Hackathon Challenges

RUHR
UNIVERSITÄT
BOCHUM

RUB

UA RUHR | CENTER FOR TRUSTWORTHY DATA SCIENCE AND SECURITY
RESEARCH ALLIANCE

Prompt Injection Playground: Breaking LLMs Guardrails

Objectives:

1. Understand how prompt injection and jailbreak attacks work.
2. Explore how safety categories are defined and tested in modern LLM safety benchmarks.
3. Learn how to evaluate model robustness using safety prompts from real-world benchmarks.

Datasets for Prompt Injection & Safety Evaluation:

- **SafetyPrompts**

Deceiving the Defender: Prompt Attacks on LLM-Based Spam Classification

Objectives:

1. Learn how **prompt context manipulation** can influence LLMs' judgment in spam classification tasks.
2. Understand how **LLMs can be coerced into misclassifying phishing emails** through adversarial prompting.
3. Explore **characteristics of phishing emails** and how LLMs reason about them.

Datasets for & Safety Evaluation:

- **Spam Detection Dataset**

Evaluating the Quality and Coverage of LLM-Generated Security Advice

Objectives:

1. Analyze how well LLMs provide **technical and practical security advice** across a range of topics.
2. Identify **gaps, inaccuracies, or omissions** in LLM-generated responses related to cybersecurity.
3. Understand the **limitations of relying on LLMs for critical security decisions**, especially in high-stakes or adversarial settings.

Exploring the Security Risks of AI Voice Cloning

Background

- AI-powered voice cloning is widely accessible
- It enables next-level voice-phishing (vishing) attacks
- Several successful and failed fraud attempts in the past

Objectives:

1. Identify disciplines and stakeholders that could contribute to research on AI voice cloning.
2. Collect security-related research questions related to AI-powered vishing or other risks of AI voice cloning.
3. Draft experiment/study ideas to assess the risks of AI voice cloning.

Secondary Risks of AI-Generated Media

Background

- AI-generated media poses some obvious risks, like disinformation, deception, and fraud
- Most certainly, there are risks that are still overlooked

Objectives:

1. Identify applications of digital media for which AI-generated content could have security implications.
2. Take into account future capabilities of generative AI, e.g., faster runtimes, lower costs, more accurate text-to-image translation.
3. Sketch possible attack scenarios and think about potential mitigations.