Impact of LLM Tokenization Choices on Cybersecurity Tasks

Project Description: Tokenization serves as a fundamental initial step in training Large Language Models (LLMs). Text related to cybersecurity exhibits distinct characteristics compared to conventional language texts, notably due to its highly specialized terminology used to delineate threat behaviors. This specificity is captured through various indicators of an attack, including but not limited to IP/MAC addresses, file names and hashes, TTP (Tactics, Techniques, and Procedures) IDs, and YARA rules. However, the tokenizers employed in LLMs are not finely tuned to parse these unique indicators accurately, as security-specific text constitutes only a minor portion of the diverse corpus on which these tokenizers are trained. The aim of our project is to investigate and measure the impact that standard tokenizers have on the effectiveness of established models when applied to downstream tasks in the cybersecurity domain.

Project Type: Research and Engineering

Internship Batch:

- **Batch 1:** May 12 to July 12, suitable for Education City students, i.e., CMUQ, TAMUQ and HBKU students

Duties/Activities:

- Evaluating the compatibility of existing tokenizers with security tasks.
- Creation of a small performance benchmark to measure tokenization's impact.
- Running automated tests on several open and closed LLMs.
- Finetuning LLMs.

Required Skills:

- Fluency in programming, e.g, Python.
- Familiarity with AI/ML programming frameworks.
- Reading and discuss research papers.

Preferred Intern Academic Level: Senior

Learning Opportunities:

- Knowledge on LLM development lifecycle.
- Learning how to interact with language models.
- Developing analysis and reporting skills.

Expected Team Size: *1*

Mentors

Name: Dr. Husrev Taha Sencar email: hsencar@hbku.edu.qa

Name: Dr. Ahmed Lekssays email: alekssays@hbku.edu.qa