Detección de ataques adversariales sobre chatbots basados en grandes modelos de lenguaje

Propuesta de Trabajo para Memoria de Título

Semestre 2024-1

Contexto General

El avance en los Grandes Modelos de Lenguaje (LLMs), como Bert, GPT, LlaMA, entre otros ha demostrado ser uno de los logros tecnológicos más revolucionarios de nuestra época, teniendo estos aún un número insospechado de diversas aplicaciones. Uno de los usos más recurrentes y de importante crecimiento es la implementación de Chatbots [1]. Pero existen problemas importantes de seguridad, pues hay flancos abiertos asociados a corregir las alucinaciones de estos modelos, defenderlos de ataques adversariales y salvaguardar al privacidad de datos [2]. Esta memoria se centra en la detección de ataques adversariales como *prompt injection* lanzados sobre un chatbot en operación. En este sentido, la memoria deberá levantar el estado-del-arte existente e implementar dos enfoque progresivos. Primero uno basado en conocimiento experto, (detección de firmas de ataque). Para luego explorar formas de desarrollar un clasificador adaptativo que tome como referencia las salidas del chatbot monitoreado y su correlación con los prompt entrantes.

Resultados esperados

Al culminar el proyecto, se espera que el estudiante haya implementado un prototipo de sistema de detección de ataques adversariales para protección de chatbots basado en bibliotecas de recursos y ataques conocidos existentes y ejecutado pruebas para caracterizar su calidad como clasificador. Además se espera que explore el desarrollo de un clasificador evolutivo entregando evidencia sobre el mejor camino para trabajos futuros.

Patrocinante y Co-patrocinante: Pedro Pinacho Davidson (ppinacho@inf.udec.cl), Fernando Gutiérrez Gómez.

Esta Memoria de Título cuenta con financiamiento ANID a través del proyecto FONDECYT 11230359, "An Immune Inspired Model of Intrusion Prevention System (IPS) for Collaborative and Distributed Environments"

[1] Connor Weeks, Aravind Cheruvu, Sifat Muhammad Abdullah, Shravya Kanchi, Daphne Yao, and Bimal Viswanath. 2023. A First Look at Toxicity Injection Attacks on Open-domain Chatbots. In Proceedings of the 39th Annual Computer Security Applications Conference (ACSAC '23). Association for Computing Machinery, New York, NY, USA, 521–534. https://doi.org/10.1145/3627106.3627122.

[2] A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly, Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Eric Sun, Yue Zhang, arXiv:2312.02003, https://doi.org/10.48550/arXiv.2312.02003.