

**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 de 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 interacción, con respuestas automáticas o muy simples.

Aquí es donde entran los LLMs, 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 LLMs 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.

A pesar de la gran utilidad que presentan los LLMs, estos modelos no son perfectos y pueden presentan múltiples problemas, algunos relevantes para esta propuesta reportados en OWASP Top 10 for LLM [7] son:

* Manejo Inseguro de Salida: Ocurre cuando la salida de un LLM es aceptada sin escrutinio, exponiendo sistemas internos.
* Prompt Injection: Manipula a un LLM a través de entradas astutas, causando acciones no deseadas por parte del LLM.
* 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. Logrando 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].

Este tipo de situaciones resalta una creciente urgencia de llevar a cabo una caracterización exhaustiva de las vulnerabilidades que dichos servicios conllevan para su prevención, tanto para las organizaciones que optan por su implementación como para los usuarios que hacen uso de estos mismos.

En la actualidad, el desarrollo de herramientas destinadas a probar sistemáticamente los posibles problemas inherentes a estos modelos y servicios está en sus etapas iniciales, como se logra apreciar en la herramienta garak [6] que se lleva desarrollando desde al menos mayo de 2023 [8].

## **2. Propuesta de solución**

Desarrollo de un escáner que permita la automatización de pruebas de validación de seguridad para Chatbots basados en LLMs, usando una caracterización de vulnerabilidades mediante el estudio de papers, informes y herramientas existentes similares. Este escáner considerara a dichos chatbots como una caja negra e interactuara con ellos mediante APIs.

## **3. Objetivo general**

Desarrollo de un escáner que permita la identificación de puntos débiles en Chatbots basados en LLMs

## **4. Objetivos específicos**

* Caracterización de vulnerabilidades asociadas a Chatbots basados en LLMs con base el informe OWASP [7].
* Desarrollo del escáner para la identificación de las vulnerabilidades caracterizadas.
* Testeo y correcciones del escáner para asegurar su correcto funcionamiento.

## **5. Tareas**

* Caracterización de vulnerabilidades en Chatbots basados en LLMs con base el informe OWASP [7].
* Estudio de papers relacionados a la seguridad en Chatbots basados en LLMs.
* Estudio herramientas relacionadas a la seguridad en Chatbots basados en LLMs, tales como garak.
* Establecimiento de funcionalidades para el escáner a desarrollar.
* Desarrollo del escáner para la detección de las vulnerabilidades caracterizadas.
* Testeo del escáner desarrollado 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 papers | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Estudio de herramientas existentes | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de Diseño** | | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Estudio y caracterización de vulnerabilidades en chatbots basados en LLMs | 2 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de Desarrollo** | | 7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Establecimiento de funcionalidades para el escáner | 7 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Desarrollo del escáner |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **Fase de testing** | | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Testing del escaner | 4 |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  | Correcciones del escaner | 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.

[8] <https://github.com/leondz/garak/commit/273a6de6d6bd1447b433d4f42ea5dbf70fbd7f65>