DIARIO DIGITAL DEV / RECURSO DE LECTURA

Prompt injection: cuando un documento intenta dar órdenes a la IA

Un agente puede leer un archivo válido que intenta cambiar su tarea. Esta guía explica cómo separar información de autorización y comprobar qué ocurre cuando la IA propone una acción fuera de alcance.

Ideas clave del artículo

- Un documento puede aportar datos sin tener autoridad para cambiar la tarea o sus destinatarios.

- Comprueba la acción que ejecuta la herramienta y los permisos que aplica la integración.

- Ensaya desvíos con información ficticia y verifica también que el trabajo autorizado siga funcionando.

Notas para tu equipo

_____

_____

_____

_____

Artículo y fuentes

https://www.digital-dev.cl/blog/prompt-injection-agentes-ia-acciones-empresas/

OWASP Gen AI Security Project: LLM01:2025 Prompt Injection
https://genai.owasp.org/llmrisk/llm01-prompt-injection/

Anthropic Engineering: How we contain Claude across products
https://www.anthropic.com/engineering/how-we-contain-claude

Anthropic · Claude Platform Docs: Mitigate jailbreaks and prompt injections
https://platform.claude.com/docs/en/test-and-evaluate/strengthen-guardrails/mitigate-jailbreaks