Guía referencial de técnicas de ataque IA/LLM Mitre ATLAS + Vulnerabilidades OWASP Top 10 LLM

Reconocimiento
AML.TA0002
ATT&CK TA0043

- Búsqueda de materiales de investigación disponibles públicamente sobre las víctimas AML.T0000
  - Revistas y actas de conferencias AML.T0000.000
  - Repositorios de pre-impresiones AML.T0000.001
  - Blogs Técnicos AML.T0000.002
- Búsqueda de análisis de vulnerabilidades disponibles públicamente AML.T0001
- Búsqueda en sitios web de la víctima AML.T0003
- Búsqueda en repositorios de aplicaciones AML.T0004
- Escaneo activo AML.T0006 T1595
- Recopilar objetivos indexados por RAG AML.T0064

Desarrollo de recursos
AML.TA0003
ATT&CK TA0042

- Adquirir artefactos públicos de ML AML.T00002
  - Datasets AML.T0002.000
  - Modelos AML.T0002.001
- Obtener Capacidades AML.T0016 T1588
  - Implementaciones de ataques de ML AML.T0016.000
  - Software y Herramientas AML.T0016.001 T1588.002
  - IA Generativa AML.T0016.002
- Desarrollar capacidades AML.T0017 T1587
  - Ataques adversos de ML AML.TA0017.000
- Adquirir infraestructura AML.T0008
  - Desarrollo de espacios de trabajo de ML AML.T0008.000
  - Hardware de consumo AML.T0008.001
  - Dominios AML.T0008.002 T1583.001
  - Contramedidas físicas AML.T0008.003
- Publicar Datasets envenenados AML.T0019
- Envenenar data de entrenamiento AML.T0020
- Crear Cuentas AML.T0021 T1585
- Publicar Modelos envenenados AML.T0058
- Publicar entidades con alucinaciones AML.T0060
- Elaboración manual de prompts LLM AML.T0065
- Crear contenido para recuperación (RAG) AML.TA0066

Acceso Inicial
AML.TA0004
ATT&CK TA0001

- Compromiso de la cadena de suministro AML.T0010
- Ataque a la cadena de suministro OWASP LLM 003:2025
  - Hardware AML.T0010.000
  - Software ML AML.T0010.001
  - Datos AML.T0010.002
  - Modelo AML.T0010.003
- Cuentas válidas AML.T0012 T1078
- Evadir el modelo de ML AML.T0015
- Explotar aplicaciones publicamente expuestas AML.T0049 T1190
- Phishing AML.T0052 T1566
  - Spearphishing a tráves de ingeniería social hecha con LLM AML.T0052.000

Acceso al modelo
AML.TA0000

- Acceso a las APIs del modelo de inferencia de IA AML.T0040
- Producto o Servicio habilitado para ML AML.T0047
- Acceso al entorno físico AML.T0041
- Acceso total al modelo de ML AML.T0044

Ejecución
AML.TA0005
ATT&CK TA0002

- Ejecución por el usuario AML.T0011 T1204
  - Artefactos inseguros de ML AML.T0011.000
  - Paquetes de software maliciosos AML.T0011.001
- Interprete de Comandos y Scripts AML.T0050 T1059
- Inyección de Prompt LLM (Prompt Injection) AML.T0051
  - Inyección de Prompt LLM (Prompt Injection) OWASP LLM 001:2025
  - Exposición del prompt de sistema OWASP LLM 007:2025
  - Directa AML.T0051.000
  - Indirecta AML.T0051.001
- Compromiso de plugins LLM AML.T0053

Persistencia
AML.TA0006
ATT&CK TA0003

- Envenenar data de entrenamiento AML.T0020
- Puerta trasera (backdoor) en el modelo de ML AML.T0018
  - Envenenamiento del modelo de ML AML.T0018.000
  - Inyectar carga útil (payload) AML.T0018.001
- Auto-replicación del prompt LLM AML.T0061
  - Excesiva autonomía OWASP LLM 006:2025
- Envenenamiento RAG AML.T0070
  - Debilidades en vectores e incrustaciones OWASP LLM 008:2025

Escalamiento de privilegios
AML.TA0012
ATT&CK TA0004

- Compromiso de plugins LLM AML.T0053
- LLM Jailbreak AML.T0054
  - Inyección de Prompt LLM (Prompt Injection) OWASP LLM 001:2025
  - Excesiva autonomía OWASP LLM 006:2025

Evasión de la defensa
AML.TA0007
ATT&CK TA0005

- Evadir el modelo de ML AML.T0015
- LLM Jailbreak AML.T0054
  - Inyección de Prompt LLM (Prompt Injection) OWASP LLM 001:2025
  - Excesiva autonomía OWASP LLM 006:2025
- Manipulación de respuestas del LLM AML.T0067
  - Manejo inadecuado de la salida OWASP LLM 005:2025
  - Citas AML.T0067.000
- Ofuscación de prompt LLM AML.T0068
- Inyección de entrada RAG falsa AML.T0071

Acceso a Credenciales
AML.TA0013
ATT&CK TA0006

- Credenciales inseguras AML.T0055 T1552

Descubrimiento
AML.TA0008
ATT&CK TA0007

- Descubrir la ontología del modelo ML AML.T0013
- Descubrir la familia del modelo ML AML.T0014
- Descubrir artefactos de ML AML.T0007
- Descubrir alucinaciones en el LLM AML.T0062
- Descubrir las salidas del modelo IA AML.T0063
- Descubrir información del sistema LLM AML.T0069
  - Set de caracteres especiales AML.T0069.000
  - Palabras clave de instrucciones de sistema AML.T0069.001
  - Prompt de sistema (System prompt) AML.T0069.002
  - Exposición del prompt de sistema OWASP LLM 007:2025

Recopilación
AML.TA0009
ATT&CK TA0009

- Recopilación de artefactos de ML AML.T0035
- Datos de repositorios de información AML.T0036 T1213
- Datos del sistema local AML.T0037 T1005

Preparación de ataques ML
AML.TA0001

- Crear un modelo de ML proxy AML.T0005
  - Entrenar el proxy mediante artefactos de ML Recopilados AML.T0005.000
  - Entrenar el proxy vía replicación AML.T0005.001
  - Usar modelos pre-entrenados AML.T0005.002
- Puerta trasera (Backdoor) en modelo de ML AML.T0018
  - Envenenamiento del modelo y la data OWASP LLM 004:2025
  - Envenenar modelo de ML AML.T0018.000
  - Inyectar carga útil (payload) AML.T0018.001
- Verificar ataque AML.T0042
- Crear data contradictoria AML.T0043
  - Optimización de caja blanca AML.T0043.000
  - Optimización de caja negra AML.T0043.001
  - Transferencia de caja negra AML.T0043.002
  - Modificación manual AML.T0043.003
  - Insertar desencadenador de puerta trasera (backdoor trigger) AML.T0043.004

Exfiltración
AML.TA0010
ATT&CK TA0010

- Exfiltración a través de la API de inferencia de ML AML.T0024
  - Éxposición de información sensible OWASP LLM 002:2025
  - Manejo inadecuado de la salida OWASP LLM 005:2025
  - Inferir pertenencia de la data de entrenamiento AML.T0024.000
  - Invertir el modelo ML AML.T0024.001
  - Extraer el modelo de ML AML.T0024.002
  - Debilidades en vectores e incrustaciones OWASP LLM 008:2025
- Exfiltración por medios cibernéticos AML.T0025
  - Consumo ilimitado OWASP LLM 010:2025
- Extracción del prompt de sistema del LLM AML.T0056
  - Exposición del prompt de sistema OWASP LLM 007:2025
- Fuga de datos del LLM AML.T0057

Impacto
AML.TA0011
ATT&CK TA0040

- Evadir el modelo de ML AML.T0015
- Denegación de servicio AML.T0029
  - Consumo ilimitado OWASP LLM 010:2025
- Spamming con datos falsos al sistema de ML AML.T0046
- Afectar la integridad del modelo de ML AML.T0031
- Incrementar costos AML.T0034
  - Consumo ilimitado OWASP LLM 010:2025
- Daño externo AML.T0048
  - Daño financiero AML.T0048.000
  - Daño reputacional AML.T0048.001
  - Daño social AML.T0048.002
  - Desinformación OWASP LLM 009:2025
  - Daño al usuario AML.T0048.003
  - Robo de propiedad intelectual AML.T0048.004
- Afectar la integridad del dataset AML.T0059

MITRE ATLAS + OWASP TOP 10 LLM

ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems) es una base de conocimiento viva y accesible a nivel mundial sobre tácticas y técnicas adversas contra sistemas habilitados para IA, basadas en observaciones de ataques del mundo real y demostraciones realistas de equipos rojos de IA y grupos de seguridad.

El proyecto OWASP Top 10 para Aplicaciones de Modelos de Lenguaje Grandes (LLM) busca educar a desarrolladores, diseñadores, arquitectos, administradores y organizaciones sobre los posibles riesgos de seguridad al implementar y gestionar Modelos de Lenguaje Grandes (LLM) y aplicaciones de IA Generativa. El proyecto proporciona diversos recursos. En particular, la lista OWASP Top 10 para aplicaciones LLM, que enumera las 10 vulnerabilidades más críticas que se observan con frecuencia en estas aplicaciones, destacando su impacto potencial, facilidad de explotación y prevalencia en aplicaciones del mundo real.

Leyenda: Mitre ATLAS · Mitre ATLAS + Referencia ATT&CK · Mitre ATLAS + OWASP LLM