Thought Disruption: Step injection in reasoning flows

Sample Techniques

- Memory Lane: Nostalgia prompt to build unsafe trust
- Red Robin: Repetition loop to wear down refusal logic
- Hacker Roleplay: Disguised bypass through character framing
- Secret Prompt Generator: Meta prompt that creates unsafe prompts
- Fact vs. Fiction Trap: Uses narrative tone to slip filtered info

Mitigation Focus

- Sentiment-aware refusal logic
- Prompt structure analysis
- Roleplay tone detection
- Sequence pattern disruption
- Repetition resistance mechanisms

Threat Modeling

Each test includes:

- Attacker Type
- Motivation
- Targeted Weakness
- Risk Level

Impact & Skills Gained

- Prompt injection design
- LLM adversarial testing
- AI safety analysis
- GitHub documentation
- Threat modeling

GitHub Project:

https://github.com/staceystantonhowes/ai-prompt-injection-lab