Project Title

Automated Benchmarking for LLM Privacy, Fairness, and Security

Project Description:

Static benchmarks for LLMs often fail to capture real-world risks due to data contamination (models memorizing evaluation data), and benchmark saturation (test sets becoming predictable and too easy). This project aims to develop a Dynamic Benchmarking Framework that automatically generates novel, high-fidelity test cases to evaluate model behavior across three critical domains: Privacy (PII leakage and privacy reasoning), Fairness & Bias (stereotypes, counterfactual consistency, demographic equity), Code Security (vulnerability generation and detection). The objective is to build an automated evaluation pipeline that continuously generates, filters, and scores new testing samples. In parallel, the team will design and implement a validation platform that enables human annotators to assess the quality, realism, and correctness of generated test cases.

Project Type: Research or Engineering

Internship Batch:

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

Duties/Activities:

1. Benchmark Analysis: Analyze existing static datasets in Privacy, Fairness, and Code Security. Identify coverage gaps, failure modes, and dataset artifacts. Propose measurable objectives for dynamic generation.
2. Pipeline Development (Test Case Generation): Design and implement algorithms for automated test generation for:
    - Privacy: synthesize realistic scenarios with synthetic PII, generate leakage probes, and contextual privacy reasoning tasks.
    - Fairness: Develop counterfactual generation pipelines (e.g., demographic term swapping). Measure consistency and bias amplification across sensitive attributes.
    - Code Security: Procedurally generate code snippets containing specific vulnerabilities (e.g., CWE categories). Create paired secure/insecure variants for detection evaluation.
3. Evaluation & Validation Platform: Design and implement a web-based platform for human annotators. Support labeling, quality scoring, and filtering of generated samples. Integrate feedback signals back into the generation pipeline.

Required Skills: React, Phyton, Basic understanding of Cybersecurity or AI Ethics.

Preferred Intern Academic Level: Undergraduate (Senior) or Graduate students in Computer Science, Cybersecurity, Data Science, or a related field.

Learning Opportunities:

- Synthetic Data Generation: Master techniques for creating high-quality training and testing data using AI agents.
- Domain Expertise: Deepen knowledge in AI Privacy standards, Fairness metrics, and Secure Coding practices.
- MLOps: Gain experience building scalable, automated evaluation pipelines.

Expected Team Size: A team of 4 members (Split across Privacy, Fairness, Code Security, and Infrastructure).

Mentors

Name: Dr. Fatih Deniz Email: fdeniz@hbku.edu.qa