Benchmarking Cybersecurity LLM Benchmarks: Rigour, Robustness, and Multilingual Coverage

Project Description: Cybersecurity benchmarks are widely used to evaluate LLM safety, security, and robustness. However, their scientific rigour, data quality, and linguistic coverage remain under explored.

This project will systematically evaluate open-source cybersecurity benchmarks across:

- Dataset size, diversity, and balance
- Label correctness and validation methodology
- Reproducibility of scoring
- Robustness to adversarial manipulation
- Statistical power and variance
- Multilingual support (e.g., Arabic, English, French)

The multilingual dimension is critical: many benchmarks are English-only, yet LLMs are deployed globally. The project will evaluate benchmark coverage and performance stability across languages, including translation-based and native multilingual datasets.

Relevant sources may align with knowledge bases such as MITRE ATT&CK and vulnerability taxonomies from OWASP.

Deliverable: a meta-evaluation framework and empirical study suitable for submission to a top-tier AI security or measurement conference.

Project Type: Research

Internship Batch: Batch 1 or Batch 2

Duties/Activities:

- Curate open-source cybersecurity benchmarks
- Develop a benchmark evaluation taxonomy
- Re-run benchmarks across multiple LLMs
- Measure scoring stability and judge agreement
- Evaluate multilingual performance degradation
- Analyze adversarial fragility
- Draft a conference paper

Required Skills:

- Python
- LLM evaluation pipelines
- Basic cybersecurity knowledge
- Data analysis and statistics

Preferred Intern Academic Level: Senior undergraduate or MSc

Learning Opportunities:

- Experimental methodology in AI security
- Multilingual evaluation design
- Reproducible research practices
- Research publication process

Expected Team Size: 2–3 students

Mentors:

- Dr. Yazan Boshmaf (yboshmaf@hbku.edu.qa)
- Dr. Mohannad Alhanahnah (malhanahnah@hbku.edu.qa)