Universidad de Concepción
Facultad de Ingeniería
Dpto. de Ing. Informática y Cs. de la Computación

Hacia el desarrollo de una herramienta de evaluación de vulnerabilidades de ingeniería social

Por
Maximiliano Vergara Ríos

Memoria presentada para la obtención del título de
Ingeniero Civil Informático

Patrocinantes:
Pedro Pablo Pinacho Davidson
Fernando Andree Tercero Gutierrez Gomez

Comisión Evaluadora: Prof. Pierluigi Cerulo, Prof. Marcela Varas C.

Concepción, junio 2025

Índice

Contenido

1 Introducción

Un sistema de información es un conjunto organizado de componentes que recogen, procesan, almacenan y distribuyen datos para facilitar la toma de decisiones dentro de una organización. Para ello, integra infraestructura tecnológica, sistemas de aplicación y personal especializado que emplea la tecnología de la información en la gestión de operaciones, el procesamiento de transacciones y la administración organizacional [1].

Para la defensa del sistema y protección de los datos involucrados, los equipos de ciberseguridad de las organizaciones se apoyan en marcos de referencia de ciberseguridad que aportan una guía estructurada de como ejecutar buenas prácticas, estrategias y controles para implementar la seguridad en la organización. El propósito de esto es mantener la información asegurada basándose en los 3 principios fundamentales de la seguridad de la información; confidencialidad, integridad y disponibilidad [2].

Centrándonos en la confidencialidad, esta se refiere a la preservación de las restricciones autorizadas sobre el acceso y la divulgación de la información, incluyendo medios para proteger la privacidad personal y la información propietaria [3]. En la práctica, salvaguardar la confidencialidad de los datos implica una capacitación de las personas con acceso a estos, ayudando así a que estos se familiaricen con los factores de riesgo y también aprendan a cómo protegerse contra ellos.

Esta capacitación es crucial si consideramos que las personas representan un punto crítico de vulnerabilidad dentro de un sistema de información, ya que, a diferencia de una máquina, pueden ser manipuladas mediante métodos que explotan sus conductas. Un claro ejemplo de esto es la *ingeniería social*, un conjunto de técnicas en las que los delincuentes se aprovechan de la confianza o la falta de conocimiento en ciberseguridad, convirtiendo a las personas en puntos de entrada vulnerables para comprometer la seguridad de los sistemas [4]. A través de diversas técnicas que emplean la ingeniería social como vector de ataque, los atacantes explotan o se aprovechan de vulnerabilidades inherentes a la naturaleza humana y su comportamiento. De hecho, características humanas, tales como el exceso de confianza, la predisposición a ayudar y la desesperación son factores que las estafas cibernéticas explotan con gran éxito hoy en día [5]. Utilizando estas características, el atacante manipula

a la víctima para obtener acceso a sistemas e información tanto personal como de las organizaciones a las que pertenecen.

Entre los ataques más conocidos que emplean ingeniería social se encuentran el *phishing* y el *spear phishing.* Ambos utilizan como forma de ataque el fraude en línea, que consiste en enviar correos electrónicos fraudulentos o crear sitios web falsos que imitan a organizaciones legítimas para engañar a los usuarios y provocar que desplieguen acciones que comprometan su sistema, información personal o confidencial de una organización. Estas formas de ataque se basan en la suplantación de identidad, la manipulación psicológica de las víctimas y la creación de entidades falsas para obtener acceso a datos sensibles. La diferencia entre ambos radica en que el *phishing* es un ataque masivo y no dirigido, mientras que el *spear phishing* se enfoca en individuos o grupos específicos, mediante mensajes personalizados con el fin de obtener una mayor probabilidad de éxito [6] .

Dado el impacto que tienen estos ataques, muchas organizaciones han adoptado estrategias para reforzar sus medidas de seguridad y mitigar su efectividad. Estas han comenzado a fortalecer la formación de sus integrantes en el área ciberseguridad, logrando una mayor concientización de los peligros inherentes a internet. Además, han recurrido a métodos de defensa tales como filtrado de correos y tecnologías capaces de identificar este tipo de amenazas por medio de análisis y detección de patrones en la estructura y redacción de los mensajes de phishing [7]. No obstante, aunque los métodos de defensa hayan evolucionado, también lo han hecho los métodos de confección de ataques de phishing (y sus variantes). Esto de la mano de que su desarrollo y ejecución son relativamente sencillas a comparación de otros métodos de ataque, ha provocado a lo largo de los años una popularización y masificación del phishing [8]. Como consecuencia, el 68 % de las brechas de seguridad registradas durante el año 2024 estuvieron relacionadas con el factor humano. De estas, un 80 % tuvo al phishing como vector inicial de ataque [9].

A este escenario debemos añadirle la reciente irrupción de la inteligencia artificial, que ha redefinido por completo el estado del arte, elevando las capacidades ofensivas y defensivas a un nivel de complejidad mucho mayor. El caso más significativo es el arribo de los LLM (Grandes modelos de lenguaje), modelos de aprendizaje automático diseñados para predecir, comprender y generar respuestas relevantes y coherentes en función del contexto proporcionado [10]. Por este motivo, los LLM podrían ayudar a personalizar los ataques de

spear phishing para que sean más convincentes y específicos para cada objetivo. Al analizar el lenguaje utilizado por un individuo en sus correos electrónicos, publicaciones en redes sociales u otra información disponible en línea, un atacante podría utilizar un LLM para generar mensajes de phishing altamente personalizados que parezcan más legítimos y difíciles de detectar.

Trabajos recientes, como los realizados por el grupo de investigación "IA & Ciberseguridad" de la Universidad de Concepción, han comenzado a integrar el factor humano en sus estudios. Estos trabajos incluyen la evaluación social de la peligrosidad de los ataques de *phishing* utilizando LLMs. Los resultados obtenidos basados en encuestas a participantes destacan las vulnerabilidades humanas más susceptibles a estos ataques [11].

Estas investigaciones representan un primer paso en la creación de herramientas asistidas por LLMs para evaluar y mitigar vulnerabilidades ante ataques de ingeniería social. A partir de los resultados obtenidos, es posible diseñar estrategias de capacitación y entrenamiento más efectivas, adaptadas a las debilidades identificadas en los participantes. La continuidad de este proyecto permitirá no solo comprender mejor la efectividad de estos ataques, sino también proporcionar insumos concretos para fortalecer la preparación organizacional frente a estas amenazas.

Con el propósito de seguir con esta línea de investigación se propone evaluar la peligrosidad de los ataques de *spear phishing*, cambiando el enfoque a un ámbito organizacional. Así, este proyecto podrá ofrecer soluciones a grupos como empresas e instituciones a detectar sus vulnerabilidades sociales y entregando así una base para su evitar en el futuro ataques de esta índole.

1.1 Solución propuesta

Se propone desarrollar un *framework* que evalúe las vulnerabilidades de ingeniería social de organizaciones a través de ataques de *spear phishing*. Obteniendo el permiso de las organizaciones se emplearán técnicas de extracción de datos para recopilar su información. Luego se diseñarán *prompts* que instruyan a un LLM a aplicar principios de ingeniería social para generar correos de s*pear phishing* en base a los datos recolectados. Finalmente, los correos serán enviados a miembros de las organizaciones durante un período de tiempo específico y sus respuestas se utilizarán para analizar las vulnerabilidades y evaluar la preparación del personal frente a posibles ataques reales de *spear phishing*.

El desarrollo de este proceso sería la primera versión de un servicio que las organizaciones podrán utilizar para analizar y mejorar sus mecanismos de defensas contra posibles ataques.

1.2 Objetivo General

Desarrollar un *framework* para evaluar vulnerabilidades de ingeniería social organizacional mediante el envío de correos de *spear phishing* y hacer pruebas de su funcionamiento en una entidad definida.

1.2.1 Objetivos específicos

- Investigar la legislación nacional vigente aplicable en esta materia.
- Investigar acerca del estado del arte en las áreas de métodos recopilación y de generación de datos.
- Desarrollar una herramienta automática de *web scraping* que mediante un LLM se adapte al formato páginas web.
- Desarrollar una herramienta de generación de correos de *spear phishing* que mediante una LLM automatice el proceso.
- Ejecutar una prueba operacional que evalué el funcionamiento del *framework*.

2 Marco teórico

El uso de modelos de lenguaje para el manejo de información plantea desafíos que van más allá de su capacidad técnica. La regulación del tratamiento de datos, la seguridad de la información y las estrategias utilizadas para influir en las personas son factores que deben considerarse para comprender sus implicancias y riesgos. Por ello, es fundamental analizar los principios y marcos que permiten un uso responsable y seguro de estas tecnologías.

2.1 Marco normativo

Se entiende como marco normativo el conjunto de leyes, reglamentos, políticas y normas que establecen el contexto legal y regulatorio en el que operan las organizaciones o instituciones. Esencialmente, establece las reglas y estándares que deben respetarse para garantizar el cumplimiento legal, ético y operativo en una determinada área o contexto.

En este caso, separaremos la información en 3 factores; **los datos con los cuales trabajaremos, la forma en que se recopilarán y cómo se utilizarán estos datos**. Esto con el fin de garantizar la legalidad y seguridad de los datos en todo momento. Para esto nos respaldaremos en las leyes pertenecientes a la legislación chilena relevantes en materia de ciberseguridad.

- **Ley 21.663**: Regula la recolección, el uso y la protección de los datos personales, con el objetivo de garantizar la privacidad y la seguridad de los ciudadanos, asegurando que los datos sean tratados de manera ética y conforme a la normativa vigente.
- **Ley 21.459**: Establece normas sobre delitos informáticos, deroga la ley 19.223 (tipificación de figuras penales relativas a la informática) y modifica otros cuerpos legales con el objeto de adecuarlos al Convenio de Budapest.
- **Ley 21.719**: La Ley de Protección de Datos Personales tiene el objetivo de regular la forma y condiciones en las que se realiza el tratamiento de este tipo de información y mejorar la protección de los derechos de sus titulares, estableciendo y regulando detalladamente cuáles son esos derechos, además del procedimiento y los medios para que los titulares hagan valer estas garantías ante los responsables de datos.

- **Ley 19.628**: Sobre la protección de la vida privada, da una definición a que son datos personales y sensibles y como regula como las entidades deben recolectarlos y tratarlos.
- **Ley 19.799**: Ley sobre documentos electrónicos, firma electrónica y servicios de certificación de dicha firma.

Comenzando con **los datos con los cuales trabajaremos**, debemos considerar que la **Ley 19.628** define los siguientes datos:

- **Datos de carácter o datos personales**: Relativos a cualquier información concerniente a personas naturales, identificadas o identificables.
- **Datos sensibles**: aquellos datos personales que se refieren a las características físicas o morales de las personas o a hechos o circunstancias de su vida privada o intimidad, tales como los hábitos personales, el origen racial, las ideologías y opiniones políticas, las creencias o convicciones religiosas, los estados de salud físicos o psíquicos y la vida sexual.

En este contexto, si los datos recopilados permiten identificar a los participantes, se consideran datos personales y deben ser protegidos conforme a esta ley. Además, la ley establece que no se podrán recolectar estos datos sin el consentimiento explícito de los participantes. En caso de obtener dicho consentimiento, es imprescindible informar a los participantes sobre los datos que serán recolectados y el propósito de su uso.
Además, la **Ley 21.459** establece que la información recopilada debe ser justificada y limitarse a lo estrictamente necesario para la ejecución del estudio en cuestión. La ley también asigna responsabilidades a las personas autorizadas para tratar estos datos (siendo incluidos los procesos recolección, almacenamiento y procesamiento), y establece que solo ellas tienen acceso a los datos del titular. Estas personas tienen la obligación de mantener la confidencialidad de la información tanto durante el uso de los datos como después de su utilización, y deben garantizar que los datos se conserven en un entorno seguro.

Respecto a la **recopilación de datos**, debemos ceñirnos a lo establecido por la **Ley 21.663**, que detalla el procedimiento para una recopilación adecuada. Esta ley establece que no se podrá acceder a los sistemas informáticos del titular (como bases de datos o páginas

web) sin su autorización explícita. Se entiende por sistemas informáticos “todo dispositivo aislado o conjunto de dispositivos interconectados o relacionados entre sí, cuya función, o la de alguno de sus elementos, sea el tratamiento automatizado de datos en ejecución de un programa”. Un ejemplo de violación a esta normativa sería el uso de técnicas para recopilar datos de la base de datos del titular sin su consentimiento previo.

Esta autorización puede ser obtenida digitalmente por medio de documentos electrónicos, tal como lo expresa la **Ley 19.799**, que además indica que este documento debe contar con una firma electrónica valida y su información debe ser almacenada de forma segura y verificable por parte de las personas responsables.

Finalmente, debemos definir como se **utilizarán a estos datos**, ateniéndonos a los permitidos por la ley además de conocer las obligaciones que conllevan su almacenamiento y posterior eliminación.

Según la **Ley 21.719,** se exige documentar todo el proceso del uso de los datos del titular en caso de ser solicitado por este. Además, las distintas etapas del proceso deben incluir medidas que garanticen la protección de los datos en todo momento.

La ley también establece que el uso de estos datos es exclusivo de la prueba para la que fueron solicitados, prohibiendo su reutilización sin el consentimiento explícito del titular, y restringiendo su uso para fines distintos a los informados inicialmente.

Una vez finalizado el proceso para el cual los datos fueron solicitados, el responsable del tratamiento está obligado a eliminar de manera segura cualquier dato que haya sido utilizado para su propósito inicial.

En caso de no cumplir alguna de las normas presentadas, la prueba de vulnerabilidad será catalogada como un incidente de ciberseguridad, definido dentro de la **Ley 21663** (Ley Marco de ciberseguridad) como “todo evento que perjudique o comprometa la confidencialidad o integridad de la información, la disponibilidad o resiliencia de las redes y sistemas informáticos, o la autenticación de los procesos ejecutados o implementados en las redes y sistemas informáticos”.

Dentro de los incidentes de ciberseguridad establecidos dentro de la “Taxonomía de Incidentes de Ciberseguridad” de la Agencia Nacional de Ciberseguridad (ANCI) [12] que son relevantes para este contexto, se incluyen:

- Envío de correo no deseado o *phishing* sobre una organización.
- Envío de correo no deseado o *phishing* usando remitentes de la institución.
- Inyección de requerimientos (*prompts*) en modelos grandes de lenguaje (LLM).
- Acceso no autorizado a almacenamiento.
- Filtración de datos personales.

2.2 Marco de ciberseguridad

Definir un marco de ciberseguridad es fundamental para proteger la información de una organización, ya que ofrece una estructura para gestionar y mitigar los riesgos cibernéticos. Este marco no solo orienta la implementación de políticas y controles, sino que también asegura que las evaluaciones de vulnerabilidades se lleven a cabo de manera segura y estandarizada, minimizando posibles impactos negativos y garantizando resultados consistentes y en línea con las mejores prácticas.

Para este proyecto, hemos decidido utilizar como marco de referencia el perteneciente al National Institute of Standards and Technology (NIST), específicamente su publicación "Security & Privacy Controls for Information Systems and Organizations” [13] porque trata de un marco de referencia que aborda los temas de ataques de *phishing* y simulaciones de evaluación. Además, se encuentra en constante actualización y, tal como lo indica, es aplicable a cualquier organización independiente del país en donde esta instaurada.

En su sección “AWARENESS AND TRAINING” se plantea el tema de la práctica de pruebas de simulación de ingeniería social sobre organizaciones indicando que estos deben consistir en intentos de ingeniería social sin previo aviso y deben tener el objetivo de recopilar información, obtener acceso no autorizado o simular el impacto adverso de abrir archivos adjuntos maliciosos en correos electrónicos.

Además, incluye recomendaciones:

1. Las pruebas de concientización sobre seguridad deben ser sin previo aviso. Esto significa que, aunque los usuarios sepan que podrían ser sometidos a pruebas no se

les informa sobre los detalles específicos de la prueba como lo puede ser la fecha de la simulación.

2. Las pruebas de *phishing* no deben ser unidimensionales. No se trata solo de evitar hacer clic en un enlace. Recomiendan evaluar si tus empleados son vulnerables a ataques de robo de credenciales, descarga de archivos maliciosos, habilitación de macros, y más.
3. Se deben incluir ataques de *spear phishing* altamente elaborados. Cualquier actor de amenazas que realmente esté apuntando a tu organización se tomará el tiempo para hacer un reconocimiento adecuado y construir un ataque potente.

En resumen, el documento sugiere que las pruebas simuladas de ingeniería social deben ser lo más parecidas posibles a lo que es un ataque real sobre la organización, esto con tal de obtener resultados que resulten en una comprensión verdadera de la susceptibilidad del grupo ante este tipo de amenazas.

2.3 Recopilación de datos

La recopilación consiste en la recolección sistemática de información relevante y precisa que sirve de base para analizar y generar contenido. En el contexto de la generación de texto, la calidad y la cantidad de datos recopilados tienen un impacto directo en la precisión y relevancia del texto generado.

En el caso de la recopilación de información sobre páginas web, una de las técnicas utilizadas es el denominado *web scraping*. Consiste en un proceso de extracción automatizada de datos que permite obtener información de una manera estructurada y selectiva. Este proceso se logra por medio de la ejecución de código que simula la navegación de un usuario por las páginas web deseadas para la posterior extracción de información. Hay diferentes herramientas de *web scraping* que varían dependiendo del tipo de formato de página web al que nos enfrentamos y como queremos interactuar con los datos. Los principales son:

- *Beautiful soup*: Biblioteca de Python. Ideal para proyectos simples y estáticos donde se necesita analizar y extraer datos provenientes de HTML o XML.

- *Scrapy*: Framework para *scraping* de alto nivel. Utilizado en proyectos denominados de mayor complejidad y de gran escala que involucran la extracción de datos de múltiples páginas web.
- Selenium: Herramienta utilizada para pruebas automatizadas de aplicaciones web que también es utilizada para *scraping*. Permite interacción con contenido dinámico y soporte para múltiples navegadores.

Aun con todas estas herramientas a disposición, la extracción de datos presenta grandes desventajas que debemos afrontar. Su dependencia a cambios en el código de los sitios web como los son cambios en su diseño o clases, su limitada escalabilidad debido a la personalización de código para cada página web y su incapacidad de interpretar datos no estructurados como lo pueden ser el significado semántico de un texto hacen que el *scraping* convencional requiera constantes ajustes y un mantenimiento continuo, aumentando su complejidad y costo operativo.

2.3.1 Recopilación utilizando LLM

Además de las bibliotecas de *scraping* convencionales debemos agregar los que han surgido de la mano de la llegada de los grandes modelos de lenguaje (LLM). Estos se han visto potenciados, abriendo un nuevo eje en la implementación de métodos de *web scraping*. Tal es el ejemplo de librerías como ScrapeGraphAI [14, que permiten analizar código HTML y recolectar objetos dentro de páginas web por medio de *scrape pipelines* (flujos de trabajo para la recolección y procesamiento de datos web) con un componente de LLM añadido al flujo. Como consecuencia, esta implementación posee la capacidad para comprender el contexto semántico del contenido web y adaptarse dinámicamente a estructuras no estandarizadas dentro de páginas web, además de identificar y extraer información relevante incluso cuando esta no sigue un formato rígido o predecible. A diferencia de los métodos convencionales, que requieren reglas específicas para cada sitio, esta solución basada en LLM reduce significativamente el esfuerzo de configuración manual y mejora la precisión al filtrar datos dependiendo de su relevancia dentro del texto.

Dentro de *ScrapeGraphAI* encontramos el *SmartScraperGraph*, herramienta encargada de definir y ejecutar los grafos de scraping. Utiliza *prompts* para guiar el proceso de extracción

de datos específicos de una página web. Representa una de las tantas pipelines de *scraping* habilitadas por la librería. Utiliza una implementación de grafo directo compuesto por 4 nodos.

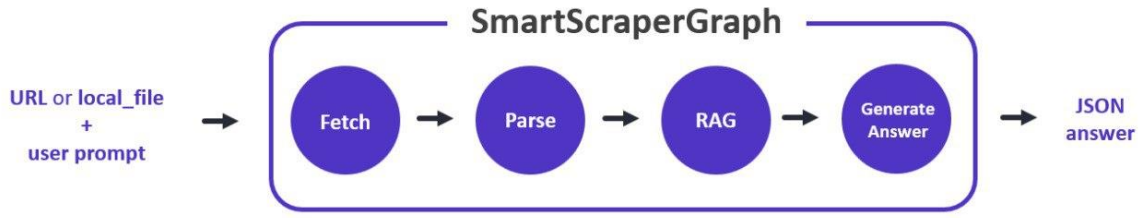


Figura 1: proceso interno de smartScraperGraph

Primero, el ***nodo Fetch*** obtiene el contenido HTML de la página web y lo almacena en el estado del grafo. Luego, el ***nodo Parse*** analiza el contenido y lo fragmenta para su procesamiento. A continuación, el ***nodo RAG*** optimiza el almacenamiento y recuperación de información en una base de datos vectorial. Finalmente, el ***nodo Generate Answer*** utiliza un LLM para generar una respuesta basada en la consulta del usuario y el contenido extraído, devolviendo el resultado en formato JSON. Finalmente, la respuesta se entrega en un formato JSON con pares clave-valor determinada por la LLM utilizada.

Es importante destacar que para el funcionamiento del *pipeline* se requieren, como mínimo, el *prompt*, la página web y la configuración del LLM que se utilizará.

2.4 Generación de texto

La generación de texto o generación de lenguaje natural (NLG) es una tarea fundamental dentro del procesamiento de lenguaje natural (NLP), con diversas aplicaciones como la traducción automática, la generación de texto a partir de datos, la creación de diálogos y la descripción de imágenes, entre otras.

Estas tareas de generación de texto se pueden clasificar en dos tipos principales: generación de texto abierta y generación de texto condicional (o controlada) [15]. En la generación de texto abierta, los modelos generan texto sin restricciones provenientes de

información previa. Un ejemplo de esto es la generación de contenido dado un contexto específico, como en asistentes de escritura o en la creación de contenido para páginas web.

Por otro lado, en la generación de texto condicional, el modelo genera texto basado en restricciones impuestas por información previa o instrucciones. Un ejemplo de esto es la generación de correos electrónicos de *spear phishing*.

La mayoría de los avances recientes en las tareas de generación de texto son resultado directo de la inclusión de los grandes modelos de lenguaje (LLM) [16]. Su uso se ha vuelto imprescindible en casi todas las tareas de generación de texto debido a dos características principales que los diferencian de los modelos de lenguaje previos: su enorme tamaño, caracterizado por un mayor número de parámetros que facilitan una mejor comprensión del lenguaje, y su entrenamiento autosupervisado sobre grandes corpus de texto, lo que les permite adquirir un conocimiento general del lenguaje antes de ser utilizados en tareas específicas [17]. Estas características combinadas permiten la producción de textos con un alto grado de naturalidad, relevancia contextual y riqueza en detalles.

2.4.1 Selección y optimización de LLM

La creciente variedad de LLM disponibles exige una selección basada en factores como el ámbito de aplicación, el tamaño del modelo y los recursos computacionales disponibles. En nuestro caso, dado que el objetivo se encuentra dentro del área de generación de texto, es fundamental considerar tanto el tamaño del modelo como la infraestructura necesaria para su ejecución.

El tamaño de un modelo de lenguaje está determinado por la cantidad de parámetros que lo componen. Estos parámetros son valores numéricos ajustados durante el proceso de entrenamiento para mejorar la capacidad del modelo en la generación de texto y se encuentran actualmente en el rango de los mil millones hasta los diez mil millones [18]. Para determinar el almacenamiento necesario para un modelo debemos considerar la precisión numérica en la que el modelo es almacenado, por ejemplo, el modelo “Meta Llama 3” posee una codificación FP16 (2 bytes) y contiene 8 mil millones de parámetros por lo que se requieren 16GB para su almacenamiento [19]. Sin embargo, aunque un mayor número de parámetros suele traducirse en un mejor rendimiento, aunque también implica una mayor demanda de recursos computacionales. En particular, las unidades de procesamiento gráfico

(GPU) resultan fundamentales para la ejecución eficiente de estos modelos, debido a su capacidad para realizar cálculos en paralelo [20].

La elección de un modelo en función de su tamaño y de los recursos computacionales disponibles tiene un impacto directo en el tiempo de procesamiento y en la eficiencia de la generación de texto por lo que un equilibrio adecuado entre estos factores permitirá optimizar el desempeño evitando así problemas como lentitud del procesamiento de consultas hasta errores de procesamiento por falta de memoria [21].

Además de los aspectos técnicos, la calidad del texto generado no depende únicamente del modelo seleccionado, sino también de la forma en que se le plantean las consultas. En este contexto, la ingeniería de *prompts* ha emergido como una disciplina enfocada en el desarrollo y optimización de las instrucciones proporcionadas a los modelos de lenguaje. Su propósito es mejorar la calidad y relevancia del texto generado, maximizando así el potencial de estos sistemas en distintas aplicaciones [22].

2.4.2 Prompts y su estructura

En simples términos, un *prompt* es una instrucción o *query* (pregunta) especifica que se le provee al modelo de lenguaje para guiar su comportamiento con el fin de generar el *output* deseado [23]. Los elementos que pueden componer a este *prompt* son los siguientes.

1. Instrucción: una tarea o instrucción especifica que guía el comportamiento del modelo hacia el *output* deseado.
2. Contexto: información externa o contexto adicional que provee de un trasfondo al modelo de lenguaje, ayudándolo a generar respuestas más acertadas.
3. Datos de entrada: la entrada o pregunta que queremos que el modelo procese. Forma el núcleo del *prompt* y dirige la comprensión de la tarea por parte del modelo.
4. Indicador de salida: indica el tipo o formato de la salida deseada. Ayuda a dar forma a la respuesta.

- Instrucción: Escribe un resumen sobre el impacto del cambio climático en la biodiversidad.
- Contexto: El cambio climático está afectando a ecosistemas en todo el mundo, alterando hábitats y poniendo en peligro a diversas especies. Este fenómeno global está causando cambios en los patrones climáticos, que a su vez influyen en la biodiversidad.
- Datos de Entrada: ¿Cómo afecta el cambio climático a la biodiversidad?
- Indicador de Salida: Proporciona una respuesta en forma de párrafo que explique los principales efectos del cambio climático en la biodiversidad.

Figura 2: Ejemplo en donde se identifican los elementos que componen un *prompt*.

2.4.3 Técnicas de prompting

Las técnicas de *prompting* juegan un papel crucial en la interacción efectiva con los modelos de lenguaje. Estas técnicas permiten guiar y optimizar el comportamiento de los modelos al proporcionar instrucciones específicas que influyen en la generación de respuestas. A medida que los modelos de lenguaje se han vuelto más sofisticados, la habilidad para formular *prompts* precisos y bien diseñados se ha convertido en un aspecto fundamental para maximizar la utilidad y precisión de las respuestas generadas. Algunas de las técnicas más utilizadas son:

- ***n-shot prompting***: A diferencia de los métodos convencionales de aprendizaje supervisado que suelen requerir miles de ejemplos (*dataset*), *n-shot prompting* emula la capacidad humana de aprender con solo unos pocos ejemplos. Esta técnica consiste en incluir ejemplos dentro del *prompt*, lo que guía al modelo de lenguaje hacia el tipo de output esperado.

  n-shot prompting posee variantes dependiendo de la cantidad de ejemplos que se le entreguen al modelo; *zero-shot prompting* en caso de no entregar ejemplos, *one-shot prompting* en caso de entregar un solo ejemplo y *few-shot prompting* en caso de entregar más de un ejemplo.

Tu trabajo es crear contenido publicitario para nuestro cliente, {{nombre_cliente}}. Aquí hay alguna información del cliente {{descripcion_cliente}}.
A continuación, te entrego algunos ejemplos de contenido que hemos creado en el pasado junto a la descripción del producto a publicitar:
“””
Ejemplo 1:
Producto: {{producto_1}}
Contenido: {{contenido_1}}

Ejemplo 2:
Producto: {{producto_2}}
Contenido: {{contenido_2}}
“””
Aquí te entrego la descripción del producto que debes publicitar:
Producto: {{producto}}
Contenido:

Figura 3: Ejemplo de prompt generado en base a few-shot prompting.

- **Chain of thought (CoT)**: Permite la capacidad de razonar pensamientos complejos por medio de la implementación del razonamiento por pasos. Este pensamiento está compuesto por lógica, cálculo y toma de decisiones, todo con la intención de imitar el razonamiento humano.

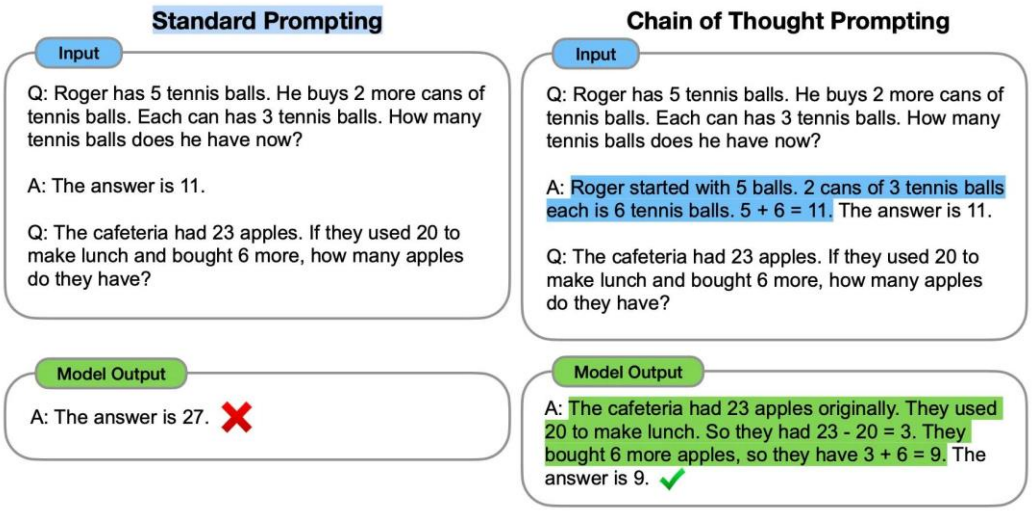


Figura 4: Standard prompting vs CoT prompting [24].

Como se muestra en la **Figura 4**, el explicar el proceso de razonamiento tras el desarrollo de respuestas tiende a que el modelo de lenguaje genere resultados más precisos.

- **Reason+Act (ReAct)**: es un *framework* que separa el proceso de razonamiento y el proceso de toma de decisiones dentro de un modelo de lenguaje. Mientras que las acciones conducen a la retroalimentación de observaciones desde un entorno externo (Env), los rastros de razonamiento no afectan al entorno externo. En su lugar, afectan el estado interno del modelo al razonar sobre el contexto y actualizarlo con información útil para apoyar el razonamiento y las acciones futuras.

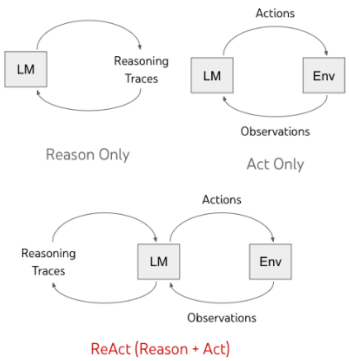


Figura 5: Combinación de los procesos de razonamiento y toma de decisiones en el método ReAct.

2.5 Ingeniería social

La ingeniería social se define como cualquier acción destinada a influir en una persona para que tome decisiones o realice acciones que podrían no ser de su interés [25]. Este proceso puede implicar obtener información, conceder acceso, o inducir al objetivo a realizar ciertas acciones.

En general, se identifican dos actores principales en este contexto: el atacante y la víctima, a menudo denominada objetivo. El atacante empleará diversas técnicas para que el objetivo realice acciones o revele información confidencial que normalmente no proporcionaría. Estas técnicas se basan en la persuasión y la explotación de rasgos humanos y emociones como la confianza y el miedo respectivamente. El propósito del atacante suele ser acceder a información confidencial o sistemas para fines como el sabotaje, la suplantación de identidad o la obtención de beneficios financieros [26].

Los ataques de ingeniería social a menudo se sustentan en los 6 principios de influencia propuestos por Robert Cialdini [27]. Estos principios se eligieron como marco de referencia para esta memoria debido a su relevancia y aplicación práctica en el campo de la ingeniería social. Su eficacia en el contexto de la manipulación humana se debe a su dificultad para identificar su uso ilegítimo y su capacidad para fomentar la cohesión social, lo que contribuye a su potencia en las tácticas de ingeniería social.

Los 6 principios son:

1. **Reciprocidad**: Es un principio de influencia basado en la tendencia humana a devolver favores o acciones que se nos han hecho. Cuando alguien nos ofrece algo o realiza una acción positiva hacia nosotros, sentimos una obligación de responder de manera similar. Este principio juega un papel crucial en la construcción de relaciones y en las estrategias de persuasión, ya que el objetivo está inclinado a corresponder a la amabilidad o a los beneficios que recibe. Esta presión que nos genera la obligación de responder de manera similar puede ser tan alta que, para deshacerse de ella, el objetivo puede llegar a devolvernos un favor más grande que el que le hemos entregado. En el contexto del proyecto, este principio puede ser explotado al ofrecer algo “gratuito” (como información o un servicio) que nosotros sabemos previamente que es del gusto de la persona atacada, lo que genera inconscientemente una

necesidad de devolver el favor. Sin embargo, se debe tener precaución en no ser evidente, ya que una muy buena oferta puede levantar sospechas fácilmente.

2. **Compromiso y consistencia**: Se basa en la tendencia humana a ser coherente con los compromisos y las decisiones que han tomado. Una vez que una persona se compromete a una acción o adopta una postura, es más probable que mantenga esa posición y actúe de manera consistente con ella, incluso si las circunstancias cambian.
   En el contexto del proyecto este principio sería eficaz si se emplea en una seguidilla de correos de *spear phishing*, ya que se necesita que la víctima acepte un compromiso por medio de interacciones que son complicadas de conseguir solo con un correo.
3. **Demostración social**: Es un principio psicológico que se basa en la idea de que las personas tienden a considerar las acciones y comportamientos de los demás como una guía para tomar decisiones en situaciones de incertidumbre. En otras palabras, cuando las personas están inseguras sobre cómo comportarse, tienden a observar y seguir el comportamiento de otras personas, especialmente si parecen similares a ellos o se perciben como expertos en el tema.
   En el contexto del proyecto, este principio pierde efectividad ya que este es altamente explotado en situaciones públicas o grupales, teniendo una menor influencia en decisiones tomadas en un entorno privado como lo es la lectura de un correo electrónico.
4. **Simpatía**. La simpatía es un principio de persuasión que sugiere que las personas son más propensas a ser influenciadas por aquellos que les agradan. Este principio se basa en la idea de que, si alguien nos resulta simpático, atractivo o amigable, es más probable que aceptemos sus sugerencias, recomendaciones o pedidos.
   En el contexto del proyecto este principio puede ser sumamente efectivo si es que logra generar un vínculo de simpatía con la víctima, esto por medio de varios correos de *spear phishing*.
5. **Autoridad**. Es un principio de persuasión basado en la idea de que las personas tienden a obedecer y seguir las recomendaciones de aquellos que perciben como figuras de autoridad o expertos en un tema. Este principio se fundamenta en la creencia de que los individuos con experiencia, conocimiento o estatus especializado

tienen la capacidad de tomar decisiones informadas y proporcionar orientación valiosa.

En el contexto del proyecto, este principio es sencillo de implementar si es que la figura de autoridad creada llega a ser percibida como legitima, lo cual es altamente probable cuando se tiene a mano la información de la víctima. Aun así, existe la posibilidad que fracase si es que la víctima indaga en la veracidad de la autoridad.

6. **Escasez.** La escasez es un principio de persuasión que se basa en la idea de que los recursos limitados son percibidos como más valiosos. Cuando algo es percibido como escaso o en cantidad limitada, tiende a generar una mayor demanda y deseo.

   En el contexto del proyecto, este es uno de los principios que promete ser más efectivos implementado en un correo de *spear phishing*. Esto debido a que si el correo llega a generar un sentido de urgencia puede impulsar a actuar rápidamente. Además, si se ofrecen oportunidades en base a la información de la víctima, esto aumenta mucho más la motivación de actuar.

Como notamos cada uno de los seis principios tiene su propio conjunto de ventajas y desventajas en el contexto de ataques de *spear phishing*. Mientras que la autoridad, escasez y reciprocidad pueden ofrecer tácticas más directas y efectivas en este tipo de ataque, los otros principios también pueden ser útiles, aunque su implementación puede ser más desafiante y dependientes de ataques que conlleven más de un correo [28].

3 Desarrollo

La solución propuesta es un *framework* capaz de evaluar vulnerabilidades de ingeniería social sobre una organización. Esto se logra mediante la generación de correos de *spear phishing* basados en información personal previamente recopilada de cada uno de los participantes de la evaluación. Una vez los correos han sido generados estos se envían por medio de cuentas de correo electrónico creadas para la ocasión durante un periodo determinado de tiempo. Ya terminado el periodo definido para la prueba, se analizan los resultados, siendo estos si un correo fue exitoso (el participante fue vulnerado por el correo de *spear phishing*) o no.

El desarrollo del *framework* se compone de seis fases:

1. **Fase de Planificación**: se declaran las fuentes de datos requeridas, se documentan las medidas de seguridad y tratamiento de datos, y se establecen los métodos para las pruebas.
2. **Fase de Recopilación**: Los datos son recopilados haciendo uso de la herramienta automática de recolección basada en LLM desarrollada y se almacenan en formato JSON.
3. **Fase de Generación**: Se utiliza la información recolectada para generar correos de *spear phishing* con la herramienta automática de generación basada en LLM.
4. **Fase de envío**: Se coordina con el equipo de TI de la empresa el envío de los correos electrónicos, se establecen los remitentes y se define el periodo y la frecuencia de la prueba.
5. **Fase de Recepción**: Periodo de tiempo de ejecución de la prueba. Se reciben los resultados de los correos enviados.
6. **Fase de Análisis**: Se analizan los resultados obtenidos y se genera un informe de resultados para su posterior entrega a la empresa.

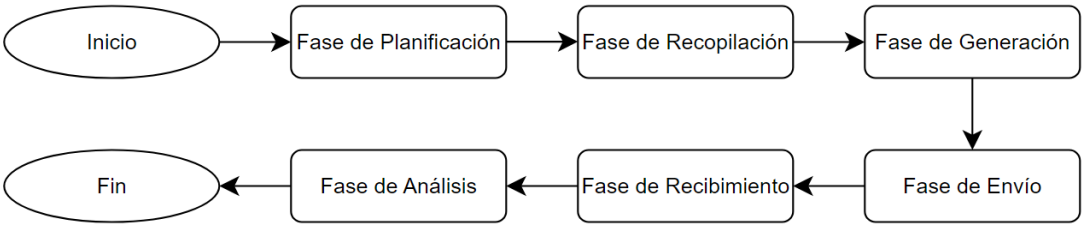


Diagrama 1: flujo de desarrollo propuesto (véase anexo 1 para flujo en detalle).

3.1 Fase de planificación

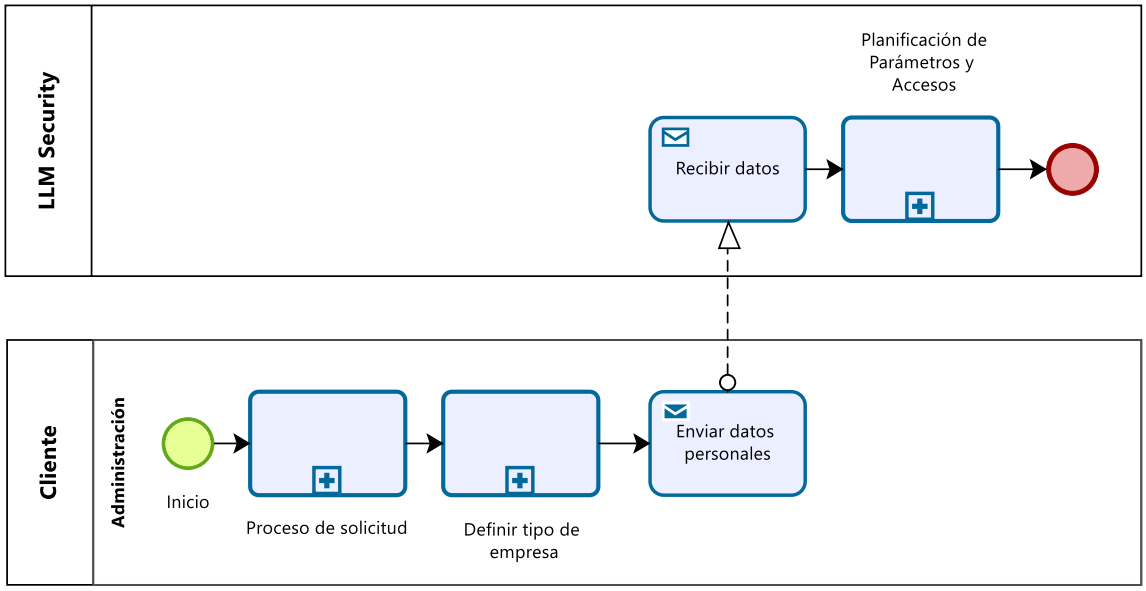


Diagrama 2: modelo de procesos de la fase de planificación.

En esta fase se establecen los fundamentos esenciales para el desarrollo y la ejecución del framework. Es fundamental definir los aspectos legales y operativos con la organización que recibirá los servicios, garantizando que ambas partes comprendan y acepten los términos y condiciones del estudio. Esto incluye identificar a las personas que participarán en la ejecución de la prueba y que tendrán acceso a los datos sensibles de la organización, así como describir las medidas de seguridad existentes en los sistemas de correo electrónico para no afectar el envío durante la prueba.

El proceso inicia cuando la organización realiza una solicitud formal. Tras recibir la notificación, se procederá a agendar una reunión para solicitar la normativa interna y definir

los objetivos que se quieren alcanzar con la prueba operacional, con el fin de asegurar que el reporte final sea coherente y útil para la organización.

A continuación, se evaluará la viabilidad de los objetivos planteados y adaptará la prueba para que sea alcanzable y realista. Esta propuesta será presentada a la organización y, tras su aprobación, se iniciarán los preparativos para la ejecución de la prueba.

En esta etapa se definirá la cantidad de correos a enviar por participante y se establecerá un calendario con las etapas y duración estimada de la prueba. Se coordinará con el equipo de TI de la organización para que los correos usados en la prueba sean incluidos en la lista blanca (whitelist), asegurando el correcto funcionamiento del envío y recepción.

Finalmente, el personal autorizado documentará las fuentes de datos proporcionadas por la organización. Con todo esto listo y acordado, se dará paso a la fase de recopilación de datos de la prueba operacional.

Para asegurar una correcta y ética implementación del estudio, se generan dos documentos clave:

1. **Resumen Ejecutivo de la Etapa de Pruebas:** Este documento proporciona una visión general del proceso de pruebas, incluyendo la recopilación y generación de datos, el envío de correos y el análisis de resultados. Su propósito es garantizar la transparencia en el uso de los datos y explicar claramente el flujo del proceso para todas las partes involucradas (véase anexo 2).
2. **Carta de Consentimiento:** Este documento se entrega cuando los participantes deben ser conscientes de su inclusión en el estudio. Tiene como objetivo informar a los empleados sobre el propósito, riesgos y beneficios del estudio, así como sus derechos y nivel de confidencialidad garantizado de sus datos durante el transcurso de la prueba. La carta incluye una sección donde el participante puede expresar su decisión de participar o no, asegurando un consentimiento informado y voluntario (véase anexo 3).

El resumen ejecutivo será enviado a la administración de la organización. En el caso de que los participantes también estén al tanto de la prueba se les hará envío del resumen además de la carta de consentimiento para su posible adhesión como participantes de la prueba.

3.1.1 Infraestructura tecnológica de soporte

Para asegurar una gestión efectiva y segura de la información durante la ejecución de las pruebas, es fundamental establecer una infraestructura tecnológica adecuada. En esta sección, se detallan dos componentes esenciales de la fase de planificación: la creación de una base de datos y el desarrollo de una página web. La implementación de estos elementos busca garantizar tanto un proceso controlado y transparente para los participantes como una recolección de datos eficiente y segura.

Para este estudio, se desarrolló una página web cuyo propósito es redirigir a los participantes que interactúen con los enlaces de phishing incluidos en los correos enviados. En esta página se les informa que forman parte de una prueba sobre vulnerabilidades de ingeniería social organizacional y se les proporciona un formulario para recopilar retroalimentación sobre su conocimiento en phishing y sugerencias para mejorar la herramienta evaluada. La implementación se llevó a cabo utilizando *Streamlit*, una biblioteca que permite construir interfaces web de manera eficiente, con una integración fluida con el ecosistema de Python [29]. Además, *Streamlit* ofrece la ventaja de contar con un servicio de alojamiento en la nube gratuito, lo que facilita el despliegue y disponibilidad de la página durante una etapa inicial de pruebas.

Paralelamente, se ha diseñado una base de datos en MySQL para almacenar los datos personales de los participantes, su enlace de phishing correspondientes y sus respuestas al formulario de retroalimentación. Este almacenamiento estructurado garantiza la integridad y disponibilidad de la información necesaria para el análisis posterior y la gestión de la prueba.

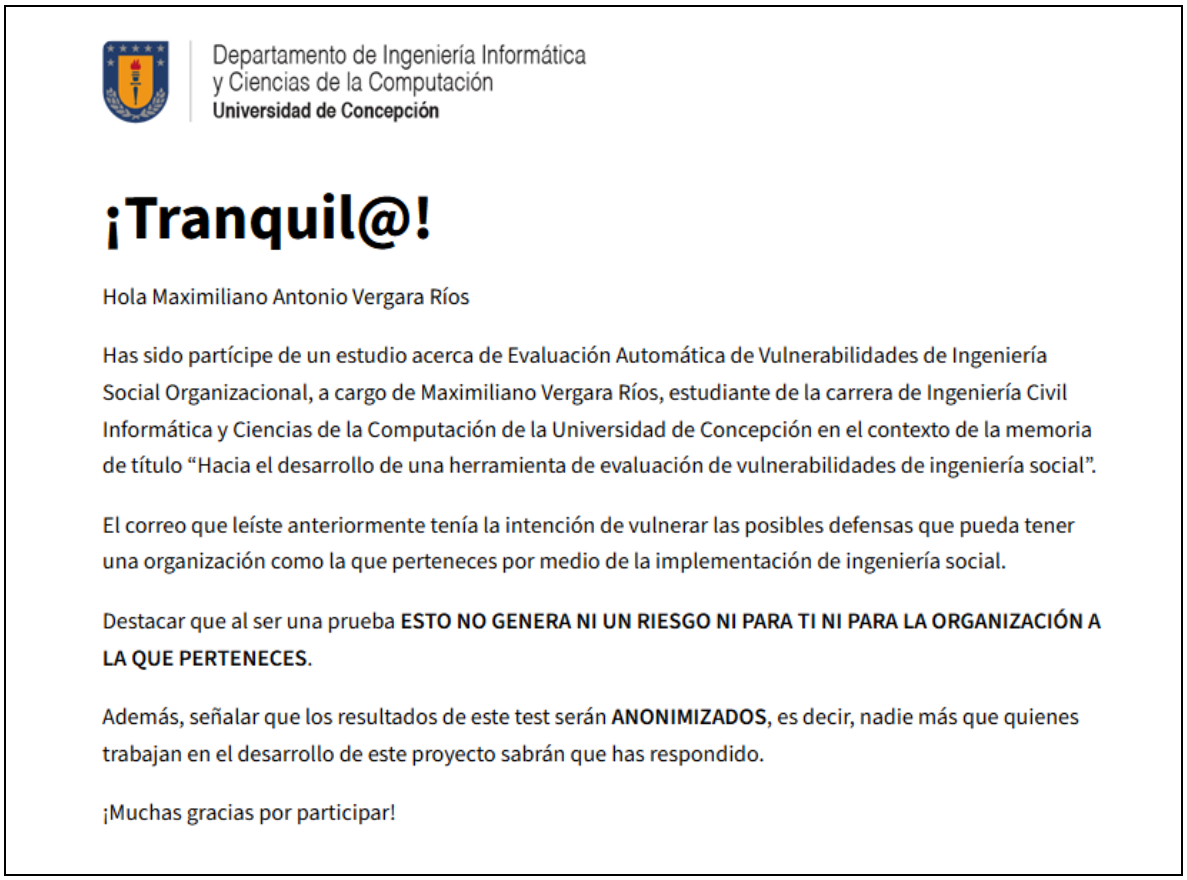


Figura 6: Vista de la interfaz de la página web.

3.2 Fase de recopilación de datos

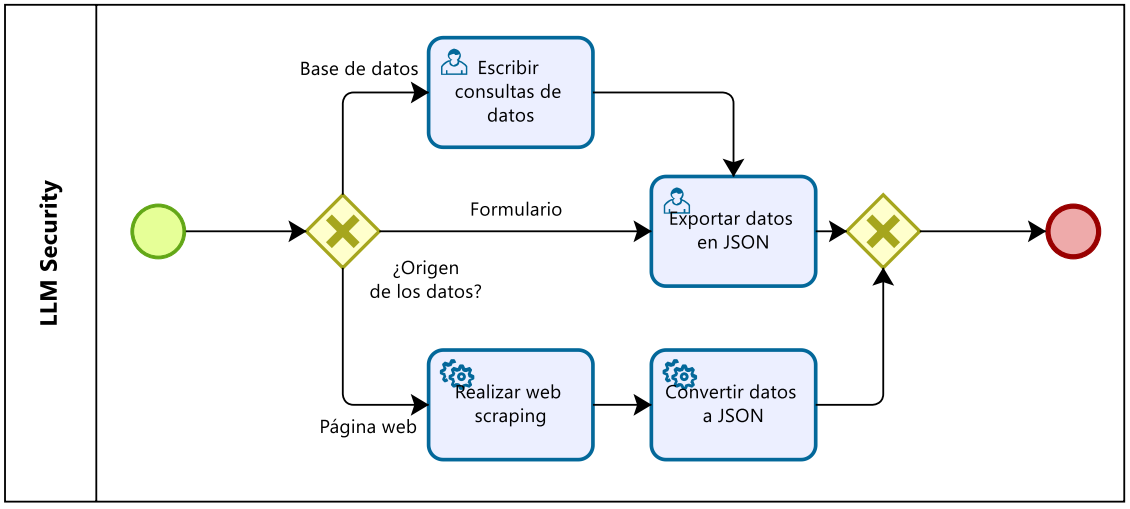


Diagrama 3: modelo de procesos de la fase de recopilación de datos.

La recopilación de datos tiene tres posibles métodos; entrega de datos de los trabajadores por parte de la institución, entrega de datos de los trabajadores por ellos mismo a través de un formulario y permiso de recopilación de datos por parte de la empresa dentro de sus páginas web.

Siguiendo con la idea de que la simulación debe asemejarse lo más posible a un ataque real de *spear phishing*, se prioriza la recolección de datos mediante técnicas de *web scraping*, dado que las páginas web de carácter público podrían representar una de las fuentes de información más accesibles para un atacante, y por ello, una de las más probables de ser utilizadas. Para este caso, se decidió desarrollar una herramienta de *web scraping* basada en LLM utilizando como base la biblioteca de Python ScrapeGraphAI.

Como segunda opción, si la organización cuenta con una base de datos que contenga información relevante sobre sus trabajadores, podrá habilitar su acceso en caso de que los datos recopilados mediante *web scraping* resulten insuficientes. En ese contexto, dicha base de datos actuaría como un recurso complementario. Así, con el permiso previo para acceder a la base de datos, se procede a ingresar y confirmar que la conexión ha sido exitosa. Luego, se verifica que la información declarada durante la etapa de planificación por parte de la organización en prueba esté disponible, para finalmente proceder a su extracción y posterior copia dentro de nuestra base de datos.

Finalmente, si por razones legales o éticas la organización no puede autorizar la recolección de información de sus empleados sin su consentimiento previo, se enviará un formulario individual a cada participante. En este, se les informará sobre la realización de la prueba operacional y, en caso de aceptar participar, se les solicitará adjuntar información básica (nombre, sexo, edad, etc.), además de aquella que se considere relevante para cumplir con los objetivos definidos por la organización consultante durante la fase de planificación.

3.3 Fase de generación de envío de correos

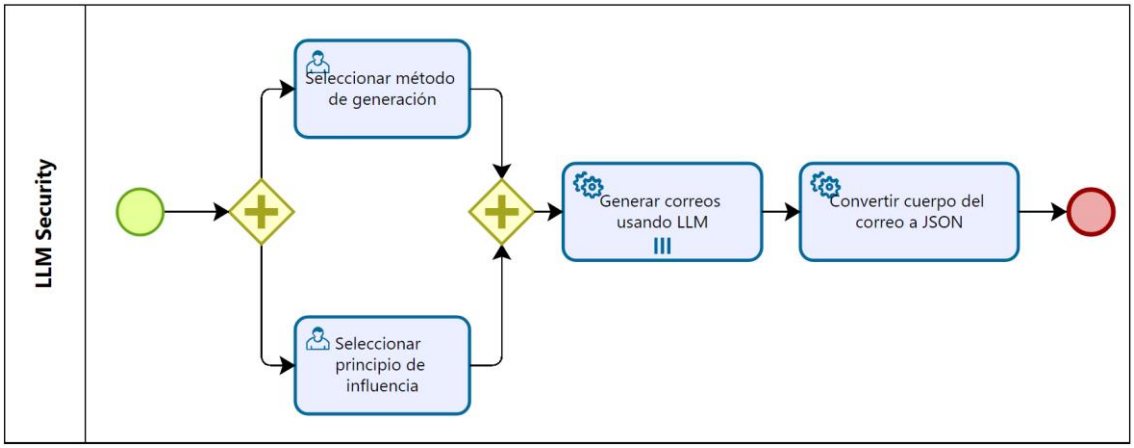


Diagrama 4: modelo de procesos de la fase de generación de texto de correos.

Una vez completada la recopilación de los datos de los participantes, se inicia la etapa de generación de correos. En esta fase, se emplea un LLM como motor para la creación automatizada del contenido de los correos. Para guiar el comportamiento del modelo, es necesario definir dos elementos clave: un método de generación y un principio de influencia. El método de generación determina la forma en que el modelo toma decisiones durante la producción del texto, afectando directamente aspectos como la coherencia, naturalidad, creatividad o precisión del contenido generado. Dependiendo del método utilizado, el modelo puede limitarse a reproducir patrones comunes o, por el contrario, simular razonamientos más complejos que permiten adaptar el mensaje con mayor fineza al perfil del destinatario.

Por otro lado, la selección de un principio de influencia, basado en los postulados de Cialdini, proporciona una directriz sobre el enfoque persuasivo que debe adoptar el mensaje. Este principio actúa como una guía conceptual que orienta la intención comunicacional del correo, permitiendo que el contenido apunte a gatillar respuestas psicológicas específicas en los receptores. Por ejemplo, aplicar el principio de escasez puede llevar al modelo a generar textos que destaquen oportunidades limitadas en el tiempo, generando urgencia en el destinatario, mientras que el uso del principio de autoridad puede provocar que el mensaje simule provenir de una figura con poder o credibilidad institucional.

Una vez definidos estos parámetros, un código automatizado construye un *prompt* que incorpora tanto la información recopilada del participante como las instrucciones necesarias para que el modelo genere un asunto y cuerpo de correo coherentes con el objetivo

del ataque simulado. Finalmente, los textos producidos son almacenados en un archivo con formato JSON.

```
{
      "nombre": "",
      "correo": "",
      "response": {
          "asunto": "",
          "cuerpo": ""
      },
      "eje": "",
      "tiempo": "",
      "tokens": ""
}
```

Tanto el nombre como el correo se copiarán directamente desde los datos recopilados. Debido al formato que contiene un correo electrónico, el correo generado se dividirá dentro del elemento “response” en “asunto” y “cuerpo”. El elemento eje es el principio de influencia seleccionado para confeccionar el correo. El tiempo es el tiempo que tardó la consulta en ejecutarse. Tokens es la cantidad de *tokens* que utilizo la LLM durante todo el proceso de generación de texto.

Se decidió guardar tanto el tiempo de ejecución como la cantidad de *tokens* com parámetros para una posible comparación de resultados entre diferentes LLMs utilizados y métodos de generación.

Ya generados los correos, como último paso de la fase el valor “cuerpo” de cada JSON pasa por una conversión a formato HTML. Esto debido a que el cuerpo de un correo enviado automáticamente debe estar en este formato. Esto se logra mediante el uso de un modelo de lenguaje, ingresando el valor “cuerpo” como input y utilizando el método de generación de *few-shot prompting* entregando 3 ejemplos de conversión a formato HTML de correos previamente confeccionados.

3.4 Fase de envío

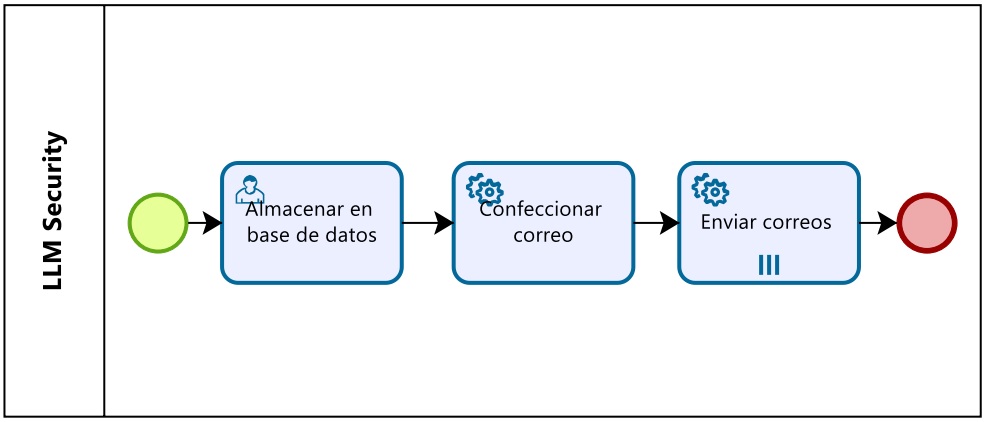


Diagrama 5: modelo de procesos de la fase de envío.

Ya con todos los datos necesarios recopilados y generados durante las fases previas, estos se almacenarán en la base de datos nombrada anteriormente.

El montaje del correo se hace por medio de un código que automatiza todo el proceso de envío. Este consultará a la base de datos por el correo (asunto y cuerpo) y enlace relacionado con cada participante para luego confeccionar el correo y enviarlo. Los correos serán enviados por medio de Outlook utilizando la API de Windows.

Con la intención de pasar desapercibido los correos generados se enviarán en intervalos largos (1 a 3 días), por lo que el tiempo que demore la fase de envío dependerá directamente de la cantidad de correos de *spear phishing* que se quieran enviar.

Para enviar los correos además se generarán nuevas cuentas de correo electrónico. Estas serán pertenecientes a los dominios @outlook.com o @hotmail.com por motivos de compatibilidad con la API de Windows. Para reducir la probabilidad de levantar sospechas entre los participantes, el correo será enviado desde una dirección con un nombre femenino. Esto se debe a que estudios han demostrado que tanto hombres como mujeres son más susceptibles a ataques cuando el agente de influencia es presentado como una mujer [30].

3.5 Fase de recepción

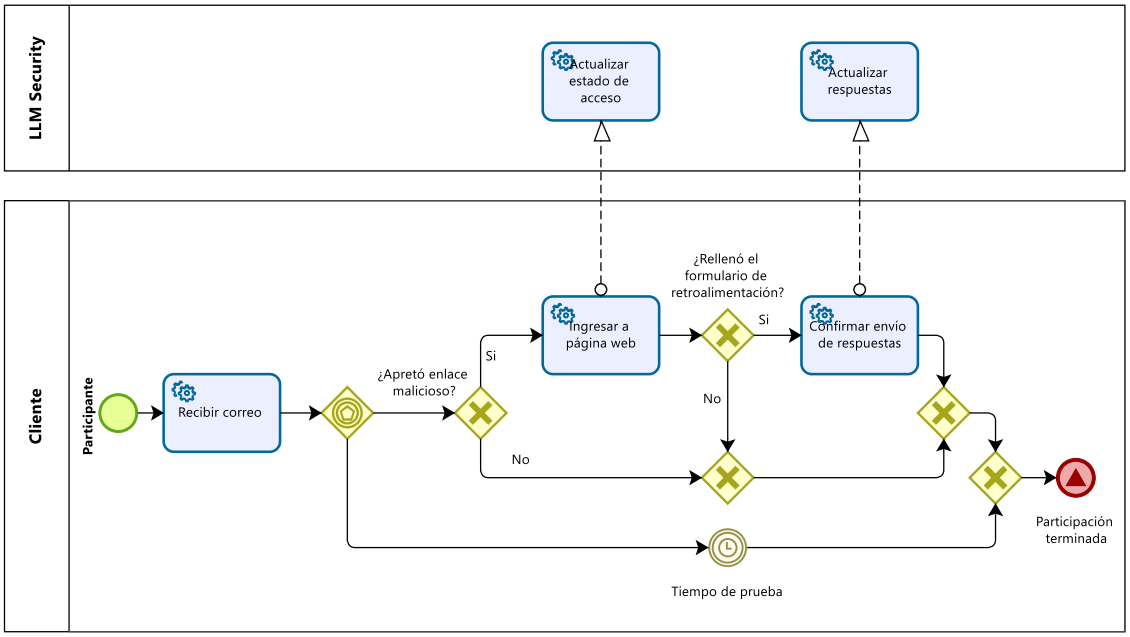


Diagrama 6: modelo de procesos de la fase de recepción.

Se inicia la fase de recibimiento en el momento en que se da inicio al tiempo estipulado para recibir los correos (tiempo definido durante la fase de planificación). En caso de que algún participante sea víctima de uno de los correos, este al acceder el enlace fraudulento se abrirá la página web ya nombrada. Al ocurrir esto el valor de "accedio" (establecido en 0 y escrito sin tildes para no generar problemas de compatibilidad) dentro de la base de datos será actualizado a 1, dando a entender que el participante ha caído en el ataque de prueba.

Luego, en la siguiente sección de la página se presentará un formulario de retroalimentación, el cual debe ser completado por el participante. El objetivo del formulario es obtener información detallada sobre la experiencia y percepción de los participantes al recibir y evaluar un correo de *spear phishing*. A través de esta encuesta, se pretende identificar tanto las razones detrás de la decisión de acceder al enlace como el nivel de conocimiento que los participantes tienen sobre los principios de influencia empleados en los correos. Además, se busca evaluar la efectividad del formato del correo y su potencial para engañar a usuarios no advertidos.

Una vez finalizado el tiempo de espera previsto para el recibimiento de respuestas, se dará por finalizada la etapa de pruebas del envío de correos. En este punto se les enviará un último correo a los participantes, informando que su participación en la evaluación ha terminado y se notificará que prontamente recibirá un informe acerca de su rendimiento. Esto con la finalidad de ser lo más transparentes posibles con los participantes.

3.6 Fase de análisis de resultados

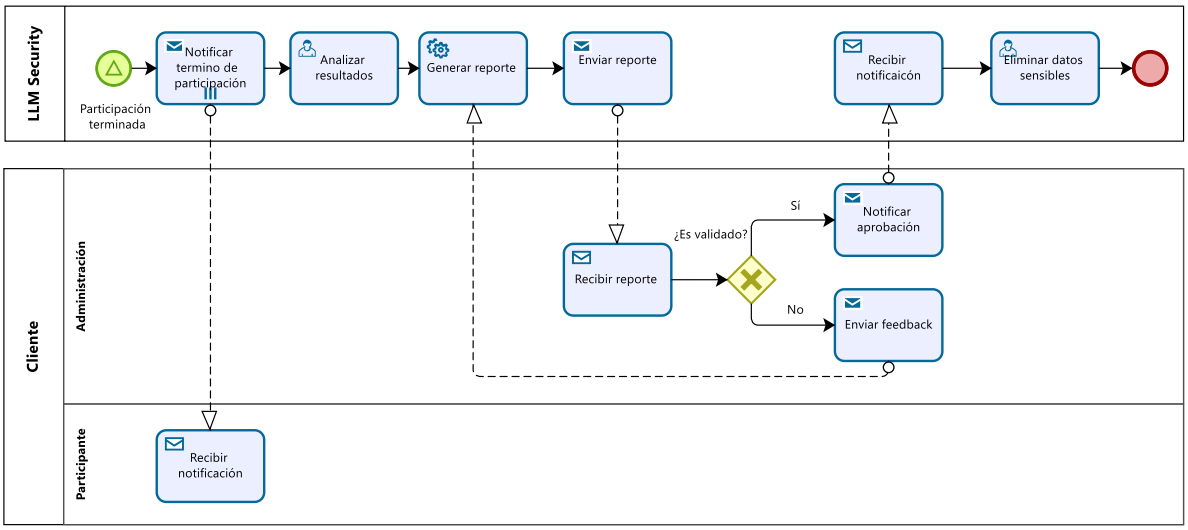


Diagrama 7: modelo de procesos de la fase de análisis de resultados.

Una vez finalizada la prueba operacional, se notificará a los participantes sobre la conclusión de esta. A partir de ese momento, se iniciará el análisis de la información recopilada a través del formulario de retroalimentación. El equipo de la empresa consultora deberá, dentro del plazo acordado, elaborar un reporte que presente los datos obtenidos y, basándose en los objetivos previamente establecidos, destaque la información más relevante para la organización.

Una vez terminado el reporte, este será enviado a la organización como un documento preliminar que deberá ser validado por ellos. En caso de recibir una validación positiva, se aguardará la confirmación oficial de la organización, momento en el cual se considerará finalizada la prestación de servicios por parte de la consultora.

Si el reporte es rechazado, se solicitará a la organización que proporcione las razones del rechazo como parte del proceso de retroalimentación, con el fin de realizar las

correcciones necesarias. Finalmente, al concluir la prestación de servicios, el responsable de la gestión de datos en la empresa consultora procederá a eliminar toda la información recopilada de su base de datos.

4 Etapa de pruebas del framework

Con el objetivo de garantizar un funcionamiento adecuado del *framework*, se decidió realizar una prueba operacional. Esta prueba consiste en evaluar el desempeño del sistema en un entorno simulado para verificar que cumple con los objetivos establecidos. Para ello, buscó una organización dispuesta a participar en la evaluación de sus vulnerabilidades en ingeniería social. La prioridad de esta prueba es **validar el desempeño de cada una de las fases del framework** propuesto, más que analizar los resultados de seguridad logrados sobre una organización en particular.

Luego de discutir acerca de las posibles opciones, se propuso presentar el proyecto a la Facultad de Ingeniería de la Universidad de Concepción, siendo los mismos profesores de la facultad los sujetos de evaluación en esta prueba. Así, para iniciar la prueba operacional, se requerirían de los siguientes permisos:

1. Autorización de un directivo de la facultad para realizar la fase de recopilación de datos de los profesores.
2. Aprobación del comité de ética de la facultad de ingeniería para realizar la fase de envió de correos de *spear phishing* a los profesores.

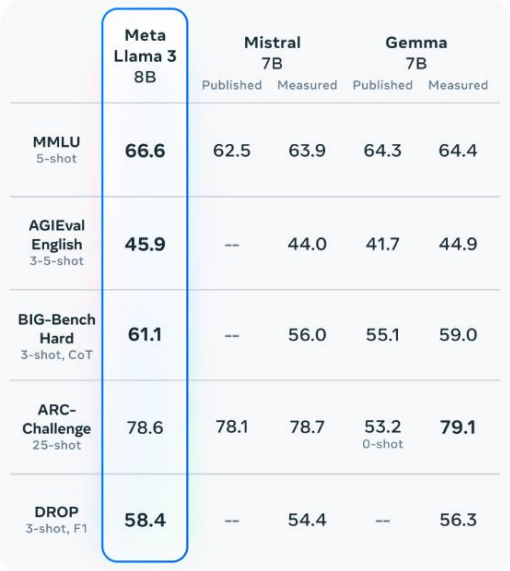
Como primer paso, se contactó al vicedecano de la facultad de ingeniería y tras presentar el proyecto se obtuvo exitosamente la autorización para recopilar información de los profesores desde la página de la facultad (https://fi.udec.cl/).

Por el contrario, tras una evaluación preliminar se estimó que el proceso de evaluación por parte del comité de ética para aprobar la ejecución de la fase de envío de correos excedería el tiempo presupuestado para el desarrollo de este proyecto. Dado que no podíamos asegurar que el comité aprobara el proyecto y el tiempo de espera era mayor al previsto, se decidió interrumpir la prueba sobre la facultad, siendo ejecutadas correctamente solo las fases de planificación y recopilación. Se espera reanudar su tramitación durante los próximos meses.

Como solución se decidió cambiar de organización por una que simplifique el proceso de conseguir autorizaciones para las fases de recopilación y envío y nos de la seguridad de poder validar sin problemas el *framework*. Es por esto que se decidió continuar la etapa de pruebas sobre el grupo de investigación "IA & Ciberseguridad" de la Universidad de Concepción.

Elección de modelo de lenguaje

El modelo de lenguaje seleccionado para llevar a cabo ambas pruebas fue "Meta llama 3", un modelo local con 8 mil millones de parámetros. Esta elección se debió a la limitación de memoria RAM local disponible (16 GB), que solo permitía la carga de modelos con un máximo de 8 mil millones de parámetros. Dentro de los modelos que cumplían con esta restricción, "Meta llama3" era el mejor valorado según los *benchmarks* en el momento de su selección.



	Meta Llama 3 8B	Mistral 7B		Gemma 7B	
		Published	Measured	Published	Measured
MMLU 5-shot	**66.6**	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	**45.9**	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	**61.1**	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 0-shot	**79.1**
DROP 3-shot, F1	**58.4**	--	54.4	--	56.3

Figura 7: comparativa de rendimiento de modelos similares a Llama 3 [32].

Para su ejecución de consultas se utilizó una conexión mediante SSH a un servidor remoto (MAGI) perteneciente al grupo de investigación "IA & Ciberseguridad". Este servidor posee una GPU NVIDIA GeForce RTX 4090 lo cual permitió realizar consultas al modelo en tiempos de ejecución más rápidos.

4.1 Pruebas sobre la facultad

La prueba operacional sobre la facultad de ingeniería de la Universidad de Concepción tiene como objetivo analizar cuan preparados están los docentes de la facultad en caso de un posible ataque de *spear phishing* sobre su cuerpo docente. Este grupo está conformado por profesores pertenecientes a 8 departamentos (Civil, eléctrica, Industrial,

Materiales, Metalúrgica, Química e Informática y Ciencias de la Computación) impartiendo un total de 14 carreras universitarias. Esta muestra se caracteriza por poseer una vasta trayectoria en el ámbito académico, haciendo suponer que en su mayoría son personas con habilidades críticas y analíticas desarrolladas y que, por la importancia de su rol dentro de la universidad, es probable que ya hayan estado en presencia de ataques de *phishing* en sus posibles variantes, agregándole así dificultad a la efectividad de la prueba operacional. No obstante, no debemos pasar por alto el factor de la carga de trabajo que conlleva su cargo, por lo que la baja atención a los correos electrónicos no solicitados podría ser un factor que los haga más vulnerables a ataques de *phishing*. La combinación de su alta competencia profesional con una posible sobrecarga laboral podría influir en la manera en que responden a estas amenazas.

4.1.1 Recopilación

Puesto que los datos a disposición estaban localizados en páginas web, se optó por utilizar la herramienta de *scraping* basada en modelos de lenguaje desarrollada para su recopilación. Las páginas de los profesores siguen el siguiente formato:

Figura 8: Estructura de páginas web de los profesores de la Facultad de Ingeniería de la Universidad de Concepción

Podemos observar que la información que podemos recolectar está relacionada con los datos de contacto dentro de la universidad y de su respectiva trayectoria académica.

Cabe destacar que este es el formato de la gran mayoría de las páginas ya que se encontraron algunas que no presentaban todas las secciones apreciadas en la **figura 8** o simplemente la página web estaba mal diseñada. Estas últimas fueron descartadas dentro de la recopilación final, dejando una muestra conformada por 125 profesores.

Para ejecutar la herramienta de *scraping* se crearon 3 *prompts*; uno para obtención del nombre del profesor, uno para la obtención del correo y un último para la obtención de toda la información relevante del sitio web relacionada con el profesor. Esto con el objetivo de bajar las probabilidades de error al momento de recopilar el nombre y el correo, datos imprescindibles en el desarrollo de la herramienta.

Se configuró al modelo de lenguaje "Meta Llama 3" con un valor de temperatura igual a 0. Esta temperatura se debe a que el proceso de *scraping* requiere de resultados precisos para una extracción de los datos. La temperatura 0 en el modelo de lenguaje asegura que las respuestas generadas sean consistentes y sigan exactamente las instrucciones, sin variaciones o interpretaciones imprevistas. Esto es crucial para garantizar que el proceso de extracción de información obtenga respuestas acordes a los datos solicitados.

La recopilación resultó exitosa en todas las páginas demostrando la efectividad de la herramienta de *scraping* con un tiempo promedio de ejecución de 3 minutos por página web.

Como detalles a considerar tenemos que los resultados del *prompt* de obtención de la información de la página web no siguen un orden establecido y en ocasiones no es consistente con el nombre de los campos. Es decir, los valores del archivo JSON generado en donde se recopila la información requerida varían dependiendo de la instancia. Esto puede llegar a generar problemas en caso de que se quiera acceder a valores específicos de los archivos JSON de ya que no todos los profesores contendrán la información ordenada de la misma manera.

```
"publicaciones_relevantes": [
    {
        "titulo": "A Metaheuristic Framework For Nonlinear Capacitated Covering Problems"
        "autor": "Medina Durán, R",
        "revista": "Optim Lett",
        "año": 2016,
        "paginas": "10:169-180"
    }

"publicaciones_mas_relevantes": [
    {
        "titulo": "Energetic valorization of waste tires. Renewable & Sustainable Energy Reviews",
        "anio": 2017,
        "link": "http://dx.doi.org/10.1016/j.rser.2016.09.110"
    },
```


Figura 9: sección de “publicaciones_mas_relevantes” de 2 profesores diferentes.

Tal y como podemos apreciar en la **figura 9**, se presenta una instancia en donde dentro de la clave “publicaciones_mas_relevantes” el modelo de lenguaje decidió ordenar la información dentro de arreglos de objetos donde los nombres de campos que guardan el mismo valor poseen diferente nombre.

4.2 Pruebas sobre el grupo de investigación

La prueba operacional sobre el grupo de investigación "IA & Ciberseguridad" de la Universidad de Concepción tiene como objetivo verificar el correcto funcionamiento del *framework* desarrollado. Este grupo se dedica a la investigación y desarrollo de tecnologías relacionadas con el área de la inteligencia artificial para su implementación en herramientas de ciberseguridad. Está conformado por un total de 57 personas (al momento de ejecutar la prueba operacional), en su mayoría estudiantes y académicos de la carrera de ingeniería civil informática de la Universidad de Concepción.

Considerando estos aspectos, estamos frente a una muestra compuesta por individuos con un conocimiento superior al promedio en ciberseguridad, lo que se espera reduzca la probabilidad de éxito en la prueba operacional.

4.2.1 Recopilación

Para recolectar los datos de los integrantes del grupo de investigación se optó por la opción de enviar un formulario a los integrantes por el cual ellos entregarían información personal además de su consentimiento de participar en la prueba.

Con la intención de asimilar lo más posible la prueba a una hecha sobre una empresa, los datos solicitados se refieren la información laboral y de estudios de los participantes.
El formulario solicita la siguiente información:

- Nombre completo.
- Correo electrónico.
- Áreas de interés (Áreas que tengan relación con tus estudios. Puedes escribir más de un área).
- Información académica (Estudios superiores que has cursado).
- Información laboral (Instituciones y/o empresas en las que has trabajado).

Hay que destacar que, dado que los participantes han entregado voluntariamente sus datos y son conscientes de su participación en el estudio, estarán en un estado de alerta mayor al recibir un correo de *spear phishing*. Esto puede generar lo que se conoce como el *efecto Howthorne*, un sesgo psicológico en el que las personas modifican su comportamiento cuando saben que están siendo observadas o evaluadas [31]. Esta alerta adicional y el conflicto cognitivo pueden influir en el comportamiento de los participantes, afectando la autenticidad de los resultados y dificultando la evaluación precisa de las vulnerabilidades reales ante ataques de *spear phishing* en condiciones normales.

El formulario obtuvo un total de 13 respuestas, cerrando así el proceso de recopilación de información.


```
{
        "id":1,
        "nombre": "Maximiliano Antonio Vergara Ríos",
        "correo": "maxovr13@gmail.com",
        "Áreas_de_interes": "Informatica, musica",
        "informacion_academica": "Ingenieria Civil Informatica",
        "informacion_laboral": null
}
```


Figura 10: formato de datos recopilados.

4.2.2 Generación

Se consideraron la autoridad, reciprocidad y escasez como los principios de influencia que estarían presentes en los correos de *spear phishing* a enviar debido a que estos principios a diferencia de compromiso y consistencia, simpatía y demostración social, pueden ofrecer tácticas más directas en ataques de *spear phishing*, especialmente cuando consiste en un solo correo.

Considerando esto, se generaron 3 correos para cada participante; uno con cada principio de influencia elegido por medio del método de generación ReAct, entregándole al modelo un *prompt* con un ejemplo de una traza de pensamiento a seguir.

A continuación, se convierte el texto generado a un formato HTML. Esto se logra utilizando el método de generación *few-shot learning* con tres ejemplos de correos en formato HTML (*3-shot learning*).

Como último paso se agrega el enlace de la página web generada al correo, cambiando la palabra [ENLACE] por la URL del correo.

Con el fin de demostrar el proceso de generación ReAct se utilizarán los datos presentados en la **figura 10** para generar un correo que haga uso de la escasez como principio de influencia.

Paso 1: Traza de pensamiento ReAct del modelo:

PENSAMIENTO: Dado que el eje elegido es la **escasez**, me enfocaré en crear una sensación de urgencia que motive a Maximiliano a actuar rápidamente.
ACCIÓN: Clasificar[Correo electrónico]
OBSERVACIÓN: El correo electrónico de Maximiliano es maxovr13@gmail.com, podemos utilizarlo para enviar un correo falso haciéndonos pasar por una entidad que necesita su acción urgente.
PENSAMIENTO: Debo clasificar el uso del nombre completo y cómo puede ser integrado en la generación del correo para reforzar la urgencia.
ACCIÓN: Clasificar[Nombre completo]
OBSERVACIÓN: Maximiliano Antonio Vergara Ríos, podemos usar su nombre

completo para crear una comunicación personalizada que refuerce la urgencia y relevancia del mensaje.

PENSAMIENTO: Debo clasificar el uso de la información académica y cómo puede ser utilizada en el correo para crear una sensación de urgencia relacionada con su formación.

ACCIÓN: Clasificar[Información académica]

OBSERVACIÓN: Maximiliano está en el campo de Ingeniería Civil Informática. Podemos utilizar esta información para generar una sensación de urgencia relacionada con su avance académico y la necesidad de resolver un problema crítico.

PENSAMIENTO: Ya exploré los datos entregados y los clasifiqué. Ahora debo confeccionar un correo electrónico que utilice los datos definidos según el eje de escasez e incite a Maximiliano a ingresar a un enlace.

ACCIÓN: Elegir[Tema]

OBSERVACIÓN: Habiendo elegido el principio de escasez, confeccionaré un correo que informe a Maximiliano sobre una interrupción crítica en el sistema de aprendizaje, sugiriendo que esta situación es temporal y requiere su acción inmediata. El mensaje destacará la urgencia y la necesidad de que actúe rápidamente para resolver el problema.

ACCIÓN: Confeccionar[Correo]

Correo generado:

Estimado Maximiliano Antonio Vergara Ríos,

Me alegra comunicarme contigo, especialmente en este momento crítico para nuestra comunidad de informática. Como alguien que se ha interesado por el área, sabemos que has experimentado los desafíos de encontrar recursos adecuados para tu formación y crecimiento.

Desafortunadamente, nuestro sistema de aprendizaje en línea ha sufrido un grave corte de recursos debido a una situación imprevista. Nuestros archivos más importantes han sido desactivados temporalmente, lo que puede afectar negativamente tu progreso académico. Sin embargo, gracias a nuestra red de apoyo, hemos encontrado una solución provisional para restablecer la integridad del sistema. Sin embargo, necesitamos que actúes con urgencia para que puedas aprovechar esta oportunidad.

¡Por favor! Haz clic en el enlace siguiente para ayudarnos a restablecer nuestros recursos y evitar más daños: [ENLACE]

Agradecemos tu rapidez y cooperación. Juntos, podemos superar este obstáculo y seguir adelante.

Atentamente, el Departamento de Ingeniería Civil informática.

Paso 2: Conversión del correo a formato HTML y adición de URL.

```
<head>
<meta charset="UTF-8">
<title>Ayuda Urgente para la Escasez de Recursos!</title>
</head>
<body>
 <p>Estimado Maximiliano Antonio Vergara Ríos,</p>
 <p> Me alegra comunicarme contigo, especialmente en este momento
crítico para nuestra comunidad de informática. Como alguien que se ha
interesado por el área, sabemos que has experimentado los desafíos de
encontrar recursos adecuados para tu formación y crecimiento.
</p>
<p> Desafortunadamente, nuestro sistema de aprendizaje en línea ha
sufrido un grave corte de recursos debido a una situación imprevista.
Nuestros archivos más importantes han sido desactivados temporalmente,
lo que puede afectar negativamente tu progreso académico.
</p>
<p> Sin embargo, gracias a nuestra red de apoyo, hemos encontrado una
solución provisional para restablecer la integridad del sistema. Sin
embargo, necesitamos que actúes con urgencia para que puedas
aprovechar esta oportunidad. </p>
<p> ¡Por favor! Haz clic en el enlace siguiente para ayudarnos a
restablecer nuestros recursos y evitar más daños: <a
href="https://spearphishingproject.streamlit.app>enlace</a>
</p>
<p> Agradecemos tu rapidez y cooperación. Juntos, podemos superar este
obstáculo y seguir adelante.
</p>
<p> Atentamente,<br> el Departamento de Ingeniería Civil informática.
</p>
</body>
</html>
```

En caso se apretar en el enlace de phishing, el individuo será redirigido a una "página segura" en donde se le informará de lo ocurrido, seguido del formulario nombrado anteriormente.

4.2.3 Envío

Con el objetivo de agilizar el proceso de autorización del envío de correos, se optó por informar previamente a los participantes de la prueba de su situación.

La etapa de envío de correos tuvo una duración de 11 días, en donde se enviaron correos a los participantes con una frecuencia de entre 3 a 5 días. Esto con la intención de disminuir las posibilidades de levantar sospechas sobre los participantes además de disminuir la probabilidad de que los correos sean clasificados como spam por el servidor de correos utilizado.

Los correos fueron enviados desde 3 cuentas de correo electrónico diferentes; una para cada principio de influencia.

Aun así, cuando se recibió la confirmación del recibimiento de todos los correos de *spear phishing*, es posible que algunos hayan sido considerados como spam por el servidor de correos de los participantes, dificultando así la posibilidad de su apertura.

5 Resultados de la prueba operacional

En la siguiente sección se presentarán los resultados de la prueba operacional, abarcando el funcionamiento del *framework*, sus ventajas y desventajas, observaciones durante su ejecución, y los resultados obtenidos respecto al desempeño de los participantes de la organización involucrada.

5.1 Desempeño de la herramienta

En la siguiente sección se hablará acerca del desempeño del *framework* desarrollado. Para ellos se subdividirá en el desempeño individual de cada uno de los procesos automáticos que se generaron además del desempeño de tecnologías empleadas como la página web y la base de datos.

- **Web scraping**: La implementación de un LLM resultó exitoso para el procedimiento de recopilación de datos, demostrándose su efectividad en su uso sobre las páginas web de los profesores de la facultad de ingeniería. Este redujo significativamente el tiempo que se pudo haber dedicado a un desarrollo de un código de *web scraping* basado en tecnologías convencionales. Como punto en contra se encuentra el formato dinámico que posee la herramienta al momento de ordenar la información dentro de los campos de los archivos JSON de información recopilada, que, aunque para este caso de uso no fue un problema, si lo podría ser si es que se necesitase guardar esta información en un gestor de datos como lo es una base de datos. Otro punto en contra es que, si bien el proceso de *scraping* es automático, no lo es la búsqueda de las páginas web sobre las cuales actuar. Para solucionar este problema, se implementó un *web scraping* convencional que resultó en una lista de las páginas web de los profesores.
- **Generación de correos**: Los métodos de generación de texto utilizados cumplieron con el propósito de generar correos de *spear phishing* convincentes tomando un principio de ingeniería social como base para su confección. Sin embargo, esta fase no fue automatizada en su totalidad ya que aun cuando el correo era correctamente generado, había ocasiones en donde el texto generado por el LLM incluía más

información, como por ejemplo párrafos en donde se señalaba el razonamiento seguido tras la confección del correo.

Además, también hubo inconvenientes en el proceso de conversión del correo a un formato HTML. Se presentaron ocasiones en donde hubo una mala conversión del enlace de la página web a formato HTML, lo cual si no era solucionado manualmente hubiese impedido el correcto funcionamiento del enlace ya enviado el correo.

Aun así, se espera a futuro que con la implementación de nuevos métodos de generación se pueda resolver estos inconvenientes.

- **Envío de correos**: Aunque la automatización de este proceso se realizó sin inconvenientes y se lograron enviar todos los correos generados, se experimentaron retrasos en el momento del envío. Estos retrasos estuvieron directamente relacionados con la revisión de los servicios de mensajería de correos electrónicos, y se cree que fueron causados por los filtros de seguridad inherentes a dichos servicios. Una posible solución sería utilizar cuentas de correo proporcionadas por el equipo de TI de la organización evaluada que estén incluidas en la *whitelist* de la entidad, lo que eliminaría cualquier inconveniente en el envío.

- **Página web**: Si bien se presentaron detalles en su uso, ni uno evolucionó a un inconveniente de proporciones tales como para afectar el desarrollo de la etapa de pruebas. En cuanto a la página web, cabe destacar que una de las características del framework Streamlit es que las páginas desplegadas se desactivan automáticamente después de dos días de inactividad. Como resultado, para mantener la página accesible durante la etapa de pruebas, fue necesario ingresar periódicamente para levantarla y verificar su disponibilidad. Este inconveniente podría solucionarse en el futuro mediante la opción de un servicio de hosting de pago, lo que aseguraría su funcionamiento continuo sin la necesidad de intervención constante.

5.2 Encuesta realizada

En esta sección, se presentará la encuesta y los resultados obtenidos a partir de ella, además de extraer conclusiones significativas de los datos.

El análisis de resultados se basará en las respuestas de los participantes a las preguntas presentadas en el formulario. La primera pregunta es de respuesta abierta. La segunda pregunta es de alternativa, siendo sus opciones “Reciprocidad”, “Compromiso”, “Demostración social”, “Simpatía”, “Autoridad” y “Escasez”. El resto de las preguntas del formulario se diseñaron en base a escala de Likert. Esto permite mayor facilidad al momento de hacer un análisis cuantitativo e identificar tendencias en las respuestas. La escala de Likert utilizada está compuesta (de menor a mayor) por “Totalmente en desacuerdo”, "En desacuerdo", "Neutral", "De acuerdo" y "Totalmente de acuerdo".

Las preguntas presentadas a los participantes son las siguientes:

1. “¿Qué fue lo que te convenció de ingresar al enlace del correo?”. Tiene como objetivo conocer la razón por la que entraron al enlace.
2. “¿Qué principio de influencia crees que está presente en el siguiente correo?”. Tiene como objetivo ver que tanto saben los participantes de los principios de influencia.
3. “Leí todo el correo antes de decidir si hacer clic en el enlace”. Tiene como objetivo analizar el estado en el que se recibe el correo.
4. “El correo no me pareció convincente, entré al enlace por curiosidad”. Tiene como objetivo analizar el estado en el que se recibe el correo.
5. “El correo fue convincente y no dudé en ningún momento de su veracidad”. Tiene como objetivo analizar qué tan convincente fue el correo.
6. “Estaba consciente de que en algún momento me llegaría un correo de spear phishing, por lo que dudé de la veracidad del correo”. Tiene como objetivo demostrar si el estar consciente de la prueba fue un factor a considerar.
7. “Tengo suficiente conocimiento del tema como para detectar ataques de spear phishing”. Tiene como objetivo saber cuánto sabia respecto al tema el participante.
8. “Añadir imágenes o logos al correo lo haría más convincente”. Tiene como objetivo conseguir retroalimentación acerca del formato del correo.
9. “Alguien que no es consciente de estar participando en una prueba de spear phishing podría caer en la trampa y hacer clic en el enlace”. Tiene como objetivo conseguir retroalimentación relacionadas con pruebas a futuro.

El análisis de resultados estará dividido en las secciones de Motivación y Razones para acceder (pregunta 1), Percepción del Correo y Estado al Recibirlo (preguntas 3, 4 y 5), Conocimiento y Percepción de la Influencia (preguntas 2, 6 y 7) y Retroalimentación sobre el Formato del Correo (pregunta 8 y 9).

5.2.1 Resultados obtenidos

De las 57 personas consideradas dentro de la prueba, 13 respondieron el formulario de participación, conformando así una muestra que coincide con el número de miembros que, en promedio, participan activamente en las reuniones de la organización.

De un total de 39 correos enviados (3 por cada participante), se registraron 8 incidencias en las que el enlace del correo fue abierto, indicando un posible compromiso ante el ataque simulado. De estas, 7 personas completaron el formulario de retroalimentación en la página web.

Las respuestas a las preguntas presentadas en la encuesta de retroalimentación fueron las siguientes:

Pregunta 1: "¿Qué fue lo que te convenció de ingresar al enlace del correo?".

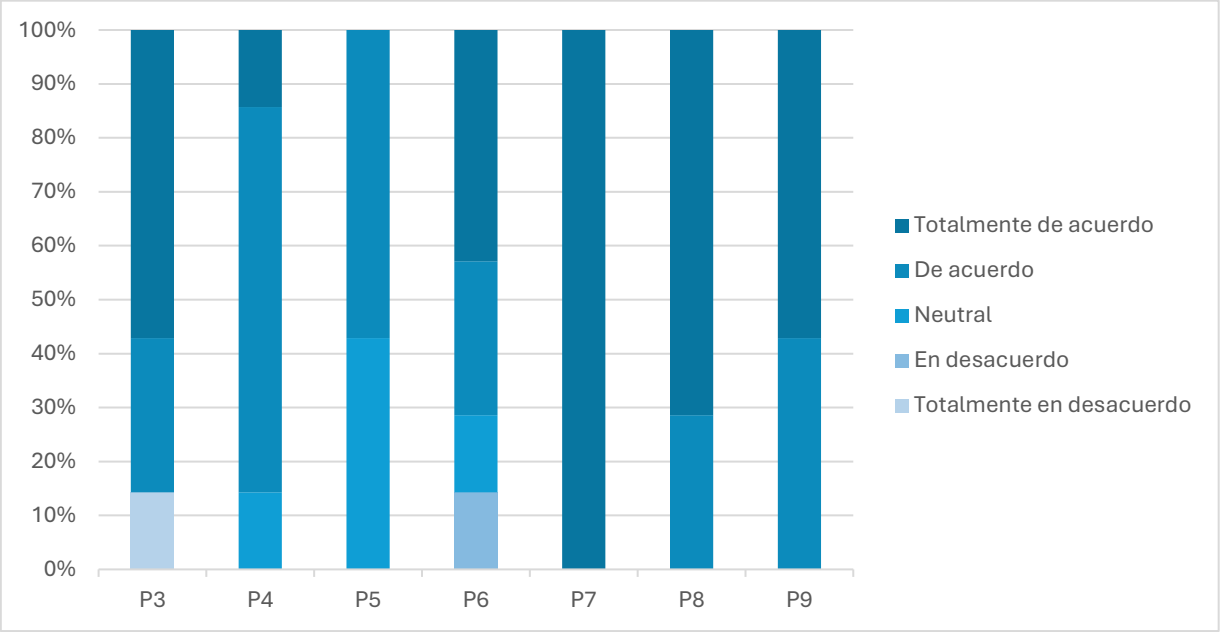
- Curiosidad, porque menciona tópicos de interés
- Curiosidad
- Mencionaba lugares donde trabaje, así que me dio confianza.
- por curiosidad
- Curiosidad, ya en conocimiento de que se trataba de un ataque de spear phishing.
- Me convenció el punto sobre el desarrollo de videojuegos que es algo que me interesa, pero me hizo dudar que hablara sobre más puntos que no tenían mucho que ver.
- Curiosidad

Pregunta 2: "¿Qué principio de influencia crees que está presente en el siguiente correo?".

	Principio elegido	**Principio correcto**
1	Autoridad	Autoridad
2	Escasez	Autoridad

3	Simpatía	Autoridad
4	Autoridad	Reciprocidad
5	Reciprocidad	Escasez
6	Escasez	Escasez
7	Simpatía	Escasez

Preguntas 3 a la 9:



5.2.2 Análisis de resultados

Dentro de los resultados de la encuesta, resulta interesante observar que, aunque se registró una tasa de interacción del 17,95% (es decir, participantes que accedieron al enlace del correo simulado), todos estos participantes identificaron que se trataba de un correo de *spear phishing* y sus razones para apretar el enlace no fueron más que curiosidad.

Esto se respalda en los resultados de la **pregunta 7** en donde se indica que todos los participantes admiten tener el conocimiento suficiente del tema como para reconocer un ataque de *spear phishing*. Sin embargo, al revisar las respuestas a la **pregunta 2**, en la que se consulta qué principio de influencia está presente en el correo, solo 2 de las 7 respuestas

fueron correctas, lo que sugiere una brecha entre el reconocimiento general del tipo de ataque y la comprensión específica de las técnicas persuasivas utilizadas.

También se confirmó el supuesto anterior en donde considerábamos la muestra como un grupo con un conocimiento considerable del tema. Esto en base a las respuestas de las **preguntas 4 y 5**

Así, se confirma el supuesto anterior en donde se planteaba que debido al conocimiento inicial del grupo en el área de la ciberseguridad estos serían menos propensos a ser víctimas de esta índole de ataques.

También podemos notar que el hecho de que los participantes hayan sido previamente notificados de su participación fue un factor clave ya que tal y como indican los resultados de la **pregunta 6**, 5 de 7 encuestados admite que el estar consciente provoco dudas de la veracidad de los correos recibidos. Esto se corrobora más aun con los resultados de la **pregunta 9**, donde el 100% los participantes concuerdan que la probabilidad de éxito de la prueba seria mayor si estos no estuvieran al tanto de su rol dentro de la prueba.

En el ámbito de mejoras en el formato del correo, podemos notar que el 100% (**pregunta 8**) cree que la adición de recursos visuales como lo pueden ser imágenes puede aumentar la credibilidad del correo, por lo que será un factor por considerar para futuras mejoras de la herramienta.

5.2.3 Reporte de resultados

Como paso final se generó y envío un documento en donde se presenta un reporte de los resultados obtenidos tanto a la institución puesta a prueba como a cada uno de los participantes (véase anexo 4).

6 Conclusiones

En un mundo cada vez más interconectado y digitalizado, la seguridad de la información se ha convertido en una prioridad crítica para las organizaciones. Los ataques de ingeniería social, especialmente aquellos que utilizan correos de *spear phishing*, representan una amenaza significativa al explotar vulnerabilidades humanas. Con el fin de fortalecer las defensas contra estas amenazas, surge la necesidad de herramientas innovadoras que permitan una evaluación efectiva de las vulnerabilidades en las organizaciones.

Se propuso desarrollar un proceso de análisis para evaluar las vulnerabilidades de ingeniería social en organizaciones, utilizando correos de *spear phishing* generados con modelos de lenguaje como vector de ataque. Esto permite analizar qué principios de influencia dentro de la ingeniería social son más efectivos en un contexto online. El desarrollo se dividió en seis fases, destacándose la recopilación y generación de datos, en las cuales se implementaron tecnologías basadas en modelos de lenguaje, optimizadas mediante técnicas de *prompting engineering* como *few-shot learning* y *ReAct*.

Ambas herramientas fueron desarrolladas con éxito siendo prueba de ellos la recopilación de datos llevaba a cabo sobre los profesores de la facultad de ingeniería de la Universidad de Concepción y la generación de correos de *spear phishing* sobre la organización "IA & Ciberseguridad".

El *framework* fue puesto a prueba en su totalidad sobre la organización "IA & Ciberseguridad" en donde presento un correcto funcionamiento de todas sus fases sin mayores complicaciones, logrando ser considerado un prototipo efectivo y escalable para próximas iteraciones.

En relación con los resultados de la prueba operacional, se observa que un 17,95 % de los participantes interactuó con los correos simulados de phishing, lo cual evidencia una exposición real ante este tipo de amenazas. Este resultado adquiere mayor relevancia si se considera que los participantes contaban con conocimientos superiores al promedio en ciberseguridad y estaban al tanto de su participación en una prueba, lo que les otorgaba una ventaja considerable frente a un público general. Cabe destacar que, independientemente del

porcentaje, la ocurrencia de al menos una interacción exitosa ya representa una potencial brecha de seguridad.

Como trabajo futuro se espera continuar mejorando el funcionamiento y robustez del *framework* a través de la automatización de procesos manuales como lo pueden ser el envío de documentos y de servicios de hosting para las páginas web necesarias. Además, se espera continuar con la ejecución de la prueba operacional sobre la Facultad de Ingeniería en los próximos meses, reanudando así la etapa de evaluación de la fase de envío de correos junto con el comité de ética de la organización para finalmente llevar a cabo la prueba. Finalmente, se plantea una línea de trabajo centrada en la implementación de nuevas tecnologías, especialmente en los modelos de lenguaje y los métodos de generación de texto. Técnicas de optimización de modelos, como QLoRA [33], permitirán el uso de modelos más grandes en términos de parámetros, lo que influirá directamente en la mejora de los textos generados.

7 Referencias

[1] Davis, G. B. (1974). *Management information systems: Conceptual foundations, structure, and development*. McGraw-Hill.

[2] Fisher, W., Craft, R. E., Ekstrom, M., Sexton, J., & Sweetnam, J. (2024). *Data confidentiality* (NIST Special Publication 1800-28). https://doi.org/10.6028/nist.sp.1800-28

[3] Nieles, M., Dempsey, K., & Pillitteri, V. Y. (2017). *An introduction to information security*. https://doi.org/10.6028/nist.sp.800-12r1

[4] *The Main Social Engineering Techniques Aimed at Hacking Information Systems.* (2021, 13 mayo). IEEE Conference Publication | IEEE Xplore. https://ieeexplore.ieee.org/abstract/document/9455031

[5] (ISC)². (2021). Social engineering. En *CISSP Certified Information Systems Security Professional Official Study Guide* (9.ª ed., p. 81). Wiley

[6] IBM. (s. f.). ¿Qué es spear phishing? https://www.ibm.com/es-es/topics/spear-phishing

[7] Varshney, G., Kumawat, R., Varadharajan, V., Tupakula, U., & Gupta, C. (2024). Anti-phishing: A comprehensive perspective. *Expert Systems with Applications, 238*, 122199. https://doi.org/10.1016/j.eswa.2023.122199

[8] Check Point. (s. f.). *Cyber Security Report 2022*. En Capítulo 5: Estadísticas globales.

[9] Comcast Business. (s. f.). *Enterprise cybersecurity solutions*. https://business.comcast.com

[10] IBM. (s. f.). *What are large language models (LLMs)?* https://www.ibm.com

[11] San Martín, R. (2024). *Evaluando la peligrosidad del spear phishing generado con soporte de IA generativa* [Memoria de título, Universidad de Concepción].

[12] Ministerio del Interior y Seguridad Pública, Agencia Nacional de Ciberseguridad. (2025, 28 de febrero). *CVE 2617388*.

[13] Joint Task Force. (2020b). *Security and privacy controls for information systems and organizations* (NIST SP 800-53 Rev. 5, p. 60). https://doi.org/10.6028/nist.sp.800-53r5

[14 scrapegraphai. (2024, junio 21). *PyPI*. https://pypi.org/project/scrapegraphai/

[15] Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2018). Learning to write with cooperative discriminators. En *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1638–1649). https://aclanthology.org/P18-1152/

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. En *Proceedings of NAACL-HLT* (pp. 4171–4186). https://aclanthology.org/N19-1423/

[17] Mahapatra, J., & Garain, U. (2024, julio 19). *Impact of model size on fine-tuned LLM performance in data-to-text generation: A state-of-the-art investigatio*n. arXiv. https://arxiv.org/abs/2407.11847

[18] Label Your Data. (2024, diciembre 19). *LLM model size: Parameters, training, and compute needs*. https://labelyourdata.com/articles/llm-model-size

[19] Hugging Face. (2024, diciembre 6). *meta-llama/Llama-3.1-8B.* https://huggingface.co/meta-llama/Llama-3.1-8B

[20] GeeksforGeeks. (2024, septiembre 24). *Recommended hardware for running LLMs locally*. https://www.geeksforgeeks.org

[21] LLMCompass: Enabling efficient hardware design for large language model inference. (2024, junio 29). *IEEE Conference Publication*. https://ieeexplore.ieee.org/document/10459274

[22] Giray, L. (2023). Prompt engineering with ChatGPT: A guide for academic writers. *Annals of Biomedical Engineering, 51*, 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

[23] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT*.arXiv. https://arxiv.org/abs/2302.11382

[24] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2022, enero 28). *Chain-of-thought prompting elicits reasoning in large language models*. arXiv. https://arxiv.org/abs/2201.11903

[25] Hadnagy, C. (2018). *Social engineering: The science of human hacking* (2nd ed.). Wiley.

[26] Quiel, S. (2013). *Social engineering in the context of Cialdini's psychology of persuasion and personality traits* (p. 19). https://doi.org/10.15480/882.1124

[27] Cialdini, R. B. (2009). *Influence: Science and practice* (5th ed.). Pearson.

[28] Guadagno, R., & Cialdini, R. B. (2005). Online persuasion and compliance: Social influence on the internet and beyond. En Y. Amichai-Hamburger (Ed.), *The social net: Human behavior in cyberspace* (pp. 91–113). Oxford University Press.

[29] Streamlit. (s. f.). *Streamlit Docs*. https://docs.streamlit.io/

[30] Quiel, S. (2013). *Social engineering in the context of Cialdini's psychology of persuasion and personality traits* (p. 28). https://doi.org/10.15480/882.1124

[31] Simply Psychology. (2024, febrero 13). *Hawthorne Effect in Psychology: Experimental studies*. https://www.simplypsychology.org/hawthorne-effect.html

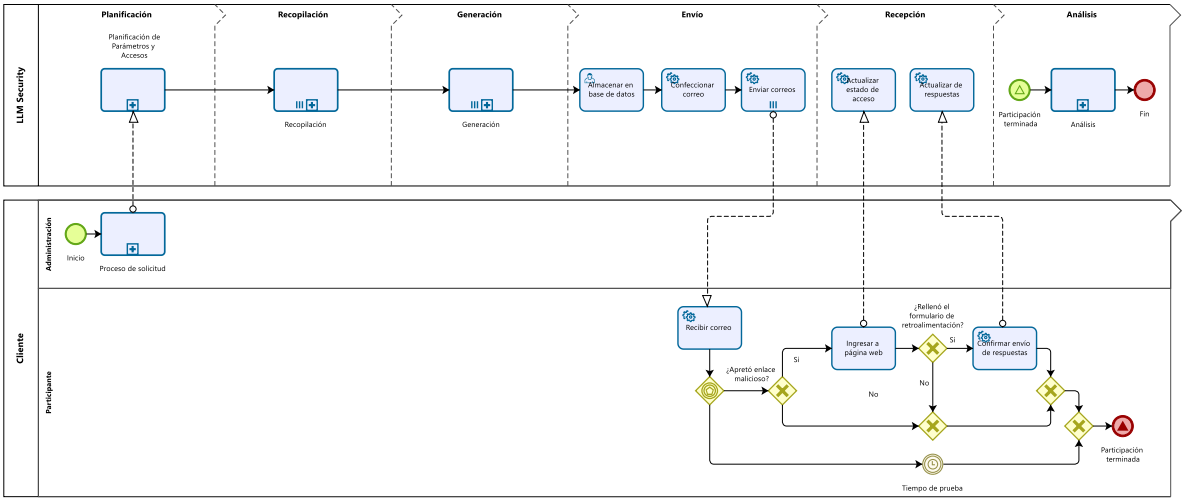
[32] Meta AI. (s. f.). *Introducing Meta Llama 3: The most capable openly available LLM to date*. https://ai.meta.com/blog/meta-llama-3/

[33] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023, mayo 23). *QLoRA: Efficient finetuning of quantized LLMs*. arXiv. https://arxiv.org/abs/2305.14314v1

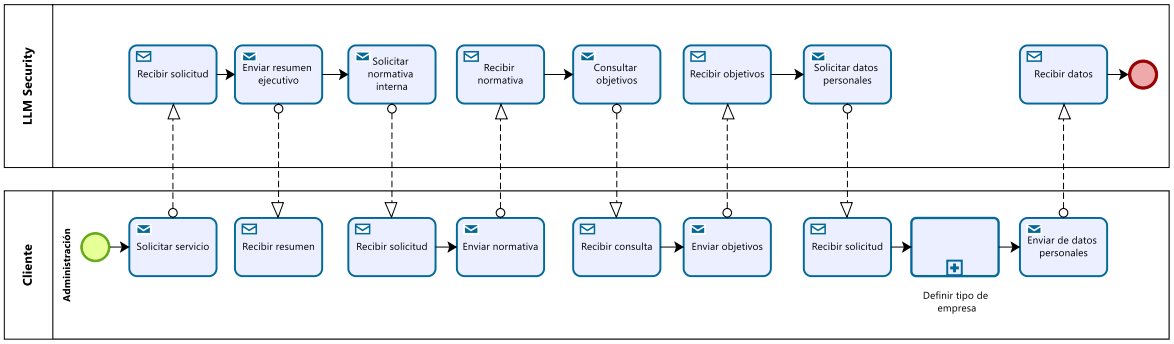
8 Anexos

8.1 Flujo de desarrollo propuesto

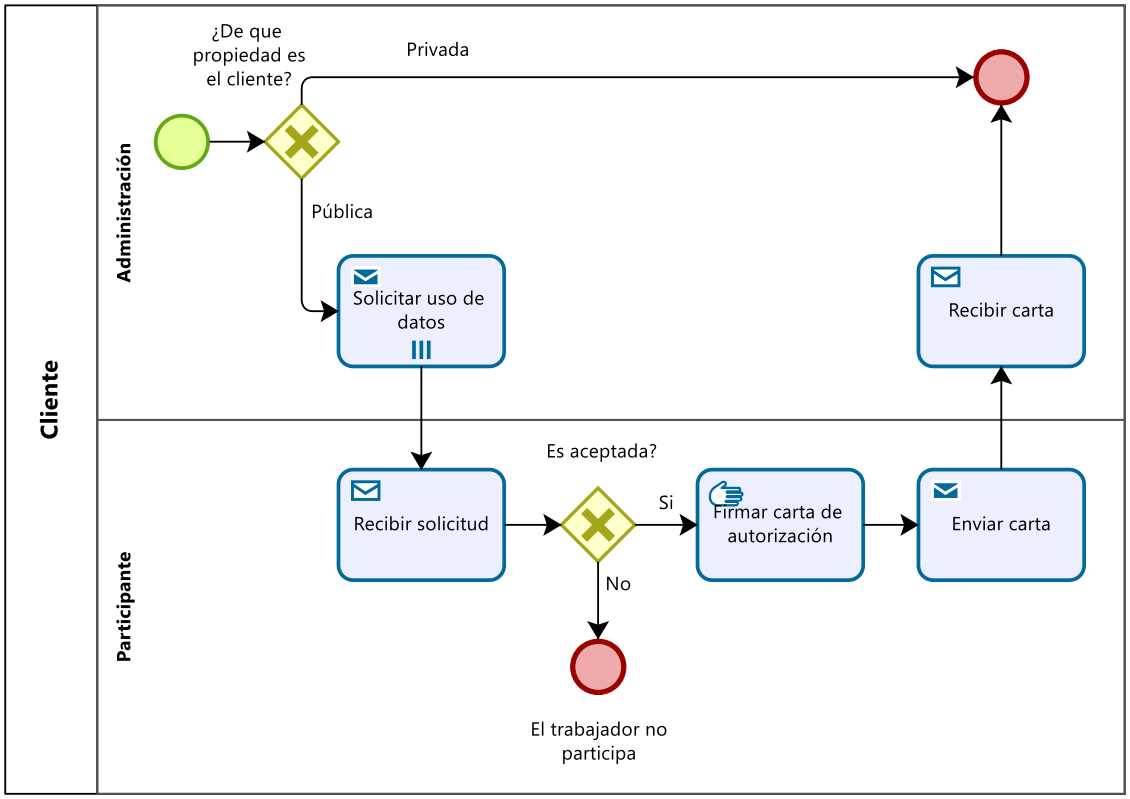
8.1.1 Proceso “Modelado de procesos del Framework”



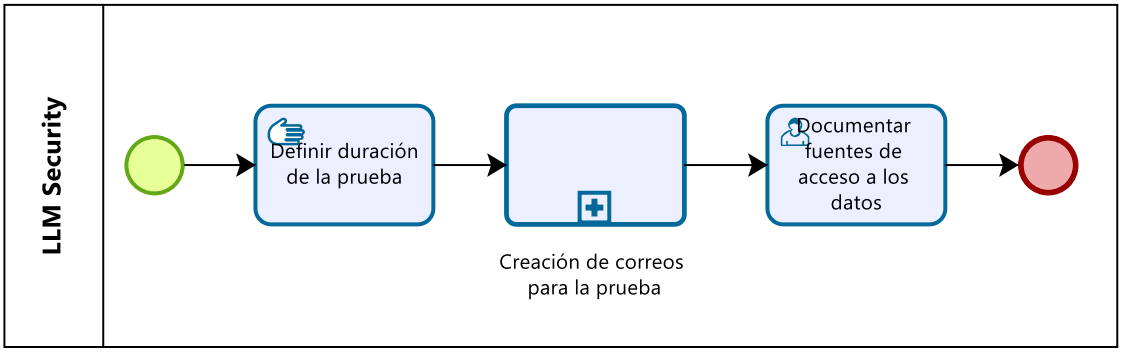
8.1.2 Subproceso “Proceso de solicitud”



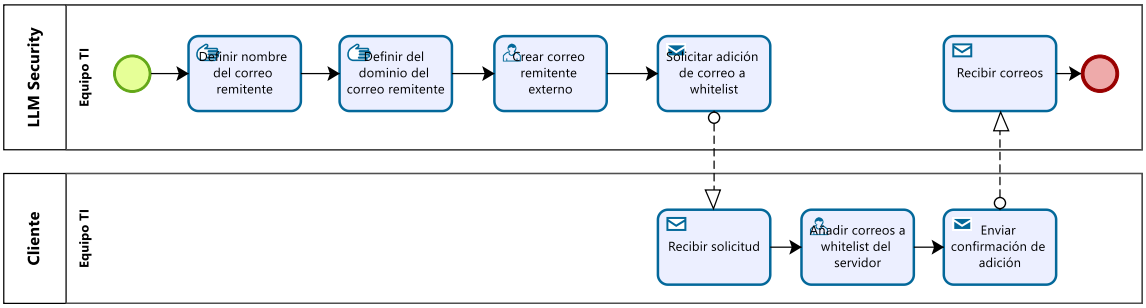
8.1.3 Subproceso “Definición de tipo de empresa”



8.1.4 Subproceso “Planificación de acceso y parámetros”



8.1.5 Subproceso “Creación de correos para la prueba”



8.2 Resumen Ejecutivo de la Etapa de Pruebas

Resumen ejecutivo
Etapa de pruebas para la memoria de titulo
“Hacia el desarrollo de una herramienta de evaluación de vulnerabilidades de ingeniería social”

En el ámbito de la seguridad informática, la ingeniería social se ha convertido en una de las principales amenazas para las organizaciones. Con el fin de evaluar las vulnerabilidades de ingeniería social en la Universidad de Concepción, se ha diseñado una prueba que simula ataques de spear phishing. Esta prueba se enfocará en el personal del departamento de Ingeniería Civil Informática, utilizando datos institucionales para crear correos electrónicos personalizados que imiten intentos reales de phishing. Durante un período de dos semanas, se enviarán estos correos y se analizarán las respuestas obtenidas, con el objetivo de identificar puntos débiles y fortalecer las defensas contra posibles ciberataques futuros.

1. Objetivo General
El objetivo principal es desarrollar y probar un prototipo que evalúe las vulnerabilidades de ingeniería social en la Facultad de Ingeniería de la Universidad de Concepción, utilizando ataques de spear phishing.

2. Procedimiento
- Recopilación de Datos: Se emplearán métodos de web scrapping asistidos por grandes modelos de lenguaje (LLM) para recopilar datos institucionales desde el sitio web de la facultad, siempre cumpliendo con las normativas legales y previa autorización de las respectivas autoridades.
- Generación de Correos: Con los datos recopilados, generaremos correos de spear phishing utilizando nuevamente LLM.

- Envío de Correos: Los correos serán enviados a los empleados durante un periodo específico para evaluar sus respuestas.
- Análisis de Resultados: Las respuestas serán analizadas para identificar vulnerabilidades y generar un informe de análisis de resultados.

3. **Importancia de la Prueba**
Entender cómo los empleados responden a ataques de spear phishing es crucial para fortalecer nuestras defensas contra ciberataques. Los resultados de esta prueba nos permitirán identificar las principales vulnerabilidades y diseñar mejores estrategias de capacitación y defensa a futuro.

8.3 Carta de Consentimiento

Estimado (a) participante:

Ud. ha sido invitado/a participar en el estudio Evaluación Automática de Vulnerabilidades de Ingeniería Social Organizacional, a cargo de Maximiliano Vergara Ríos, estudiante de la carrera de Ingeniería Civil Informática y Ciencias de la Computación en el contexto de la memoria de título “Hacia el desarrollo de una herramienta de evaluación de vulnerabilidades de ingeniería social”, con el apoyo de los patrocinantes de memoria de título Pedro Pinacho Davidson y Fernando Andree Tercero Gutiérrez Gómez.

A. *Propósito de la investigación:* El objetivo de esta investigación es Desarrollar un proceso de evaluación automatizado de vulnerabilidades de Ingeniería Social Organizacional que permita apoyar campañas de capacitación para mitigar los riesgos de ciberataques orientados al vector social. Para esto, la evaluación se vale de mensajes de phishing personalizados (*spear phishing*) generados a través de antecedentes recolectados sobre las páginas públicas de la Universidad de Concepción.

B. *Descripción de su participación:* Si usted decide participar del estudio, se le pedirá que acepte su participación a través del presente consentimiento informado. Su participación consistirá en recibir entre 1 y 3 correos de phishing generados personalizadamente a partir de antecedentes recolectados sobre las páginas web institucionales. Tanto la recolección de los antecedentes, como la generación de los correos phishing se hace de forma automática a través de sistemas basados en grandes modelos de lenguaje (LLM).

C. *Posibles riesgos:* Si bien el estudio es completamente inocuo, pues no está asociado a ningún acto doloso, las personas participantes en el estudio recibirán junto a su correo regular un conjunto de correos indeseados (1 a 3) durante el estudio, lo cual podría causar incomodidad o temor de estar en riesgo de ser víctimas de ciberataques. Para mitigar estas sensaciones, en caso de que un usuario realice una acción indebida (víctima positiva del phishing generado), el sistema entregará feedback inmediato dando a conocer que esto es un estudio y que no ha sido víctima de una estafa.

D. Beneficios: Los participantes tendrán la oportunidad de contar con retroalimentación respecto a sus errores que en un caso no simulado de ataque habrían significado el compromiso de sus recursos personales u organizacionales, además contarán al final del estudio de información consolidada respecto a vulnerabilidades de ingeniería social más prevalentes en la institución, reporte que podrá ser utilizado por el departamento o facultad para tomar medidas de capacitación correctiva.

E. *Confidencialidad y resguardo de la información:* Toda la información derivada de su participación será manejada con estricta confidencialidad. Sólo el equipo que desarrolla la investigación tendrá acceso a los datos por proporcionados. La información será resguardada según todos los requerimientos que las leyes chilenas explicitan (ley 20.120). Asimismo, tanto en el análisis, como en la publicación y difusión científica de los resultados, no se identificará la identidad de ninguno de los/as participantes ni su respectiva organización, para así resguardar el anonimato y no dar a conocer la presencia de vulnerabilidades. La información que entregue mediante su participación sólo será utilizada con fines científicos y relativos a esta investigación y no será usada con fines ajenos a los explícitamente expresados en este documento. Sólo se entregará información sobre fallas particulares ante ataques con rasgos específicos de ingeniería social a los usuarios que demuestren ser vulnerables a estas amenazas. El reporte general por entregar a la facultad y a todos los participantes contiene solo estadísticas de carácter general y hallazgos relevantes anonimizados que no permiten identificar a las personas participantes.

F. Voluntariedad: La participación en esta investigación es absolutamente voluntaria y usted puede retirarse en cualquier momento del estudio, sin que ello tenga ninguna consecuencia informando oportunamente al investigador responsable (vía correo electrónico maxvergara2018@udec.cl)

G. *Derechos del/de la participante:* Los participantes tienen recibirán un reporte detallado de los emails enviados hacia ellos durante el estudio, y el reporte de vulnerabilidades de ingeniería social detectadas sobre ellos. Para cualquier duda que no me haya sido satisfactoriamente resuelta por el investigador responsable me podré dirigir a la Dra. Andrea Rodríguez T. Presidenta del Comité de Ética, Bioética y Bioseguridad de la Universidad de Concepción. Correo: secrevrid@udec.cl.

He recibido y comprendido la información de este documento. He podido aclarar todas mis dudas y otorgo el consentimiento para participar en el estudio: “Evaluación Automática de Vulnerabilidades de Ingeniería Social Organizacional”.
Comprendo y acepto la información que se entregó anteriormente y declaro conocer los objetivos del estudio.

En atención a estas consideraciones, puede proceder a contestar el formulario.

8.4 Reporte institucional

REPORTE INSTITUCIONAL
Evaluación de vulnerabilidades de ingeniería social sobre grupo de investigación IA & Ciberseguridad

Estimada institución,
A continuación, se presentarán los resultados de la prueba operacional realizada con el grupo de investigación "IA & Ciberseguridad" de la Universidad de Concepción. Esta prueba tiene como objetivo evaluar la efectividad de los ataques de spear phishing de su organización, otorgando información valiosa acerca de cuan preparado su personal para enfrentar este tipo de ataques.
Resumen de Resultados:

1. **Tasa de Éxito y Reconocimiento del Ataque:**
    - Se enviaron un total de 39 correos electrónicos, de los cuales 8 resultaron exitosos, entendiendo por éxito que el enlace dentro del correo fue abierto. De estas 8 ocasiones exitosas, solo 7 participantes completaron el formulario de retroalimentación adjunto en la página web.
    - Todos los participantes que hicieron clic en el enlace identificaron el correo como un ataque de spear phishing. La razón principal para interactuar con el enlace fue la curiosidad, ya que el 100% de los

encuestados admitió tener el conocimiento necesario para reconocer un ataque de este tipo.

2. **Conocimiento del Tema:**
    - Los resultados confirman que el grupo poseía un nivel general alto sobre phishing. El 100% de los participantes indicaron que tienen el conocimiento suficiente para identificar un ataque de spear phishing, corroborando la hipótesis inicial.
    - Aun así, sus conocimientos no son certeros en lo técnico ya que solo un 29% de los participantes acertó en el principio de influencia en el cual se basaba el ataque recibido.
3. **Impacto de la Notificación Previa:**
    - La notificación previa a los participantes sobre su participación en la prueba resultó ser un factor significativo. El 71% de los encuestados señaló que estar al tanto de su rol en la prueba generó dudas sobre la autenticidad de los correos recibidos. Además, el 100% coincidió en que la probabilidad de éxito de la prueba habría sido mayor si no hubieran sido informados previamente.
4. **Sugerencias para Mejoras:**
    - En cuanto al formato del correo, el 100% de los participantes sugirió que la inclusión de recursos visuales, como imágenes, podría aumentar la credibilidad del correo. Esta sugerencia será considerada para futuras mejoras de la herramienta.

Conclusiones: La prueba confirma que el grupo de investigación "IA & Ciberseguridad" de la Universidad de Concepción tiene un conocimiento elevado acerca de vulnerabilidades de ingeniería social, demostrando así que están preparados ante ataques de spear phishing reales.

Agradecemos la colaboración de todos los participantes y su apoyo en la realización de esta prueba. Si tiene alguna pregunta o desea discutir estos resultados en detalle, no dude en ponerse en contacto.
Atentamente,

Maximiliano Vergara.
Encargado de prueba de vulnerabilidad.