Universidad de Concepción
CAMPUS CHILLÁN

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].

2. Propuesta de solución

Realizar pruebas de concepto con una herramienta la cual use información recopilada de múltiples personas para generar correos de ataque phishing personalizados de manera automática, teniendo la capacidad de recopilar información de las respuestas. De esta manera evaluar el peligro que puede generar una herramienta de este tipo.

3. Objetivo general

- Generar una herramienta que pueda crear correos spear phishing a partir de datos de personas.

4. Objetivos específicos

Los objetivos específicos son:

1. Estudiar la evolución y amenaza de ataques de este tipo.
2. Buscar o generar conjunto de datos de prueba que contengan información específica de cada usuario para la generación de los correos.
3. Implementar herramienta de generación spear phishing.
4. Validar la calidad del phishing a través de encuestas realizadas a los sujetos de prueba.

5. Tareas

Las tareas a realizar son:

1. Establecer taxonomía actual para el phishing y proponer una extensión.
2. Definir qué datos personales pedir a los usuarios.
3. Obtener voluntarios para el entrenamiento del modelo más los permisos para usar la información solicitada a través de un documento.
4. Generar múltiples biografías personales de los sujetos de prueba.
5. Definir modelos de lenguaje a utilizar.
6. Desarrollar herramienta que genere ataques personalizados para cada individuo.
7. Evaluar la calidad del phishing generado usando encuestas donde el usuario puntúe qué tan probable era que diera su información.
8. Elaboración del informe.

6. Planificación temporal

LLMs Spear Phishing		2023																			
		Agosto				Septiembre				Octubre				Noviembre				Diciembre			
Etapas	**Duracion**	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4	S1	S2	S3	S4
Fase Analisis	**3S**																				
Definir taxonomia	1S																				
Definir documento de consentimiento	1S																				
Obtener voluntarios	1S																				
Fase Preparación	**2S**																				
Generar biografias	1S																				
Definir modelos de lenguaje	1S																				
Fase Desarrollo	**4S**																				
Generar herramienta de ataque	4S																				
Fase Ejecucución Ataque	**7S**																				
Realizar ataque	5S																				
Evaluar respuestas	1S																				
Evaluar informacion recolectada	1S																				
Fase Elaboración del informe	**1S**																				
Fase Correcciones	**1S**																				
Fase Conclusiones	**1S**																				

7. 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

[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)

[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

[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