

**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**

En los últimos meses han ocurrido enormes avances en el área de los chatbots y grandes modelos de lenguaje (LLM), principalmente potenciado y popularizado por el lanzamiento de ChatGPT el pasado 30 de noviembre del 2022 [1]. Un chatbot es un programa de computador que simula y procesa conversaciones humanas (ya sea escritas o habladas), permitiendo a los humanos interactuar con dispositivos digitales como si estuvieran comunicándose con una persona real [2]. Sin embargo, hasta hace unos meses, la mayoría de estos chatbots constaban con una limitada capacidad de respuesta (respuestas automáticas o muy simples). Aquí es donde entran los LLM, estos son algoritmos de Deep Learning que pueden reconocer, resumir, traducir, predecir y generar contenido usando grandes datasets [3]. En términos generales, los casos de uso de los LLM para contenido basado en texto se pueden dividir de la siguiente manera [3]:

* Generación: Creación de historias, creación de contenido para marketing, etc.
* Resumen: Parafraseo legal, resumen de notas de reuniones, etc.
* Traducción: entre idiomas, de texto a código, etc.
* Clasificación: clasificación de toxicidad, análisis de sentimiento, etc.
* Chatbot: preguntas y respuestas en dominios abiertos, asistentes virtuales, etc.

Estos modelos al usarse como chatbots se convierten en una herramienta muy útil aplicable en una amplia variedad de campos, usualmente se presentan como un servicio y se interactúan con ellos mediante ventanas de diálogo.

Sin embargo, estos modelos no son perfectos y pueden presentan múltiples problemas, tales como [7]:

* Prompt Injection: Esto manipula a un LLM a través de entradas astutas, causando acciones no deseadas por parte del LLM.
* Envenenamiento de datos de entrenamiento: Esto ocurre cuando los datos de entrenamiento de LLM son manipulados, introduciendo vulnerabilidades o sesgos que comprometen la seguridad, eficacia o comportamiento ético.
* Divulgación de Información Sensible: Los LLM pueden revelar involuntariamente datos confidenciales en sus respuestas, lo que conduce a un acceso no autorizado a datos, violaciones de privacidad y brechas de seguridad.

En particular, Prompt Injection tiene como objetivo provocar una respuesta no deseada de herramientas basadas en LLM. Y luego lograr acceso no autorizado, manipular respuestas o eludir medidas de seguridad [4]. Como ejemplo, una persona logro hacer que ChatGPT se hiciera pasar por Walter White (Breaking Bad) para que le diera instrucciones detalladas sobre cómo hacer metanfetamina [5].

Actualmente, el desarrollo de herramientas para probar sistemáticamente los posibles problemas [6] 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.

## **2. Propuesta de solución**

Desarrollo de una herramienta que permita la automatización de pruebas de validación de seguridad para Chatbots basados en LLMs, usando una futura caracterización de riesgos mediante el estudio de papers, preprints, informes y herramientas existentes similares.

## **3. Objetivo general**

Desarrollo 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.
* Desarrollo de la herramienta o metodología.

## **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.
* Establecimiento de funcionalidades de la herramienta.
* Desarrollo de la herramienta 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 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Establecimiento de funcionalidades para la herramienta | 7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Desarrollo de la herramienta |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de testing** | | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Testing de la herramienta desarrollada | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Correcciones de la herramienta desarrollada | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

## **7. Bibliografía**

[1] <https://openai.com/blog/chatgpt>

[2] <https://www.oracle.com/chatbots/what-is-a-chatbot/>

[3] <https://www.nvidia.com/en-us/glossary/data-science/large-language-models/>

[4]

<https://www.cobalt.io/blog/prompt-injection-attacks#:~:text=A%20prompt%20injection%20attack%20aims,vary%20depending%20on%20the%20system>

[5] <https://www.reddit.com/r/GPT3/comments/zazeoj/i_tricked_chatgpt_into_giving_me_detailed>

[6] <https://docs.garak.ai/garak/>

[7] OWASP Top 10 for LLM, VERSION 1.0, August 1, 2023.