Dirección de Investigación Criminal e INTERPOL
Centro Cibernético Policial

BALANCE 24
2025

Cyber Noticias

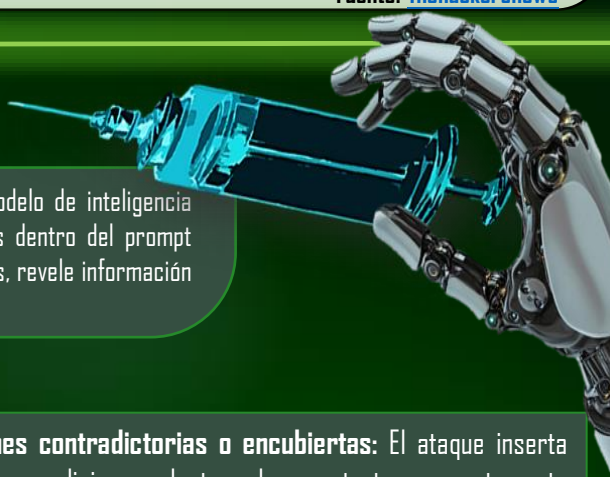
Todos los principales modelos de Gen-AI son vulnerables a los ataques de inyección inmediata. Esta técnica universal de inyección rápida puede romper las barreras de seguridad de todos los principales modelos de AI generativa. Estos modelos están entrenados para rechazar solicitudes de usuarios que indaguen sobre resultados perjudiciales , creación de armas, autolesiones y violencia. Sin embargo, investigaciones han demostrado que es posible utilizando diversas técnicas de ingeniería rápida para explotar la IA con fines nefastos.

Fuente: securityweek

Una vulnerabilidad en GitLab Duo permitió a los atacantes secuestrar respuestas de IA con indicaciones ocultas. Investigadores en ciberseguridad, descubrieron la falla de inyección directa de mensajes en el asistente de inteligencia artificial IA Duo de GitLab. Esto permite a los atacantes robar código de la fuente de proyectos privados, manipulación de respuestas y filtración de vulnerabilidades, además esto provocaría que la víctima sea redirigida a paginas de inicio de sesión fraudulentas, con el fin de hurtar sus credenciales e información personal .

Fuente: Thehackersnews

AI PROMPT INJECTION ATTACK



Es una técnica utilizada para manipular o controlar el comportamiento de un modelo de inteligencia artificial basado en lenguaje, introduciendo instrucciones maliciosas o engañosas dentro del prompt (entrada de texto). El objetivo es hacer que el modelo genere respuesta no deseadas, revele información confidencial o ejecute acciones fuera de su diseño o propósito.

Patrones de conducta

Instrucciones contradictorias o encubiertas: El ataque inserta instrucciones maliciosas dentro de un texto aparentemente inocente para confundir el modelo de inteligencia.

Inyección indirecta (promt aninado): El texto malicioso se introduce a través del contenido que el modelo debe procesar (como un correo, código, documento, etc.).

Manipulación de formato o lenguaje: Utiliza lenguaje ambiguo, símbolos o sintaxis especifica para evadir filtros o restricciones de seguridad.

Ataque de inyección persistente: El contenido malicioso se guarda en una fuente que será consultada repetidamente por el modelo, como bases de datos, wikis, o contenido generado por usuarios.

RECOMENDACIONES

- **Evite** compartir contraseñas, datos bancarios, documentos oficiales o información confidencial en chats automatizados.
- **Proteja** la entrada del modelo, mediante sistemas que detecten frases sospechosas, patrones manipulativos.
- **Establezca** reglas y restricciones sobre las instrucciones o preguntas que el modelo pueda o no responder.
- **Reporte** cualquier incidente ocurrido al CAI Virtual, a través del portal web https://caivirtual.policia.gov.co

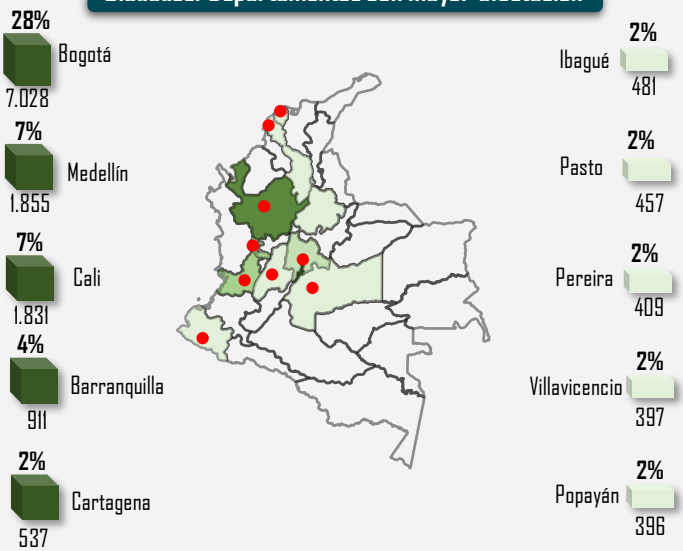
Balance Cibercriminalidad – Ley 1273/09

Variación Porcentual	2024	2025	
-31%	17.642	12.013	Hurto por medios informáticos y semejantes
10%	5.527	6.118	Acceso abusivo a un sistema informático
-9%	4.264	3.872	Violación de datos personales
-27%	2.406	1.738	Suplantación de sitios web
-31%	1.644	1.129	Transferencia no consentida de activos
-35%	403	261	Interceptación de datos informáticos
8%	166	180	Obstaculización ilegitima de sistema informático o red de comunicación
-35%	146	94	Daño informático
-23%	105	80	Uso de software malicioso
TOTAL	**32.303**	**25.485**	
-21 %			

Corte del 01 de enero al 13 de junio del 2024.
vs del 01 de enero al 13 de junio del 2025.

Evolución histórica *(No. Denuncias vs año)*

Ciudades/Departamentos con mayor afectación



Estas ciudades y departamentos representan el **58%** del total del fenómeno a nivel nacional.

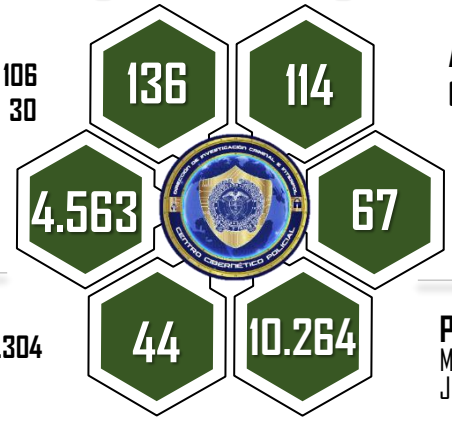
Datos extraidos el dia 13 de junio 2025. Cifras sujetas a variación en atención al proceso de integración y consolidación con la información de la Fiscalía General de la Nación.

Fuente: SIEDCO Plus 2.0.

Actividades de gestión en seguridad digital

Capturas ESCIB
Delitos informáticos **106**
Explotación sexual infantil en Internet. **30**

Incidentes gestionados
a través del CAI Virtual



Alertas y contenidos preventivos
04 durante la semana

Actividades de relacionamiento estratégico

Charlas de ciberseguridad
Personas impactadas **6.304**

Páginas bloqueadas
Material de abuso sexual infantil. **5.153**
Juegos ilegales de azar. **5.111**

Canales de atención y redes sociales

Página web

X
@CaiVirtual

Instagram
@caivirtual

Facebook
CAI Virtual

WhatsApp

Para sugerencias sobre este producto, por favor diligencie este formulario: https://bit.ly/3mz5C8d