SEGURIDAD EN GENAI

¿POR DONDE EMPEZAR?

aws SECURITY UG

"Al 68% de los líderes empresariales les preocupa que la IA generativa sobrepase rápidamente la habilidad de su organización para entender y mitigar los riesgos asociados"

Gartner®

aws SECURITY UG

Temas de Seguridad en GenAI

Seguridad ante las amenazas que trae GenAI

¿Cómo pueden adversaries usar GenAI en su contra?

Seguridad de sus aplicaciones GenAI

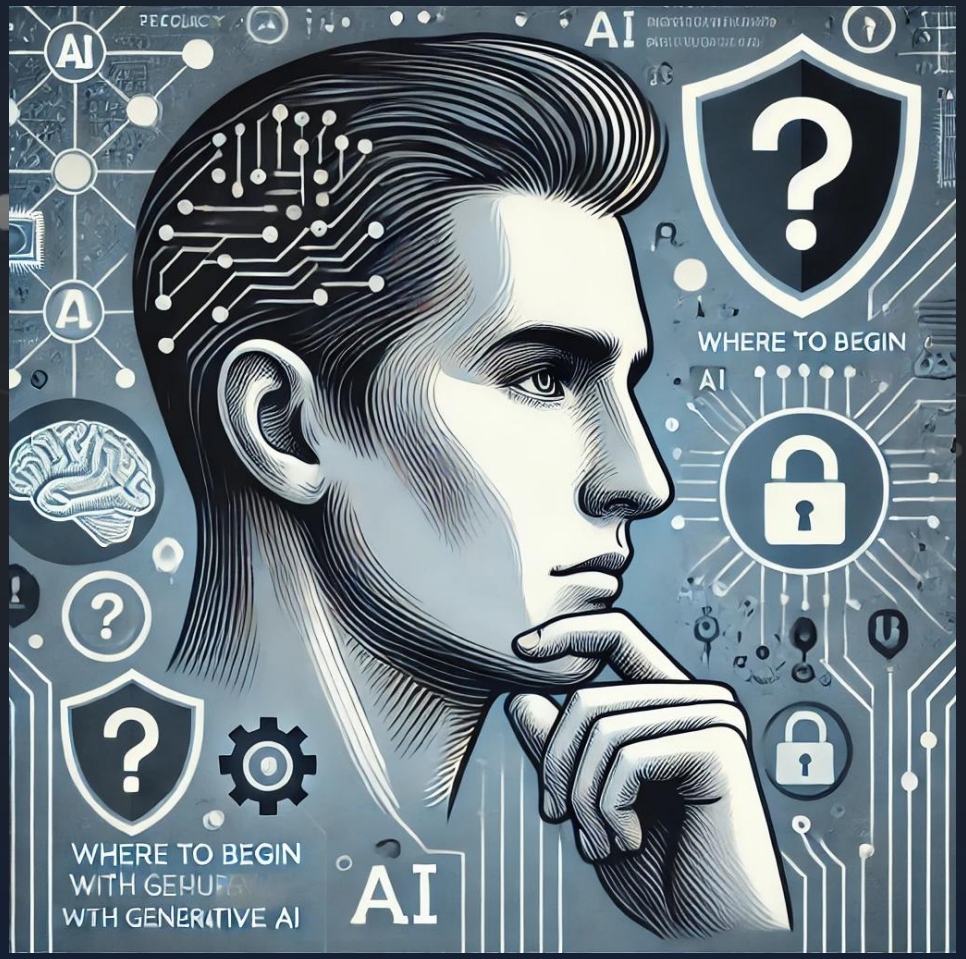
¿Cómo aseguro mis aplicaciones de negocio que usan IA Generativa?

Reforzando su Seguridad con GenAI

¿Cómo puedo usar IA generativa para reducir el esfuerzo y lograr mejores resultados?

aws SECURITY UG

Entonces ... por donde empezamos?



OWASP Top 10 para Aplicaciones de LLM

LLM01

Prompt Injection

Esta manipulación afecta a un modelo (LLM) a través de entradas astutas, provocando acciones no intencionadas por parte del LLM. Las inyecciones directas sobrescriben los prompts del sistema, mientras que las indirectas manipulan las entradas de fuentes externas.

LLM02

Insecure Output Handling

Esta vulnerabilidad ocurre cuando la salida de un LLM se acepta sin verificación, exponiendo los sistemas de backend. El mal uso puede llevar a consecuencias graves como XSS (Cross-Site Scripting), CSRF (Cross-Site Request Forgery), SSRF (Server-Side Request Forgery), escalada de privilegios o ejecución remota de código.

LLM03

Training Data Poisoning

Esto ocurre cuando se manipulan los datos de entrenamiento de un LLM, introduciendo vulnerabilidades o sesgos que comprometen la seguridad, efectividad o comportamiento ético.

LLM04

Model Denial of Service

Los atacantes provocan operaciones que consumen muchos recursos en los LLMs, lo que lleva a la degradación del servicio o a altos costos. La vulnerabilidad se magnifica debido a la naturaleza intensiva en recursos de los LLMs y la imprevisibilidad de las entradas de los usuarios.

LLM05

Supply Chain Vulnerabilities

El ciclo de vida de la aplicación de un LLM puede verse comprometido por componentes o servicios vulnerables, lo que conduce a ataques de seguridad. El uso de conjuntos de datos de terceros, modelos preentrenados y complementos puede añadir vulnerabilidades.

LLM06

Sensitive Information Disclosure

Los LLMs pueden revelar inadvertidamente datos confidenciales en sus respuestas, lo que conduce a acceso no autorizado a datos, violaciones de privacidad y brechas de seguridad. Es crucial implementar la sanitización de datos y políticas de usuario estrictas para mitigar esto.

LLM07

Insecure Plugin Design

Los complementos de LLM pueden tener entradas inseguras y control de acceso insuficiente. Esta falta de control de la aplicación los hace más fáciles de explotar y puede resultar en consecuencias como la ejecución remota de código.

LLM08

Excessive Agency

Los sistemas basados en LLM pueden realizar acciones que conducen a consecuencias no intencionadas. El problema surge del exceso de funcionalidad, permisos o autonomía otorgados a los sistemas basados en LLM.

LLM09

Overreliance

Los sistemas o personas que dependen excesivamente de los LLMs sin supervisión pueden enfrentar desinformación, mala comunicación, problemas legales y vulnerabilidades de seguridad debido al contenido incorrecto o inapropiado generado por los LLMs.

LLM10

Model Theft

Esto implica el acceso no autorizado, la copia o la exfiltración de modelos LLM propietarios. El impacto incluye pérdidas económicas, ventaja competitiva comprometida y acceso potencial a información sensible.

Source: https://owasp.org/www-project-top-10-for-large-language-model-applications/

aws SECURITY UG

Generative AI Security Scoping Matrix

UN MODELO MENTAL PARA CLASIFICAR LOS CASOS DE USO

SCOPE 1	SCOPE 2	SCOPE 3	SCOPE 4	SCOPE 5
Consumer App	**Enterprise App**	**Pre-trained Models**	**Fine-tuned Models**	**Self-trained Models**
Using 'public' generative AI services	Using an app or SaaS with generative AI features	Building your app on a versioned model	Fine-tuning a model on your data	Training a model from scratch on your data
Ex: PartyRock, ChatGPT, Midjourney	*Ex: Salesforce Einstein GPT, Amazon CodeWhisperer*	*Ex: Amazon Bedrock base models*	*Ex: Amazon Bedrock customized models, Amazon SageMaker JumpStart*	*Ex: Amazon SageMaker*

Asegurando Generative AI

Gobierno & cumplimiento | Legal & Privacidad | Gestión de riesgos | Controles | Resiliencia

https://aws.amazon.com/blogs/security/securing-generative-ai-an-introduction-to-the-generative-ai-security-scoping-matrix/

https://aws.amazon.com/blogs/security/securing-generative-ai-data-compliance-and-privacy-considerations

aws SECURITY UG

Generative AI Security Scoping Matrix

CÓMO PRIORIZAR LOS RIESGOS EN OWASP TOP 10

SCOPE 1	SCOPE 2	SCOPE 3	SCOPE 4	SCOPE 5
LLM01	LLM01	LLM01	LLM01	LLM01
LLM02	LLM02	LLM02	LLM02	LLM02
LLM03	LLM03	LLM03	LLM03	LLM03
LLM04	LLM04	LLM04	LLM04	LLM04
LLM05	LLM05	LLM05	LLM05	LLM05
LLM06	LLM06	LLM06	LLM06	LLM06
LLM07	LLM07	LLM07	LLM07	LLM07
LLM08	LLM08	LLM08	LLM08	LLM08
LLM09	LLM09	LLM09	LLM09	LLM09
LLM10	LLM10	LLM10	LLM10	LLM10

- LLM01: Prompt Injection
- LLM02: Insecure output handling
- LLM03: Training data poisoning
- LLM04: Model denial of service
- LLM05: Supply chain vulnerabilities
- LLM06: Sensitive information disclosure
- LLM07 Insecure plugin design
- LLM08: Excessive Agency
- LLM09: Overreliance
- LLM10: Model Theft

aws SECURITY UG

¿Cómo mitigar estos riesgos?

1. Mejore los programas de concienciación sobre seguridad explicando el arte de lo posible con GenAI
2. Implemente MFA / two person controls / refuerce los mecanismos de aprobación
3. Reduzca los privilegios al mínimo
4. Limite las acciones derivadas de aplicaciones GenAI para evitar exceso de confianza y agregue aprobaciones humanas

Novedades de Seguridad en AWS de las última semana

- IAM: Centralización de acceso root a la cuentas miembros de la organziación https://aws.amazon.com/blogs/aws/centrally-managing-root-access-for-customers-using-aws-organizations/
- Nuevo tipo de política de autorización – Resource Control Policy – RCP https://aws.amazon.com/blogs/aws/introducing-resource-control-policies-rcps-a-new-authorization-policy/
- Amazon VPC Block Public Access https://aws.amazon.com/blogs/networking-and-content-delivery/vpc-block-public-access/
- Customización de scopes en IAM Access Analyzer unused access analysis https://aws.amazon.com/about-aws/whats-new/2024/11/customize-scope-iam-access-analyzer-unused-access-analysis/

aws SECURITY UG

MUCHAS GRACIAS