



NARSIL

INTELLIGENCE

ESTUDIOS NARSIL · 02

Tu asistente puede trabajar para otro

La inyección de instrucciones: cómo un correo, una web o un documento pueden dar órdenes a tu IA, qué casos reales se conocen y cómo se limita el daño

OCTUBRE DE 2026

Resumen

1. **La IA no distingue con fiabilidad entre datos y órdenes.** Un texto escondido en un correo, una web o un documento puede darle instrucciones, y el asistente puede obedecerlas. Es la inyección de instrucciones, primer riesgo de la lista de OWASP para aplicaciones con modelos de lenguaje.
2. **Ya hay casos reales en productos de uso diario.** Microsoft 365 Copilot (EchoLeak), Gemini con invitaciones de calendario, navegadores con IA como Comet y agentes de programación en GitHub.
3. **Los fabricantes no lo dan por resuelto.** OpenAI, Anthropic, Microsoft, Meta y el centro de ciberseguridad británico coinciden: se puede reducir, no eliminar.
4. **Los filtros no bastan.** Un estudio de 2025 superó la mayoría de doce defensas publicadas con más de un 90 % de éxito.
5. **El riesgo crece con lo que el agente puede hacer.** Un asistente que lee datos privados, procesa contenido de terceros y puede enviar algo fuera reúne la combinación más peligrosa.
6. **La defensa es de diseño.** Mínimos permisos, contenido externo tratado como no fiable, confirmación humana para lo irreversible y registro de todo lo que hace el agente.

01

Qué es la inyección de instrucciones

Un asistente de IA recibe dos tipos de texto: las instrucciones de quien lo usa y el contenido que procesa para cumplirlas, sea un correo, una página web, un documento o el resultado de una herramienta. El problema es que, para el modelo, todo es texto. No existe una separación técnica fiable entre lo que debe obedecer y lo que solo debe leer.

La **inyección de instrucciones** aprovecha esa debilidad. OWASP, la organización de referencia en seguridad de aplicaciones, la sitúa en el primer puesto de su lista de riesgos de 2025 para aplicaciones con modelos de lenguaje y distingue dos variantes. En la **directa**, es el propio usuario quien intenta manipular el modelo con lo que escribe. En la **indirecta**, la más peligrosa para una empresa, las órdenes llegan escondidas en fuentes externas: un correo, una web, un fichero. El usuario no ve nada raro; el asistente lee las instrucciones y las ejecuta.

El NIST estadounidense describe una variante concreta, el **secuestro de agentes**: un atacante coloca instrucciones maliciosas en datos que un agente de IA va a leer, y consigue que realice acciones dañinas que nadie le había pedido. Y señala la causa de fondo: la falta de una separación clara entre las instrucciones de confianza y los datos externos que no lo son.

02

Por qué ahora importa más

Mientras un asistente solo respondía preguntas, una inyección podía, como mucho, hacerle decir algo incorrecto. Hoy los asistentes leen el correo, consultan documentos internos, navegan, rellenan formularios, ejecutan código y envían mensajes. Cada nueva capacidad convierte una manipulación del texto en una acción en el mundo real.

Simon Willison, uno de los analistas más citados en este campo, resumió en 2025 la combinación que convierte un agente en un riesgo serio y la llamó **trifecta letal**: acceso a datos privados, exposición a contenido que puede controlar un tercero y capacidad de comunicarse hacia fuera. Un agente que reúne las tres está a un mensaje de enviar información confidencial a quien no debe. Meta formuló poco después una regla práctica en la misma línea: que un agente no combine nunca más de dos de esas capacidades sin supervisión.

03

Casos reales

EchoLeak: un correo contra Microsoft 365 Copilot

En enero de 2025, el equipo de Aim Security comunicó a Microsoft una vulnerabilidad que permitía, con un solo correo manipulado y sin que el usuario hiciera nada, que Copilot extrajera información de la organización. Microsoft la corrigió en sus servidores en mayo, sin que los clientes tuvieran que intervenir, y la hizo pública el 11 de junio de 2025 como CVE-2025-32711, con una puntuación crítica de 9,3 sobre 10 según la propia Microsoft. No consta que se explotara antes de corregirse.

Una invitación de calendario contra Gemini

En agosto de 2025, investigadores de SafeBreach presentaron en Black Hat «Invitation Is All You Need»: instrucciones escondidas en el título de un evento de Google Calendar que Gemini ejecutaba cuando el usuario le preguntaba por su agenda. Habían avisado a Google en febrero. La compañía respondió con varias capas de defensa, entre ellas confirmaciones del usuario para acciones sensibles, saneamiento de enlaces y clasificadores para detectar inyecciones.

Navegadores con IA

En agosto de 2025, Brave publicó cómo un comentario de Reddit con instrucciones ocultas conseguía que Comet, el navegador con IA de Perplexity, al resumir la página, entrara en la cuenta del usuario, leyera un código de un solo uso en su correo y lo enviara fuera. Perplexity aplicó correcciones, aunque Brave advirtió después de que el ataque no estaba mitigado del todo. En octubre, Brave documentó variantes con instrucciones casi invisibles en capturas de pantalla y en otros navegadores, y concluyó que no es un problema aislado, sino un desafío sistémico para toda la categoría. Opera corrigió una vulnerabilidad similar en su navegador Neon ese mismo mes.

OpenAI, por su parte, explicó en diciembre de 2025 cómo había reforzado su navegador con IA tras encontrar, con su propio atacante automatizado, un caso ilustrativo: un correo malicioso hacía que el agente, en lugar de redactar una respuesta de ausencia, enviara una carta de dimisión al consejero delegado del usuario.

Agentes de programación

En mayo de 2025, Invariant Labs mostró que una incidencia maliciosa en un repositorio público de GitHub podía secuestrar a un agente conectado al servidor MCP de GitHub y llevarlo a filtrar datos de repositorios privados. Su conclusión fue que no se trataba de un fallo del código de ese servidor, sino de un problema de arquitectura que no se resuelve con un parche.

En febrero de 2026, el investigador Adnan Khan hizo público el caso conocido como Clinejection: el título de una incidencia de GitHub se insertaba directamente en las instrucciones de un agente de IA que clasificaba incidencias del proyecto Cline, y con ello se le podían hacer ejecutar órdenes. Encadenado con otra debilidad, permitió robar credenciales de publicación, y poco después un atacante publicó una versión manipulada de la extensión, que fue retirada. En mayo de 2026, Microsoft explicó cómo una vulnerabilidad en su marco Semantic Kernel convertía una inyección en ejecución de código en el equipo: según la compañía, una sola instrucción bastó para lanzar un programa en la máquina donde corría el agente.

04

Lo que admiten fabricantes y agencias

Pocas veces un sector reconoce con tanta claridad que tiene un problema sin solución completa.

«La inyección de instrucciones, igual que las estafas y la ingeniería social en la web, difícilmente llegará a estar resuelta del todo.»

OpenAI, diciembre de 2025 (traducción propia)

«La inyección de instrucciones está lejos de ser un problema resuelto, sobre todo a medida que los modelos realizan más acciones en el mundo real.»

Anthropic, noviembre de 2025 (traducción propia)

Microsoft la describe como un riesgo inherente a cómo funcionan los modelos de lenguaje, y reconoce que detectarla de forma determinista sigue siendo un problema de investigación abierto. Meta la califica de debilidad fundamental y no resuelta de todos los modelos. El centro nacional de ciberseguridad del Reino Unido (NCSC) fue más allá en diciembre de 2025: como no hay una distinción intrínseca entre datos e instrucciones, es muy posible que nunca se mitigue del todo, como sí se consiguió con la inyección SQL, y seguirá siendo un riesgo residual que no se elimina comprando un

producto. En mayo de 2026, las agencias de ciberseguridad de Estados Unidos, el Reino Unido, Australia, Canadá y Nueva Zelanda publicaron una guía conjunta que pide a las organizaciones asumir que los sistemas de IA agéntica pueden comportarse de forma inesperada.

05

Las cifras

Los datos disponibles proceden de pruebas controladas, no de incidentes reales, pero apuntan en la misma dirección. En enero de 2025, el NIST probó un agente basado en un modelo comercial con un banco de pruebas público: los ataques nuevos que diseñó elevaron la tasa de éxito del 11 % del mejor ataque conocido al 81 %, y la tasa media pasó del 57 % al 80 % cuando el atacante repetía los intentos.

Los fabricantes miden avances reales. Anthropic informó en 2025 de que las defensas de su extensión para Chrome redujeron el éxito de los ataques de su batería de pruebas del 23,6 % al 11,2 %, y en noviembre situó en torno al 1 % el éxito de un atacante adaptativo contra su modelo más reciente, con la advertencia expresa de que un 1 % sigue siendo un riesgo significativo.

El trabajo más incómodo para las defensas basadas en filtros es «The Attacker Moves Second», publicado en octubre de 2025 por un grupo de investigadores entre los que figuran Milad Nasr, Nicholas Carlini y Florian Tramèr. Pusieron a prueba doce defensas recientes, la mayoría de las cuales había publicado tasas de éxito de los ataques cercanas a cero, y las superaron con más de un 90 % de éxito en la mayoría de los casos. La lección: una defensa que solo se ha probado contra ataques conocidos dice poco de cómo resistirá a un atacante que se adapta.

06

Cómo se limita el daño

Si el modelo no puede distinguir con fiabilidad entre órdenes y datos, la protección no puede depender solo de él. Las recomendaciones de OWASP, el NCSC, los fabricantes y las agencias coinciden en un enfoque de diseño:

1. **Mínimo privilegio.** Cada agente, con las herramientas, accesos y datos imprescindibles para su tarea, y ni uno más. Un asistente que resume correos no necesita poder enviarlos.
2. **No reunir las tres capacidades.** Si un agente procesa contenido de terceros y accede a datos sensibles, no debería poder comunicarse hacia fuera sin supervisión.
3. **Contenido externo marcado como no fiable.** Separar e identificar lo que viene de fuera ayuda al modelo, aunque no basta por sí solo.
4. **Límites deterministas.** Controles fuera del modelo que restringen lo que el sistema puede hacer, decida lo que decida: listas de destinos permitidos, bloqueo de envíos de datos, cuotas.

5. **Confirmación humana para lo irreversible.** Enviar, pagar, borrar o compartir fuera de la organización debe pasar por una persona.
6. **Sin sesión iniciada cuando no haga falta.** Un agente que navega sin las credenciales del usuario tiene mucho menos que perder.
7. **Registro y vigilancia.** Todas las entradas, salidas y llamadas a herramientas, registradas y revisables.
8. **Pruebas adversarias** antes de desplegar y cada vez que cambie el sistema, con atacantes que se adapten.

SOBRE LOS DETECTORES

Los clasificadores que detectan inyecciones son útiles como una capa más, pero los resultados de 2025 muestran que un atacante decidido los supera. No deben ser la única barrera.

07

Qué debería hacer una empresa

Lo primero es saber qué hay. Muchas organizaciones tienen ya asistentes conectados al correo, al almacenamiento en la nube o a herramientas de gestión sin que nadie haya revisado qué pueden leer y qué pueden hacer. Un inventario de agentes, extensiones y conectores, con sus permisos, es el punto de partida.

Después, aplicar a cada uno cinco preguntas: qué datos puede leer, qué contenido de terceros procesa, si puede enviar algo fuera, qué acciones exigen confirmación y dónde queda registrado lo que hace. Con esas respuestas se recortan permisos, se separan funciones y se decide qué no debe automatizarse todavía. Los navegadores con IA merecen especial cautela en equipos corporativos: navegan con la sesión del usuario por sitios que nadie controla.

La inyección de instrucciones no es un motivo para renunciar a los agentes, sino para diseñarlos como se diseña cualquier sistema que maneja información sensible: suponiendo que alguna vez algo intentará engañarlo, y asegurando que, si lo consigue, el daño sea pequeño y quede registrado.

Anexo

Fuentes

Páginas consultadas el 2 de octubre de 2026, salvo las noticias, fechadas en su publicación.

1. OWASP. *LLM01:2025 Prompt Injection*. genai.owasp.org/llmrisk/llm01-prompt-injection
2. NIST CAISI. *Technical Blog: Strengthening AI Agent Hijacking Evaluations (17-1-2025)*. nist.gov
3. NVD. *CVE-2025-32711 (11-6-2025)* y Microsoft MSRC. msrc.microsoft.com/update-guide/vulnerability/CVE-2025-32711
4. Reddy, P. y Gujral, A. S. *EchoLeak: The First Real-World Zero-Click Prompt Injection Exploit in a Production LLM System*. arxiv.org/abs/2509.10540
5. SafeBreach. *Invitation Is All You Need (6-8-2025)*. safebreach.com/blog/invitation-is-all-you-need-hacking-gemini
6. Brave. *Agentic Browser Security: Indirect Prompt Injection in Perplexity Comet (20-8-2025)* y *Unseeable prompt injections (21-10-2025)*. brave.com/blog
7. Invariant Labs. *GitHub MCP Exploited (26-5-2025)*. invariantlabs.ai/blog/mcp-github-vulnerability
8. OpenAI. *Continuously hardening ChatGPT Atlas against prompt injection (22-12-2025)*. openai.com/index/hardening-atlas-against-prompt-injection
9. Adnan Khan. *Clinejection (9-2-2026)*. adnanthekhan.com/posts/clinejection; Simon Willison (6-3-2026).
10. Microsoft Security. *When prompts become shells (7-5-2026)*. microsoft.com/en-us/security/blog
11. Anthropic. *Mitigating the risk of prompt injections in browser use (24-11-2025)* y *Claude for Chrome (2025)*. anthropic.com
12. Microsoft MSRC. *How Microsoft defends against indirect prompt injection attacks (29-7-2025)*.
13. Meta. *Agents Rule of Two*. ai.meta.com/blog/practical-ai-agent-security
14. NCSC. *Prompt injection is not SQL injection (it may be worse) (8-12-2025)*. ncsc.gov.uk
15. CISA, NSA y socios. *Careful Adoption of Agentic AI Services (1-5-2026)*. cisa.gov
16. Simon Willison. *The lethal trifecta for AI agents (16-6-2025)*. simonwillison.net
17. Nasr, M. et al. *The Attacker Moves Second (10-10-2025)*. arxiv.org/abs/2510.09023



NARSIL INTELLIGENCE

Diseñamos, construimos e implantamos sistemas de inteligencia artificial en la infraestructura de cada organización, y formamos a los equipos que los usan.

NARSILINTELLIGENCE.COM