

# DESARROLLO DE UN CHATBOT RAG CON MEMORIA



David Fernández Prieto 44663898Z [davidvispo4211@gmail.com](mailto:davidvispo4211@gmail.com)  
Daniel Barandela Cachafeiro 45163395C [danielbaracacha@gmail.com](mailto:danielbaracacha@gmail.com)

## **Descripción del problema**

El objetivo es desarrollar un chatbot que use RAG y que a partir de una consulta del usuario sobre un ataque cibernético sea capaz de diagnosticar qué técnicas de MITRE ATT&CK se están utilizando y de ofrecer contramedidas a un escenario descrito.

Para desarrollarlo se plantean dos opciones: un RAG simple y otro que mantenga un memoria del historial de mensajes. En nuestro caso nos decidimos por la segunda opción, el asistente deberá ser capaz de mantener el contexto de la conversación y responder a preguntas sobre respuestas anteriores.

## Descripción de la implementación

Contamos con un archivo JSON que contiene técnicas y subtécnicas de MITRE ATT&CK. De cada técnica, se extraen atributos clave como la descripción, los indicadores de detección y los métodos de mitigación (nombre y descripción). Estos datos se estructuran en documentos, dividiendo el contenido principal y los metadatos para su posterior procesamiento.

Para realizar búsquedas semánticas eficientes, se genera un vector-store utilizando Chroma. Los documentos se representan mediante embeddings creados con el modelo **text-embedding-3-small** de OpenAI, lo que permite medir similitudes semánticas entre las consultas de los usuarios y las técnicas almacenadas.

La función **generate\_context** se encarga de recuperar las técnicas más relevantes del vector-store, tomando las cinco (por defecto) más similares a la consulta. Luego, construye un contexto que combina las descripciones, los indicadores de detección y los métodos de mitigación de estas técnicas, que será incluido en el prompt para facilitar al modelo una respuesta más correcta.

El sistema crea automáticamente el prompt a partir de la query del usuario, usando la respuesta de la función anterior junto con la propia query. También se añade un system prompt explicando que debe simular que es un experto en ciberataques y que evite responder a preguntas que no estén relacionadas con el tema.

La función **call\_model** es el núcleo que procesa las consultas del usuario, mandando los prompts al modelo y guardando la respuesta de este. Hacemos uso de StateGraph, que nos permite crear un flujo de trabajo (workflow) como un grafo, en el que cada nodo es una interacción con el modelo y que permite simular una memoria al 'recordar' todos los nodos que le suceden (usando MemorySaver).

El sistema integra direct-prompting para definir al asistente como experto en ciberseguridad y delimitar el alcance de sus respuestas.

## Ejemplo de ejecución

>> ¿Qué técnicas puedo utilizar en contra del phishing?

===== Ai Message =====

Para combatir el phishing, puedes implementar varias técnicas de detección y mitigación que son fundamentales para proteger tu organización. Aquí te presento algunas:

### ### Detección:

1. **\*\*Monitorización de correos electrónicos sospechosos\*\***: Establece sistemas para analizar y detectar actividad inusual en correos electrónicos, como mensajes recibidos desde un remitente desconocido o inusual que llegue a varios destinatarios a la vez.
2. **\*\*Filtrado basado en DKIM y SPF\*\***: Implementa la verificación mediante DKIM (DomainKeys Identified Mail) y SPF (Sender Policy Framework) para identificar correos electrónicos que podrían estar siendo enviados desde un dominio falsificado.
3. **\*\*Inspección de URLs\*\***: Realiza inspecciones de los enlaces dentro de los correos electrónicos, asegurándote de expandir cualquier enlace acortado y verificar si redirigen a sitios maliciosos conocidos o no categorizados.
4. **\*\*Monitoreo de tráfico en redes sociales\*\***: Presta atención a cualquier actividad inusual en redes sociales, como cuentas sospechosas o mensajes que soliciten información sensible.

### ### Mitigación:

1. **\*\*Entrenamiento de usuarios\*\***: Capacita a los empleados para que sean conscientes de los intentos de acceso o manipulación. Esto es clave para reducir el riesgo de ataques de spearphishing y otras técnicas de ingeniería social.
2. **\*\*Configuración de software\*\***: Asegúrate de que las configuraciones de seguridad de tu software estén adecuadas para minimizar los riesgos asociados a su uso.
3. **\*\*Restricción de contenido web\*\***: Limita el acceso a sitios web determinados y bloquea descargas o archivos adjuntos que no sean necesarios. Esto también incluye la restricción de ciertas características en navegadores.
4. **\*\*Auditoría de sistemas\*\***: Realiza auditorías regulares de los sistemas y configuraciones para identificar y mitigar posibles vulnerabilidades.
5. **\*\*Uso de antivirus y antimalware\*\***: Implementa software que utilice firmas o heurísticas para detectar ficheros maliciosos y comportamientos sospechosos en los sistemas.

La combinación de estas estrategias de detección y mitigación puede fortalecer significativamente la postura de ciberseguridad de tu organización contra las amenazas de phishing y otros vectores de ataque similares.

## >> ¿Cómo funciona la técnica Valid Accounts?

===== Ai Message =====

La técnica de **Valid Accounts** (T1078) se refiere a los métodos que utilizan los adversarios para obtener acceso a cuentas válidas en un sistema o entorno comprometido. Esta técnica es crucial para atacar porque permite a los adversarios realizar múltiples acciones maliciosas, como el acceso no autorizado, el movimiento lateral y la elevación de privilegios. A continuación, se detallan algunos aspectos clave de cómo funciona esta técnica.

### ### Funcionamiento de la técnica Valid Accounts:

1. **Enumeración de cuentas**: Los atacantes pueden intentar obtener una lista de cuentas de usuario, nombres de usuario o direcciones de correo electrónico dentro de un sistema. Usan herramientas y comandos integrados, así como configuraciones erróneas que pueden filtrar información sobre nombres de cuentas y sus permisos. Por ejemplo, en entornos de nube, las interfaces generalmente brindan acceso fácil a listas de usuarios.
2. **Acceso a cuentas válidas**: Una vez que el atacante tiene acceso a las credenciales de cuentas válidas (ya sea a través de técnicas como phishing, uso de contraseñas débiles o explotación de credenciales previamente robadas), puede actuar como el usuario legítimo, lo que les permite realizar acciones como copia de archivos, administración de recursos, o cambiar configuraciones.
3. **Uso de cuentas para movimiento lateral y escalación de privilegios**: Los adversarios pueden utilizar cuentas válidas para moverse lateralmente en la red, acceder a recursos compartidos y ejecutar comandos en sistemas remotos. Esto les permite expandir su acceso incluso más que el inicialmente comprometido.

### ### Detección:

Para detectar el uso inadecuado de cuentas válidas, se pueden implementar las siguientes medidas:

- **Monitoreo de logs de autenticación**: Realizar un seguimiento de los intentos de inicio de sesión fallidos para obtener información que pueda indicar un ataque de fuerza bruta o un intento de acceso no autorizado.
- **Detección de actividad sospechosa**: Vigilar el uso de comandos que puedan enumerar cuentas de usuario, como ``net.exe`` o ``net1.exe``, especialmente si se ejecutan en rápida sucesión.

### ### Mitigación:

Las estrategias para mitigar los riesgos asociados con la técnica Valid Accounts incluyen:

1. **Gestión de cuentas de usuario**: Es fundamental manejar adecuadamente la creación, modificación, uso y permisos de las cuentas de usuario, asegurando que solo las personas adecuadas tengan acceso a los recursos necesarios.

2. **\*\*Implementación de la autenticación multifactor\*\***: Utilizar múltiples métodos de verificación para autenticar a los usuarios reduce el riesgo de accesos no autorizados, incluso si un atacante tiene las credenciales.
3. **\*\*Políticas de contraseñas robustas\*\***: Establecer y hacer cumplir políticas que requieran contraseñas fuertes y cambios regulares de las mismas.
4. **\*\*Configuraciones de seguridad del sistema operativo\*\***: Fortalecer la configuración del sistema operativo y cerrar vulnerabilidades que puedan ser explotadas por adversarios para obtener acceso a cuentas.

Entender y abordar la técnica Valid Accounts es crucial para mantener una buena postura de seguridad cibernética y proteger a la organización contra accesos no autorizados y otros comportamientos maliciosos asociados.

### **>> ¿Qué es "Privilege Escalation" y cómo afecta a la seguridad de los sistemas?**

===== Ai Message =====

La **\*\*escalación de privilegios\*\*** ("Privilege Escalation") es un proceso mediante el cual un adversario obtiene derechos o permisos más altos en un sistema del que originalmente tenía. Esto a menudo permite al atacante acceder a información sensible, ejecutar comandos con más autoridad, o realizar cambios en la configuración del sistema. Esta técnica puede tener un impacto significativo en la seguridad de los sistemas por varias razones:

#### **### Cómo afecta a la seguridad de los sistemas:**

1. **\*\*Acceso no autorizado\*\***: La escalación de privilegios permite a un atacante eludir las medidas de seguridad establecidas, accediendo a recursos que deberían estar protegidos. Esto puede resultar en la exposición de datos sensibles o confidenciales.
2. **\*\*Control total del sistema\*\***: Al obtener privilegios de administrador o "root", un adversario puede realizar cualquier acción en el sistema, incluyendo el cambio de configuraciones de seguridad, la eliminación de registros o el acceso a otros sistemas de la red.
3. **\*\*Instalación de software malicioso\*\***: Con privilegios elevados, un atacante puede instalar malware, como troyanos o herramientas de administración remota, lo que le permite mantener el acceso al sistema a largo plazo.
4. **\*\*Movimientos laterales\*\***: La escalación de privilegios a menudo es un paso necesario para que un atacante se mueva lateralmente a otros sistemas dentro de la red, expandiendo su alcance y potencialmente comprometiendo más activos.
5. **\*\*Dificultad en la detección\*\***: Las actividades realizadas bajo cuentas de alto privilegio son más difíciles de detectar porque a menudo son imperceptibles para los mecanismos de monitoreo. Esto significa que los ataques pueden permanecer ocultos durante períodos más prolongados.

### ### Estrategias de Detección y Mitigación:

#### **\*\*Detección\*\*:**

- **\*\*Monitoreo de Logs\*\*:** Implementar un monitoreo en tiempo real de los logs del sistema para detectar cambios en la actividad de cuentas o en los permisos de archivos.
- **\*\*Comportamiento Anómalo\*\*:** Establecer alertas basadas en patrones de comportamiento inusuales que puedan indicar que un atacante ha escalado sus privilegios.

#### **\*\*Mitigación\*\*:**

1. **\*\*Actualización de Software\*\*:** Realizar actualizaciones regulares de software para corregir vulnerabilidades que podrían ser aprovechadas para escalar privilegios.
2. **\*\*Seguridad del Sistema Operativo y Control de Acceso\*\*:** Configurar adecuadamente el sistema operativo y aplicar estrictas políticas de control de acceso para restringir el alcance de las cuentas de usuario.
3. **\*\*Separación de Privilegios\*\*:** Implementar el principio de menor privilegio, otorgando a los usuarios solo los accesos necesarios para cumplir sus tareas.
4. **\*\*Autenticación Multifactor\*\*:** Utilizar autenticación multifactor (MFA) para añadir una capa adicional de seguridad a las cuentas de alto privilegio.
5. **\*\*Auditoría de Seguridad\*\*:** Realizar auditorías periódicas de seguridad para identificar y remediar configuraciones inseguras que puedan facilitar la escalación de privilegios.

En resumen, la escalación de privilegios es una técnica crítica que puede comprometer la seguridad general de un sistema. La implementación de medidas de detección y mitigación es esencial para proteger los activos de información y mantener una postura de seguridad robusta.

### **>> Desarrolla la primera medida de mitigación para el phishing que mencionaste.**

===== Ai Message =====

Desarrollemos la primera medida de mitigación contra el phishing, que es **\*\*la capacitación de usuarios\*\***. Esta medida es crucial, ya que gran parte del éxito de los ataques de phishing depende de la interacción humana. Aquí está un enfoque detallado de cómo implementar un programa efectivo de capacitación para usuarios:

### ### Capacitación de Usuarios en Prevención de Phishing

#### 1. **\*\*Objetivos del Programa de Capacitación\*\*:**

- Aumentar la concienciación sobre las amenazas cibernéticas, específicamente el phishing y sus variantes.
- Proporcionar a los empleados las herramientas y el conocimiento necesarios para identificar intentos de phishing.
- Fomentar una cultura de seguridad en la organización.

#### 2. **\*\*Contenido del Programa\*\*:**

- **\*\*Introducción al Phishing\*\*:** Explicar qué es el phishing, cómo funciona y sus diferentes formas (como spearphishing, whaling, y otros).

- **\*\*Técnicas Comunes Utilizadas por los Adversarios\*\***: Incluir ejemplos prácticos de correos electrónicos y mensajes de phishing que podrían recibir, resaltando tácticas como la suplantación de identidad, el uso de urgencia y la manipulación emocional.

- **\*\*Señales de Alerta\*\***: Enseñar a los empleados a reconocer indicadores de un ataque de phishing, tales como:

- Remitentes desconocidos o direcciones de correo electrónico sospechosas.
- Errores gramaticales o de redacción en el texto del mensaje.
- Enlaces que llevan a sitios web no relacionados o maliciosos.
- Solicitudes inusuales de información personal o de inicio de sesión.

- **\*\*Prácticas de Respuesta\*\***: Instruir sobre qué hacer al recibir un mensaje sospechoso:
  - No hacer clic en enlaces ni descargar archivos adjuntos.
  - Verificar la autenticidad del mensaje comunicándose directamente con el remitente a través de canales conocidos.

- Informar al departamento de TI o al equipo de ciberseguridad sobre el mensaje.

### 3. **\*\*Metodología de Capacitación\*\***:

- **\*\*Sesiones Presenciales/Virtuales\*\***: Organizar talleres donde se puedan compartir ejemplos reales y realizar discusiones.

- **\*\*Materiales en Línea\*\***: Proporcionar acceso a módulos de e-learning, videos informativos y guías.

- **\*\*Simulaciones de Phishing\*\***: Realizar ejercicios prácticos donde se envíen correos electrónicos simulados de phishing y se evalúe cómo responden los empleados. Esto puede ayudar a poner en práctica el conocimiento adquirido y a ajustar el programa según los resultados.

### 4. **\*\*Frecuencia de Capacitación\*\***:

- **\*\*Capacitación Inicial\*\***: Proporcionar una sesión de capacitación introductoria a todos los nuevos empleados tan pronto como se incorporen a la organización.

- **\*\*Refrescos Periódicos\*\***: Realizar sesiones de actualización al menos una vez al año, y considerar el uso de recordatorios periódicos sobre temas de seguridad.

### 5. **\*\*Evaluación de la Efectividad\*\***:

- **\*\*Feedback y Evaluaciones\*\***: Al final de las sesiones, recopilar comentarios de los participantes y realizar evaluaciones para medir el nivel de comprensión.

- **\*\*Monitoreo de Incidentes\*\***: Evaluar la tasa de incidentes de phishing exitosos después de la capacitación para determinar su efectividad y realizar ajustes según sea necesario.

## ### Conclusión

La capacitación de usuarios es una defensa fundamental en la lucha contra el phishing, ya que educar a los empleados sobre cómo reconocer y responder a estos ataques puede reducir significativamente el riesgo de éxito de los intentos de phishing. Al empoderar a los usuarios con conocimiento y habilidades, una organización puede fortalecer su postura de seguridad y fomentar una cultura de vigilancia continua frente a amenazas cibernéticas.



# Pruebas

## Primera prueba:

Con  $K=5$ .

### ¿Qué técnicas puedo utilizar en contra del phishing?

Responde directamente con medidas de detección y mitigación.

### ¿Cómo funciona la técnica Valid Accounts?

Da una descripción, esta vez más amplia, sobre lo que se le ha preguntado y explica en pasos cómo funciona dicha técnica. A pesar de que no se le ha preguntado sobre técnicas relacionadas aporta medidas de detección y mitigación.

### ¿Qué es "Privilege Escalation" y cómo afecta a la seguridad de los sistemas?

Primero responde correctamente a las preguntas formuladas y después vuelve a aportar medidas de detección y mitigación a pesar de que no se le pregunta sobre ello.

### Desarrolla la primera medida de mitigación para el phishing que mencionaste.

Responde con objetivos, metodologías y conclusiones; siendo una descripción muy completa de la medida mencionada. Queda comprobada así la capacidad del modelo para recordar respuestas anteriores.

## Resumen:

Responde con una estructura que facilita la comprensión de las distintas partes de la respuesta y la mantiene para todas las preguntas. Cuenta con bastante información y responde correctamente pero tiende a aportar métodos de mitigación y detección aunque no se le pregunte sobre ello.

## Segunda prueba:

Con  $K=1$ .

### ¿Qué técnicas puedo utilizar en contra del phishing?

En este caso responde de una manera similar al anterior pero con menos técnicas y con explicaciones más breves.

### ¿Cómo funciona la técnica Valid Accounts?

En esta respuesta se aprecian más diferencias, la descripción es mucho más breve y menos detallada, además en este caso no aporta el funcionamiento separado en pasos. Sigue aportando técnicas de mitigación y detección a pesar de que no se le pregunte por ello.

### ¿Qué es "Privilege Escalation" y cómo afecta a la seguridad de los sistemas?

La diferencia más destacable en esta respuesta es que a la hora de responder a cómo afecta a la seguridad solamente se limita a decir en qué formas afecta pero no las explica, simplemente hace una mención.

**Resumen:**

Al reducir el valor de k notamos como el modelo cuenta con menos información para las respuestas y a pesar de aportar información correcta podemos decir que su capacidad de respuesta es más limitada.

**Tercera prueba:**

Con K=5 sin system prompt.

**¿Qué técnicas puedo utilizar en contra del phishing?**

En este caso la respuesta es más informal pero se mantiene correcta.

**¿Cómo funciona la técnica Valid Accounts?**

La respuesta es ligeramente más breve que en la primera prueba pero sigue siendo correcta.

**¿Qué es "Privilege Escalation" y cómo afecta a la seguridad de los sistemas?**

La respuesta es similar en extensión a la primera prueba pero vemos un cambio en la estructura de la respuesta.

**Resumen:**

La diferencia que notamos al cambiar el system prompt no está tan relacionada con la calidad de la información, que sí empeora ligeramente, si no con el tono y sobre todo con la estructura en las respuestas que no está tan organizada y varía entre las diferentes preguntas.

## Conclusiones

A raíz de las distintas pruebas informales que hemos llevado a cabo podemos concluir lo siguiente:

El rendimiento del modelo es satisfactorio, es capaz de responder a las preguntas que se le formulan con información relevante y de forma clara, también mantiene el contexto de la conversación, lo que nos permite hacerle preguntas para que aclare respuestas anteriores o que desarrolle algunos puntos.

Al modificar el valor de K (el número de casos más similares de los que extrae la información) impactamos directamente a la calidad y cantidad de información con la que contará el modelo para la respuesta. Al reducirla el modelo cuenta con información de gran calidad pero la respuesta no será extensa. Al aumentarla tenemos el efecto contrario, cuenta con más información para la respuesta pero con valores excesivamente altos responderá con información que podría ser irrelevante para el caso.

Finalmente, al modificar el system prompt no notamos un gran cambio en la calidad de la información, que se mantiene similar, si no que las mayores diferencias las apreciamos en la estructura y el tono de las respuestas. Gracias al system prompt es capaz de mantener una misma estructura en todas las respuestas, además dicha estructura facilita en gran medida la comprensión de la información. También notamos que al suprimir el system prompt utiliza un tono mucho más informal. Con todo esto queda demostrada la importancia que tienen direct prompting y role prompting en los LLM, siendo una gran herramienta para mejorar la calidad de las respuestas.

## **Problemas encontrados**

Durante el desarrollo de este sistema nos encontramos con varios problemas, todos relacionados con el código y la relativa novedad que supone emplear LLMs para nosotros.

Cabe destacar que tuvimos problemas con las versiones de las librerías y las estructuras para usarlas y también que nos costó un poco entender el funcionamiento de Chroma para poder mantener un historial de interacciones.

## Bibliografía

-[www.stackoverflow.com](https://www.stackoverflow.com)

-[https://api.python.langchain.com/en/latest/vectorstores/langchain\\_chroma.vectorstores.Chroma.html](https://api.python.langchain.com/en/latest/vectorstores/langchain_chroma.vectorstores.Chroma.html)