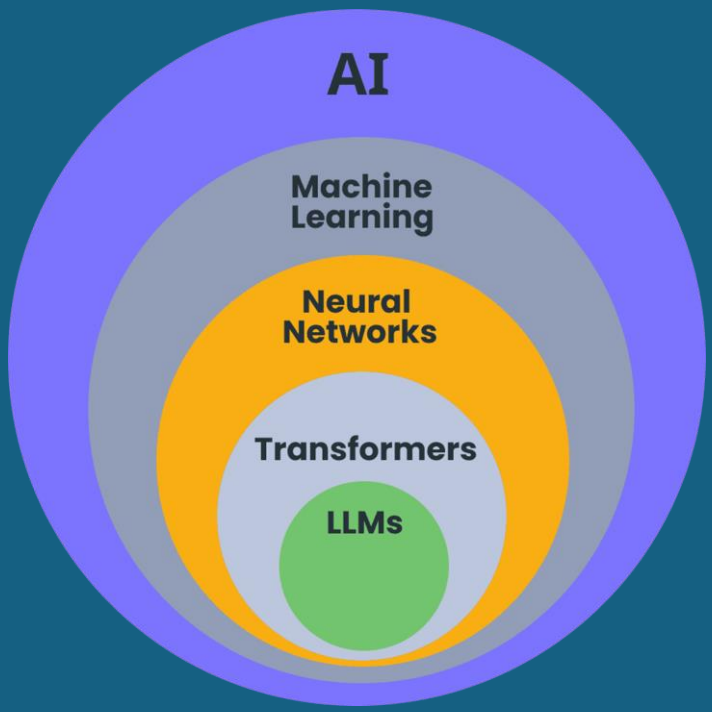
OWASP
Top 10 Vulnerabilidades en la inteligencia artificial Large Language Model (LLMs)

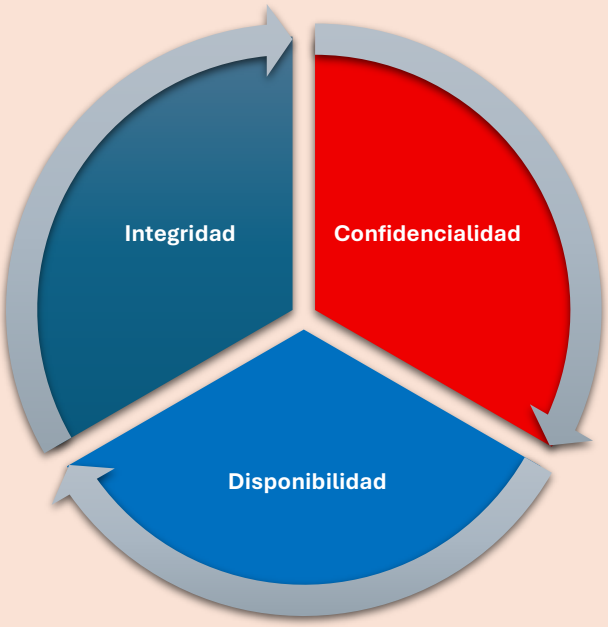


CONTENIDO

Introducción al OWASP Top 10 para LLMs 2025

El proyecto OWASP Top 10 para Aplicaciones de Modelos de Lenguaje Grande comenzó en 2023 como un esfuerzo impulsado por la comunidad para destacar y abordar problemas de seguridad específicos de aplicaciones de IA. Desde entonces, la tecnología ha continuado expandiéndose a través de industrias y aplicaciones, y con ella, los riesgos asociados.

Esta versión 2025 refleja una mejor comprensión de los riesgos existentes e introduce actualizaciones críticas sobre cómo se utilizan los LLMs en aplicaciones del mundo real hoy en día buscando preservar la confidencialidad, integridad y disponibilidad de la información.



Novedades en el Top 10 de 2025

- **Consumo Sin Límites:** Se expande a Unbounded Consumption para incluir riesgos de gestión de recursos y costos inesperados
- **Vectores y Embeddings:** Nueva categoría que responde a solicitudes de la comunidad sobre RAG y métodos basados en embeddings
- **Fuga de System Prompt:** Área con exploits del mundo real altamente solicitada por la comunidad
- **Agencia Excesiva:** Expandida dado el mayor uso de arquitecturas agénticas con mayor autonomía

Beneficios de Implementar estas Recomendaciones

La implementación de las prácticas de seguridad descritas en este documento proporciona beneficios tangibles y estratégicos:

Beneficios Técnicos

- Reducción del 70-85% en incidentes de seguridad relacionados con IA
- Detección temprana de vulnerabilidades antes de la producción
- Arquitecturas más robustas y resilientes
- Menor superficie de ataque en sistemas de IA

Beneficios Operacionales

- Disminución de costos de remediación post-incidente
- Menor tiempo de inactividad por brechas de seguridad
- Procesos de desarrollo más eficientes con seguridad integrada
- Cumplimiento simplificado con regulaciones (AI Act, GDPR)

Beneficios Estratégicos

- Confianza aumentada de clientes y stakeholders
- Ventaja competitiva en mercados regulados
- Protección de propiedad intelectual y datos sensibles
- Reputación de marca fortalecida en IA responsable

LLM01:2025 - Inyección de Prompts

Descripción

Una vulnerabilidad de inyección de prompts ocurre cuando los prompts del usuario alteran el comportamiento o salida del LLM de maneras no intencionadas. Estas entradas pueden afectar al modelo incluso si son imperceptibles para humanos, por lo que las inyecciones de prompts no necesitan ser visibles/legibles por humanos, siempre que el contenido sea analizado por el modelo.

Tipos de Inyección de Prompts

Inyecciones Directas de Prompts

Ocurren cuando la entrada de prompt de un usuario altera directamente el comportamiento del modelo de maneras no intencionadas o inesperadas. La entrada puede ser intencional (actor malicioso) o no intencional (usuario inadvertidamente).

Inyecciones Indirectas de Prompts

Ocurren cuando un LLM acepta entrada de fuentes externas, como sitios web o archivos. El contenido puede tener datos en el contenido externo que, cuando son interpretados por el modelo, alteran su comportamiento de maneras no intencionadas.

Estrategias de Prevención y Mitigación

- **1. Restringir comportamiento del modelo:** Proporcionar instrucciones específicas sobre rol, capacidades y limitaciones del modelo
- **2. Definir y validar formatos esperados:** Especificar formatos claros de salida y usar código determinístico para validar cumplimiento
- **3. Implementar filtrado de entrada y salida:** Definir categorías sensibles y construir reglas para identificar y manejar contenido
- **4. Control de privilegios y acceso mínimo:** Proporcionar tokens API propios y manejar funciones en código
- **5. Requerir aprobación humana:** Implementar controles human-in-the-loop para operaciones privilegiadas
- **6. Segregar contenido externo:** Separar claramente contenido no confiable para limitar influencia
- **7. Testing adversarial:** Realizar testing de penetración y simulaciones regulares

Beneficios de Implementación

- Protección contra manipulación maliciosa del comportamiento del LLM
- Reducción de riesgos de fuga de información sensible
- Mayor confiabilidad en salidas del modelo
- Cumplimiento con requisitos de IA confiable

LLM02:2025 - Divulgación de Información Sensible

Descripción

Los LLMs, especialmente cuando están integrados en aplicaciones, corren el riesgo de exponer datos sensibles, algoritmos propietarios o detalles confidenciales a través de su salida. Esto puede resultar en acceso no autorizado a datos, violaciones de privacidad y brechas de propiedad intelectual.

Ejemplos Comunes de Vulnerabilidad

- **1. Fuga de PII:** La información personal identificable (PII) puede ser divulgada durante interacciones con el LLM
- **2. Exposición de Algoritmos Propietarios:** Las salidas del modelo mal configuradas pueden revelar algoritmos o datos propietarios
- **3. Divulgación de Datos Empresariales:** Las respuestas generadas podrían incluir inadvertidamente información confidencial del negocio

Estrategias de Prevención

Sanitización:

- Integrar técnicas de sanitización de datos para prevenir que datos de usuario entren al modelo de entrenamiento
- Aplicar métodos estrictos de validación de entrada para detectar y filtrar datos potencialmente dañinos

Controles de Acceso:

- Aplicar controles de acceso estrictos basados en el principio de mínimo privilegio
- Limitar acceso del modelo a fuentes de datos externas

Aprendizaje Federado y Técnicas de Privacidad:

- Utilizar aprendizaje federado para minimizar recolección centralizada de datos
- Incorporar privacidad diferencial para dificultar ingeniería inversa de puntos de datos individuales

Educación del Usuario y Transparencia:

- Proporcionar orientación sobre evitar la entrada de información sensible
- Mantener políticas claras sobre retención, uso y eliminación de datos

Beneficios de Implementación

- Cumplimiento con GDPR, CCPA y regulaciones de privacidad
- Protección de secretos comerciales y propiedad intelectual
- Confianza aumentada del cliente en manejo de datos
- Reducción de riesgos de multas regulatorias (hasta €20M o 4% de ingresos globales bajo GDPR)

LLM03:2025 - Cadena de Suministro

Descripción

Las cadenas de suministro de LLM son susceptibles a varias vulnerabilidades que pueden afectar la integridad de datos de entrenamiento, modelos y plataformas de despliegue. Estos riesgos pueden resultar en salidas sesgadas, brechas de seguridad o fallas del sistema.

Ejemplos Comunes de Riesgos

- **1. Vulnerabilidades de Paquetes de Terceros:** Componentes desactualizados o obsoletos que los atacantes pueden explotar
- **2. Riesgos de Licenciamiento:** Diferentes licencias imponen requisitos legales variados
- **3. Modelos Desactualizados:** Modelos ya no mantenidos conducen a problemas de seguridad
- **4. Modelos Pre-entrenados Vulnerables:** Modelos pre-entrenados pueden contener sesgos ocultos, backdoors o características maliciosas
- **5. Procedencia Débil del Modelo:** Actualmente no hay garantías fuertes sobre el origen del modelo
- **6. Adaptadores LoRA Vulnerables:** Adaptadores LoRA maliciosos comprometen la integridad del modelo base
- **7. Desarrollo Colaborativo:** Procesos compartidos de merge de modelos pueden ser explotados
- **8. Vulnerabilidades de Cadena de Suministro en Dispositivo:** LLMs en dispositivo aumentan la superficie de ataque

Estrategias de Prevención

- Evaluar cuidadosamente fuentes de datos y proveedores, incluidos T&Cs y políticas de privacidad
- Aplicar técnicas de Hacking Ético sobre IA (AI Red Teaming) al seleccionar modelos de terceros
- Mantener inventario actualizado usando Software Bill of Materials (SBOM)
- Usar solo modelos de fuentes verificables con verificaciones de integridad de terceros
- Implementar monitoreo y auditoría estrictos para entornos de desarrollo colaborativo
- Detección de anomalías y pruebas de robustez adversaria en modelos y datos suministrados

Beneficios de Implementación

- Reducción de riesgos de backdoors y malware en modelos
- Cumplimiento de licencias y evitación de litigios de propiedad intelectual
- Mayor confiabilidad y trazabilidad de componentes de IA
- Protección contra ataques de cadena de suministro (estimados en $4.35M por brecha en promedio)

LLM04-10: Resumen de Vulnerabilidades Críticas

LLM04: Envenenamiento de Datos y Modelo

El envenenamiento de datos ocurre cuando se manipulan datos de pre-entrenamiento, fine-tuning o embedding para introducir vulnerabilidades, backdoors o sesgos.

Prevención clave:

- Rastrear orígenes de datos con OWASP CycloneDX o ML-BOM
- Verificar rigurosamente proveedores de datos
- Implementar sandboxing estricto
- Usar control de versiones de datos (DVC)

LLM05: Manejo Inadecuado de Salidas

Se refiere a validación, sanitización y manejo insuficientes de salidas generadas por LLMs antes de pasarlas a otros componentes y sistemas.

Prevención clave:

- Tratar al modelo como cualquier otro usuario con enfoque zero-trust
- Seguir directrices OWASP ASVS para validación de entrada
- Codificar salidas del modelo de vuelta a usuarios
- Usar consultas parametrizadas para operaciones de base de datos

LLM06: Agencia Excesiva

Vulnerabilidad que permite que se realicen acciones dañinas en respuesta a salidas inesperadas, ambiguas o manipuladas de un LLM.

Prevención clave:

- Minimizar extensiones que los agentes LLM pueden llamar
- Limitar funciones implementadas al mínimo necesario
- Evitar extensiones abiertas (shell commands, fetch URL)
- Requerir aprobación humana para acciones de alto impacto

LLM07: Fuga de System Prompt

Riesgo de que los prompts de sistema usados para guiar el comportamiento del modelo contengan información sensible no destinada a ser descubierta.

Prevención clave:

- Separar datos sensibles de los system prompts
- Evitar dependencia en system prompts para control estricto de comportamiento
- Implementar guardrails fuera del LLM mismo
- Asegurar que controles de seguridad se apliquen independientemente del LLM

LLM08: Debilidades en Vectores y Embeddings

Vulnerabilidades en cómo se generan, almacenan o recuperan vectores y embeddings en sistemas RAG (Retrieval Augmented Generation).

Prevención clave:

- Implementar controles de acceso granulares y permission-aware
- Validar pipelines de datos robustos
- Monitorear y logging detallado de actividades de recuperación

LLM09: Desinformación

Ocurre cuando LLMs producen información falsa o engañosa que parece creíble, incluyendo alucinaciones donde el modelo genera contenido que suena correcto pero es fabricado.

Prevención clave:

- Usar Retrieval-Augmented Generation (RAG)
- Implementar fine-tuning del modelo con datos verificados
- Fomentar verificación cruzada de salidas con fuentes confiables
- Implementar mecanismos de validación automática

LLM10: Consumo Sin Límites

Ocurre cuando una aplicación LLM permite a usuarios realizar inferencias excesivas y no controladas, llevando a riesgos como denegación de servicio (DoS), pérdidas económicas, robo de modelo y degradación de servicio.

Prevención clave:

- Implementar validación estricta de entrada
- Aplicar limitación de tasa y cuotas de usuario
- Gestionar asignación de recursos dinámicamente
- Establecer timeouts y throttling para operaciones intensivas en recursos

Beneficios Integrales de Implementar OWASP Top 10 para LLMs

Retorno de Inversión (ROI) Cuantificable

La implementación de estas prácticas de seguridad genera retornos medibles:

Reducción de Costos

- Costo promedio de brecha de datos: $4.45 millones (IBM 2023)
- Prevención de brechas relacionadas con IA puede ahorrar 70-85% de estos costos
- Reducción de costos de remediación post-incidente en 60-75%
- Menor tiempo de inactividad (promedio de 280 días reducido a 45-60 días)

Eficiencia Operacional

- 40-50% reducción en incidentes de seguridad de IA
- 30-40% mejora en tiempo de detección de vulnerabilidades
- Automatización de 60-70% de controles de seguridad
- Disminución de falsos positivos en 50-60%

Cumplimiento y Regulación

- Alineación con AI Act de la Unión Europea (obligatorio 2025-2027)
- Cumplimiento simplificado con GDPR, CCPA, y regulaciones sectoriales
- Reducción de riesgos de multas regulatorias (hasta €20M o 4% de ingresos)
- Preparación para auditorías de IA con documentación completa

Ventaja Competitiva

- Diferenciación en mercado mediante IA confiable y segura
- Mayor confianza de clientes (78% considera seguridad IA factor decisivo)
- Acceso a contratos gubernamentales y corporativos que requieren seguridad certificada
- Atracción de inversión (15-20% mayor valuación para startups con IA segura)

Roadmap de Implementación: De la Teoría a la Práctica

Fase 1: Evaluación Inicial (Semanas 1-2)

- Inventariar todos los sistemas y aplicaciones que utilizan LLMs
- Identificar superficies de ataque y puntos de integración
- Evaluar madurez actual de seguridad de IA
- Priorizar vulnerabilidades según impacto al negocio

Fase 2: Planificación y Diseño (Semanas 3-4)

- Definir políticas de seguridad específicas para IA
- Establecer arquitectura de seguridad por capas
- Seleccionar herramientas y frameworks de testing
- Formar equipo de seguridad de IA (AI Red Team)

Fase 3: Implementación de Controles Críticos (Meses 2-4)

- Implementar controles para Top 3 vulnerabilidades prioritarias
- Desplegar sistemas de monitoreo y detección
- Establecer procesos de validación de entrada/salida
- Configurar logging y auditoría completos

Fase 4: Implementación Completa (Meses 5-6)

- Extender controles a todas las 10 vulnerabilidades OWASP
- Integrar seguridad en CI/CD pipelines
- Capacitar equipos de desarrollo y operaciones
- Establecer procesos de respuesta a incidentes de IA

Fase 5: Mejora Continua (Continuo)

- Realizar testing adversarial regular
- Actualizar controles según amenazas emergentes
- Participar en comunidad OWASP para compartir conocimientos
- Medir y reportar KPIs de seguridad de IA

Conclusiones: Construyendo el Futuro de la IA Segura

Reflexiones Finales

El OWASP Top 10 para Aplicaciones LLM 2025 representa un hito fundamental en la maduración de la seguridad de IA. A medida que los modelos de lenguaje grande se integran cada vez más profundamente en sistemas críticos de negocio, la seguridad deja de ser una consideración secundaria para convertirse en un requisito fundamental.

Lecciones Clave

- **1. Enfoque Sistémico:** La seguridad de IA requiere un enfoque holístico que abarca todo el ciclo de vida
- **2. Más Allá de lo Técnico:** Los controles técnicos deben complementarse con gobernanza, capacitación y cultura organizacional
- **3. Prevención sobre Reacción:** La detección temprana ahorra millones; invertir en prevención es 10x más económico que remediación
- **4. Seguridad como Proceso:** Nuevas vulnerabilidades emergen constantemente; la vigilancia debe ser permanente
- **5. Colaboración Comunitaria:** Compartir conocimientos y experiencias fortalece la defensa colectiva

El Imperativo de Actuar Ahora

La adopción de IA está acelerándose exponencialmente. Organizaciones que implementen estas prácticas tempranamente obtendrán ventajas significativas:

- Serán considerados líderes en IA responsable y confiable
- Estarán preparados para regulaciones inminentes (AI Act, leyes sectoriales)
- Tendrán menor deuda técnica de seguridad que remediar
- Habrán desarrollado expertise interno valioso y escaso

Mirada al Futuro

La seguridad de LLMs evolucionará rápidamente. Tendencias a observar:

- **IA Defensiva:** Automatización de controles mediante IA para seguridad de IA
- **Regulación Aumentada:** Mayores requisitos de verificación y certificación de modelos
- **Estándares Emergentes:** Colaboración entre industria, academia y gobierno para estándares
- **Modelos Multimodales:** Más capacidades y mayor necesidad de controles sofisticados
- **Ecosistema de Herramientas:** Democratización de herramientas de seguridad de IA

Palabras Finales

La inteligencia artificial tiene el potencial de transformar positivamente nuestra sociedad de maneras que apenas comenzamos a imaginar. Sin embargo, para realizar este potencial de manera responsable y sostenible, debemos construir sobre fundamentos de seguridad sólidos.

Este documento es más que una lista de vulnerabilidades—es una hoja de ruta hacia la IA confiable. Cada vulnerabilidad prevenida, cada control implementado, cada desarrollador capacitado, nos acerca un paso más a un futuro donde la IA no solo sea poderosa, sino también segura, ética y digna de confianza.

Recursos Adicionales

Enlaces y Referencias Clave

- Sitio oficial del proyecto OWASP Top 10 LLM: https://genai.owasp.org
- Repositorio GitHub con herramientas y ejemplos: https://github.com/OWASP/www-project-top-10-for-large-language-model-applications
- MITRE ATLAS: Framework de tácticas adversariales en ML: https://atlas.mitre.org/
- NIST AI Risk Management Framework: https://www.nist.gov/itl/ai-risk-management-framework

Comunidad y Contribución

Este proyecto es impulsado por la comunidad. Para contribuir o participar:

- Únase a las discusiones en el canal Slack de OWASP
- Participe en reuniones mensuales del proyecto
- Contribuya con casos de estudio y ejemplos reales
- Ayude a traducir y adaptar el contenido a diferentes contextos

Agradecimientos

Este documento en español ha sido traducido y enriquecido para proporcionar valor adicional a la comunidad hispanohablante de seguridad de IA. Agradecemos al proyecto OWASP original y a todos los contribuyentes que hicieron posible esta valiosa guía.

El momento de actuar es ahora. El futuro de la IA segura comienza con las decisiones que tomamos hoy.