Large Language Models and Cybersecurity

Project description

Large language models (LLMs) are a type of artificial intelligence (AI) that are trained on massive datasets of text and code. They can now complete many complex text-based tasks rapidly and effectively, including generating text, translating languages, and writing different kinds of creative content.

In this work we would like to explore the interaction between LLMs and cybersecurity. LLMs can be applied in various ways to enhance cybersecurity. For example, LLMs are capable of analysing vast amounts of textual data, including forums, blogs, and news articles, to identify emerging cyber threats and trends. They can also automate the generation of security documentation and reports, helping organizations communicate effectively about their security posture. On the other hand, while LLMs becoming increasingly popular, they pose a number of security risks. The OWASP foundation identifies ten common security vulnerabilities that needs to be considered for developing LLMs' applications.

Within the context, here is two exemplary topics that a PhD applicant may like to consider.

Topic 1: Data leakage and privacy in use of Large Language Models

With this topic we are going to focus on the security risk associated with data leakage and privacy. For example, an LLM might learn from your prompts and offer that information to others who query for related things. Another example is that personally identifiable information (PII) is used in training a language model which might unintentionally generate outputs that reveal aspects of the PII. We plan to investigate a new way of sanitising sensitive information and developing security policies to mitigate this. We conduct an experiment by developing an LLM application and evaluating our security models with respect to how effective on preventing data leakage for the application.

Topic 2: Leveraging Large Language Models for generating access control models

Access control policies (ACPs) are typically part of security requirements, often written in natural language. ACPs articulate strict rules describing how access is managed, who may access which resources, and under what conditions. For instance, a doctor can change a patient's record in the healthcare system, but the nurse can only view the patient's record. These ACPs are needed to be translated into a formal representation (e.g. attribute-based access control (ABAC) model) which is then implemented as an access control mechanism (service) in computer systems to enforce the security controls.

In this work we would like to explore how to leverage LLMs to extract essential policy elements from ACP sentences and how to use these policy elements to generate a particular model that fits within a context in which the ACP sentences come from. We will define a set

of criteria or properties that the generated model must satisfy, and run experiments to evaluate the degree of satisfaction of the properties.

Candidate profile

The ideal candidate should have a first or upper-second class degree in Computer Science/Cybersecurity/Mathematics (or equivalent overseas qualification), and/or a merit Master's degree (or equivalent experience/qualifications). Prior scientific publications are particularly desirable but not essential. You should demonstrate a solid theoretical background and excellent software development skills. Strong commitment to reaching research excellence and achieving assigned objectives is required, as well as an ability to work in a collaborative environment.

Supervision Team

Dr. Liang Chen

Dr. Alexios Mylonas

31/01/2024