

Universidad de Concepción

Ingeniería Civil Informática

Propuesta de memoria

Generación de ataques Spear Phishing usando Grandes Modelos de Lenguaje

Rodrigo Ignacio San Martin Garcia

Patrocinantes:

Pedro Pablo Pinacho Davidson

Fernando Andree Tercero Gutierrez Gomez

Concepción - Agosto 2023

1. Descripción

El spear phishing es una forma de ataque cibernético que utiliza correos electrónicos personalizados y engañosos para robar información confidencial. A diferencia del phishing tradicional, que se dirige a un gran número de usuarios al azar, el spear phishing se enfoca en objetivos específicos, como empleados de una organización, clientes de un banco o figuras públicas [1]. Los atacantes investigan previamente a sus víctimas para diseñar mensajes convincentes que se aprovechen de su confianza, curiosidad o miedo. El spear phishing representa una grave amenaza para la seguridad de los individuos y las entidades ya que puede causar graves daños a las víctimas, como robo de identidad, fraude o pérdida de dinero. Por eso, es importante conocer las características y los métodos de los ataques [3].

Los ataques phishing presentan un proceso de desarrollo el cual ayuda a entender los pasos para llegar a obtener los datos privados de las personas. Dentro de los pasos se encuentra la preparación del ataque, donde se tienen en cuenta cosas como, definir el medio de comunicación, establecer el objetivo del ataque, que técnicas se estarán usando y preparar el material a usar, luego se pasa a un proceso de ejecución donde se aplica lo antes definido y se obtienen datos personales para finalmente explotar estos datos [1].

Actualmente los grandes modelos de lenguaje (LLMs) pueden asistir en tareas como el reconocimiento y la generación de texto [4], gracias a esto se pueden automatizar procesos del desarrollo del ataque antes mencionado, donde ya teniendo el material a usar se le puede dar de contexto al modelo y que este genere los correos de ataque, siendo completamente personalizados para el objetivo, además gracias al reconocimiento de texto del modelo se puede estudiar como funciona un ataque persistente, donde se pide al objetivo responder los correos y que el modelo al leer las respuestas pueda obtener más información del usuario y contestarle durante el tiempo que sea necesario [2].

Investigar cómo se desarrollan ataques de este tipo usando LLMs podría ayudar a generar herramientas que prevengan futuros daños [4].

1. Propuesta de solución

Realizar pruebas de concepto con un modelo el cual use información recopilada de múltiples personas y genere correos personalizados de manera automática en el tiempo, teniendo la capacidad de mantener un diálogo claro con el objetivo a la vez que recopila información de las respuestas. De esta manera evaluar el peligro que puede generar una herramienta de este tipo.

1. Objetivo general

* Generar un modelo que pueda crear correos spear phishing a partir de datos de personas, y que mantenga el contacto con estas ante posibles respuestas.

1. Objetivos específicos

Dentro de los objetivos específicos se encuentran:

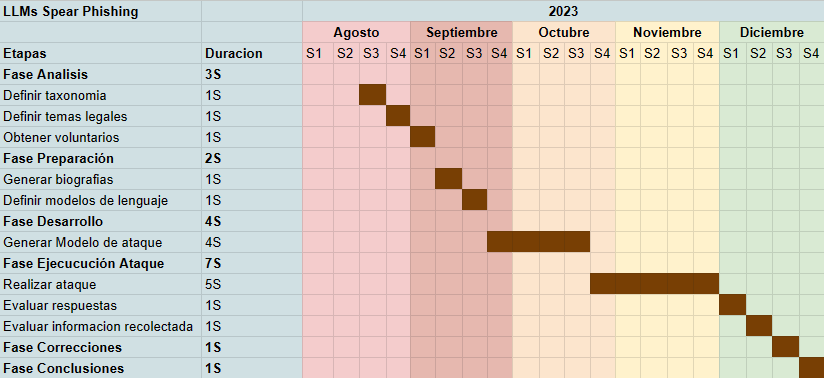
* Implementar modelo generación spear phishing que mantenga el diálogo con el receptor.
* Buscar o generar conjunto de datos de prueba que contengan información específica de cada usuario para la generación de los correos .
* Estudiar la evolución y amenaza de ataques de este tipo.
* Validar la calidad del phishing viendo la cantidad de información que es capaz de recolectar de cada persona.

1. Tareas

Las tareas a realizar son:

* Establecer taxonomía actual para el phishing y proponer una extensión.
* Definir temas legales.
* Obtener voluntarios para el entrenamiento del modelo más los permisos para usar su información.
* Generar múltiples biografías personales de los sujetos de prueba (usando contexto personal más rasgos psicológicos).
* Definir modelos de lenguaje a utilizar.
* Entrenar modelos que generen ataques personalizados para cada individuo.
* Evaluar el proceso de respuesta del modelo, viendo que mantenga un contexto apropiado según la respuesta del sujeto de prueba.
* Evaluar la cantidad y calidad de la información recolectada para ver el rendimiento del modelo.

1. Planificación temporal



1. Referencias

[1] Aleroud, A., & Zhou, L. (2017). Phishing environments, techniques, and countermeasures: A survey. *Computers & Security*, *68*, 160-196.

[Phishing environments, techniques, and countermeasures: A survey - ScienceDirect](https://www.sciencedirect.com/science/article/pii/S0167404817300810)

[2] Hazell, J. (2023). Large language models can be used to effectively scale spear phishing campaigns. *arXiv preprint arXiv:2305.06972*.

[[2305.06972] Large Language Models Can Be Used To Effectively Scale Spear Phishing Campaigns (arxiv.org)](https://arxiv.org/abs/2305.06972)

[3] Gomes, V., Reis, J., & Alturas, B. (2020, June). Social engineering and the dangers of phishing. In *2020 15th Iberian Conference on Information Systems and Technologies (CISTI)* (pp. 1-7). IEEE.

[Social Engineering and the Dangers of Phishing | IEEE Conference Publication | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/9140445)

[4] Gupta, M., Akiri, C., Aryal, K., Parker, E., & Praharaj, L. (2023). From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*.

[From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy | IEEE Journals & Magazine | IEEE Xplore](https://ieeexplore.ieee.org/abstract/document/10198233)