Improving IoT Intrusion Detection Resilience through LLM-based Adversarial Training

Deep Learning and other Machine Learning algorithms have been extensively used in recent research literature to facilitate traffic classification and anomaly detection-based intrusion detection systems for the Internet of Things (IoT). To protect such systems against adaptive adversarial attacks in which the attacker seeks to find attack samples that fool the detection model to classify them as benign, so-called *adversarial training* is employed in which specially-crafted adversarial perturbations of benign data points are included in the training dataset as negative examples in order to enhance the model's ability to reliably distinguish between positive and negative examples.

The task in this topic would be to investigate the use of pre-trained foundation LLM models for the purpose of augmenting existing IoT Intrusion Detection datasets for the purpose of adversarial training of IoT intrusion detection systems. The goal would be to implement and evaluate a framwork for enhancing the performance on state-of-the-art IoT intrusion detection systems with the goal of increasing the overall resilience of these systems through LLM-based adversarial learning.

To be able to successfully work on this topic, you should have a basic understanding about recent AI algorithms like Large Language Models, some familiarity with AI frameworks like PyTorch[1] or TensorFlow [2], and a willingness to learn and quickly adopt knowledge on AI, IoT security and intrusion detection.

Note: This represents only a preliminary high-level description of the topic. Details and actual scoping of the thesis will be agreed on with separately with the supervisor.

Thesis supervisor:

Prof. Dr.-Ing. Markus Miettinen
markus.miettinen@fra-uas.de
+49 69 1533 3969

Thesis language: English or German

Topic updated: June 11, 2025

[1] `https://pytorch.org/`
[2] `https://www.tensorflow.org/`