Internship Proposal 2026

Project Title: Evaluating AI Agents for Automated Vulnerability Discovery

Project Description

Large Language Model (LLM)-based agents are increasingly being explored as autonomous vulnerability discovery tools, capable of invoking security analysis instruments such as fuzzers, static analyzers, and symbolic execution engines. However, there is limited empirical understanding of how effective these agents are, how they reason about security, and how robust their findings are compared to traditional approaches.

This project will design and execute a rigorous evaluation framework for AI agents tasked with vulnerability discovery. Interns will: (1) set up controlled benchmarking environments with known-vulnerable software (e.g., Magma, LAVA-M, or real-world CVE targets); (2) equip LLM agents (e.g., Claude, GPT-4, open-source models) with access to security tools and instrument their reasoning traces; (3) evaluate the agents across multiple dimensions including vulnerability detection rate, code coverage achieved, reasoning trajectory quality (step-by-step analysis of agent decision-making), and robustness (consistency of results across repeated runs and varied prompting strategies).

The expected outcome is a comprehensive benchmark dataset and evaluation report, along with a research paper suitable for submission to a top-tier security or AI venue. The work will contribute to the community's understanding of where AI agents excel and where they fall short in security analysis tasks.

Project Type: Research

Internship Batch

- Batch 1: May 10 to July 9, suitable for Education City students, i.e., CMUQ, TAMUQ and HBKU students

Duties/Activities

- Set up benchmarking environments with known-vulnerable targets (Magma, LAVA-M, real-world CVE reproductions)
- Configure and instrument LLM agents with access to security analysis tools (fuzzers, static analyzers, debuggers)
- Design evaluation metrics for reasoning trajectory quality, code coverage, and vulnerability detection
- Execute systematic experiments across multiple LLM agents and prompting strategies

- Analyze agent reasoning traces to identify common failure modes and successful patterns
- Evaluate robustness through repeated trials and prompt variation experiments
- Contribute to writing a research paper documenting the evaluation framework and findings

Required Skills

- Proficiency in Python and/or C/C++ programming
- Familiarity with software security concepts (common vulnerability classes, CVEs, exploitation basics)
- Experience with Linux environments and command-line tools
- Basic understanding of LLMs and AI agents (prompt engineering, API usage)
- Familiarity with at least one security analysis tool (e.g., AFL, AddressSanitizer, CodeQL, Ghidra)

Preferred Intern Academic Level: Junior or Senior undergraduate, or Master's student in Computer Science, Cybersecurity, or Software Engineering

Learning Opportunities

- Deep understanding of LLM agent architectures and tool-use paradigms in security contexts
- Hands-on experience with vulnerability discovery tools and benchmark frameworks
- Skills in designing rigorous empirical evaluations and reproducible experiments
- Exposure to state-of-the-art research at the intersection of AI and cybersecurity
- Experience with academic research methodology and paper writing

Expected Team Size: 2 interns

Mentors

Name: Ahmed Lekssays **Email:** alekssays@hbku.edu.qa