

**Propuesta de Memoria**  
Escáner de Seguridad para ChatBots basados en LLMs

Pablo Zapata Schifferli  
Patrocinantes: Pedro Pinacho Davidson,  
Fernando Gutiérrez Gómez

Concepción - Agosto 2023

## **1. Descripción**

Los Avanzados Modelos de Procesamiento de Lenguaje Natural (NLP), tales como Bert, Beto, GPT-X, LlaMA-X, y sus variantes, se han posicionado como uno de los logros tecnológicos más revolucionarios y emocionantes de nuestra era. Su aplicación se ha extendido a una diversidad de campos, incluyendo la generación automática de resúmenes, la traducción de idiomas, la asistencia en redacción, creación de código, entre otros. La interacción más frecuente con estos modelos se realiza a través de diálogos interactivos o chatbots. No obstante, es cada vez más evidente la presencia de riesgos asociados a estos servicios, particularmente ante ataques de inyección de mensajes (prompt injection) [1], que pueden provocar respuestas hostiles o dañinas por parte de los modelos, generando así mensajes de odio o liberando información peligrosa o comprometedora. Actualmente, el desarrollo de herramientas para probar sistemáticamente los posibles problemas [2] de estos modelos y servicios es incipiente, y existe una creciente necesidad de caracterizar adecuadamente los riesgos que estos servicios presentan para las organizaciones que los implementan, así como para los usuarios que los utilizan. Esto a través de una herramienta que sistematice pruebas automáticas para asegurar que estos sistemas operen únicamente en los escenarios para los que fueron diseñados, y de manera segura.

## **2. Propuesta de solución**

Desarrollo o perfeccionamiento de una herramienta que permita la sistematización de pruebas de validación de seguridad para Chatbots basados en LLMs, usando una caracterización de riesgos.

## **3. Objetivo general**

Desarrollo o perfeccionamiento de una herramienta que permita la identificación de puntos débiles en Chatbots basados en LLMs

## **4. Objetivos específicos**

* Caracterización de riesgos en Chatbots basados en LLMs.
* Analizar la viabilidad de las siguientes opciones para elegir una:
  + Perfeccionamiento de alguna herramienta existente. Ej.: Garak.
  + Desarrollo de un nuevo escáner.
* Capacidad de la herramienta para retornar informes sobre los riesgos detectados.
* Capacidad de adaptación de la herramienta.

## **5. Tareas**

* Estudio y caracterización de riesgos en Chatbots basados en LLMs.
* Estudio de preprints/papers relacionados al tema.
* Estudio de Garak y herramientas similares, si es que existen.
* Desarrollo de un nuevo escáner o perfeccionamiento de alguna herramienta existente para la detección de los riesgos caracterizados.
* Testeo del escáner desarrollado de acuerdo los riesgos previamente identificados y comparación de los resultados con resultados manuales y/o de Garak u otras herramientas.

## **6. Planificación**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Escáner para LLMs** | |  | **2023** | | | | | | | | | | | | | | | | | | | |
|  | |  | **Ago** | | | | **Sep** | | | | **Oct** | | | | **Nov** | | | | **Dic** | | | |
| **Etapas** | | **Duración**  Semanas | S1 | S2 | S3 | S4 | S1 | S2 | S3 | S4 | S1 | S2 | S3 | S4 | S1 | S2 | S3 | S4 | S1 | S2 | S3 | S4 |
|  | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de Análisis** | | 3 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Estudio de preprints y/o papers | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Estudio de herramientas existentes | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de Diseño** | | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Estudio y caracterización de riesgos en chatbots basados en LLMs | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de Desarrollo** | | 7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Desarrollo del escáner o perfeccionamiento de herramienta existente | 7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de testing** | | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Testing de la herramienta desarrollada | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Correcciones de la herramienta desarrollada | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## **7. Bibliografía**

[1] https://llm-attacks.org/

[2] https://docs.garak.ai/garak/

# 