Antes de que tu app hecha con IA atienda a un cliente

Una extensión falsa de la herramienta con la que hoy se construyen estas apps estuvo unas once horas a la vista de todos, y estas son las preguntas que hago antes de poner a atender a una app o una automatización hecha con IA, las mismas para una pyme que para un banco.

Por Oscar Obando Chaves · Chief AI Officer · Legal AI Engineer · Noviembre 2026 · Documento de distribución libre · Publicado originalmente en LinkedIn

Once horas en el marketplace de Cursor

Dos versiones falsas de la extensión de Claude Code estuvieron unas once horas en el marketplace de Cursor, mandando datos de los equipos donde se instalaron e intentando abrir una puerta trasera, según el aviso de seguridad que Cursor publicó el 29 de octubre de 2026. Microsoft confirmó al día siguiente que nunca llegaron a su propio marketplace.

Lo que más me llama la atención es la ruta, puesto que el riesgo llegó como actualización de una herramienta de confianza, la misma con la que hoy se construye una app con IA en semanas. La misma semana Cloudflare contó que 155 respuestas de un modelo de su plataforma llevaron datos de solicitudes de otros clientes, y Warp reconoció que, por un proveedor suyo, se accedió a trazas de sus agentes de hasta el 1 % de sus usuarios.

Las preguntas que hago antes de que una app o una automatización hecha con IA atienda a un cliente van desde dónde están las claves hasta quién aprueba una acción que no tiene vuelta atrás y qué pasa si falla un viernes en la noche. Son las mismas para una pyme que para un banco, y casi ninguna es técnica.

Fuentes: Cursor, aviso de seguridad (29 de octubre de 2026); Microsoft Threat Intelligence (30 de octubre de 2026); Cloudflare, informe del incidente de Workers AI (26 de octubre de 2026); Warp, aviso de incidente (26 de octubre de 2026).

La lista de riesgos de 2026

La lista de riesgos de las apps con modelos de lenguaje que el OWASP GenAI Security Project anunció el 1 de septiembre de 2026 abre con la inyección de instrucciones, la filtración de información sensible y la agencia excesiva, y sus primeros cuatro riesgos, contados para una

empresa, se leen así.

1. Prompt Injection (LLM01), que alguien le escriba a tu app lo que no debía obedecer, de frente o escondido en un correo o en un PDF.

2. Sensitive Information Disclosure (LLM02), que tu app muestre datos de tus clientes o de tu empresa a quien no debía verlos.

3. Excessive Agency (LLM03), que la IA tenga más permisos de los que su tarea necesita.

4. Supply Chain (LLM04), que el riesgo entre por un modelo, una extensión o una librería de la que depende tu app.

El Centro Criptológico Nacional de España, en la metodología CCN-STIC-2011 que publicó el 29 de octubre de 2026 para evaluar la seguridad de productos con IA, puso como principios la supervisión humana antes de ejecutar acciones críticas o irreversibles y la trazabilidad para reconstruir después la actuación de la IA.

Fuentes: OWASP GenAI Security Project, Top 10 para aplicaciones con LLM de 2026 (anunciado el 1 de septiembre de 2026); Centro Criptológico Nacional de España, metodología CCN-STIC-2011 (29 de octubre de 2026).

Las claves y los permisos

La primera pregunta es dónde están las claves de tus proveedores, sean las del modelo, el banco, el correo o la base de datos, que no deberían vivir en el código ni en un chat, quién puede verlas y cómo se cambian el día que se filtran.

La segunda es quién puede hacer qué dentro de la app, con cada usuario y cada agente con el permiso justo. Las acciones que no tienen vuelta atrás, como pagar, borrar, enviar o firmar, necesitan una persona que las apruebe antes de que ocurran, que es lo que pide el principio de supervisión humana del Centro Criptológico Nacional.

Fuentes: Centro Criptológico Nacional de España, metodología CCN-STIC-2011 (29 de octubre de 2026).

Lo que entra y lo que sale

La tercera pregunta es qué lee tu app y qué muestra, puesto que un texto escondido en un correo o en un PDF puede ordenarle lo que no debía hacer, y una mala configuración puede mostrarle a un cliente lo que escribió otro.

Eso último le pasó a Cloudflare, que informó el 26 de octubre de 2026 que 155 respuestas de un modelo de su plataforma Workers AI llevaron datos de solicitudes de otros clientes, por una configuración propia y un error del motor de código abierto que usaba.

Fuentes: Cloudflare, informe del incidente de Workers AI (26 de octubre de 2026).

Los proveedores de tus proveedores

La cuarta pregunta mira el modelo, las extensiones y las librerías de las que depende tu app, y lo que dice tu contrato con el proveedor del modelo sobre tus datos, los cambios de modelo y la responsabilidad.

Le pasó a Warp, que informó el 26 de octubre de 2026 que, por un proveedor de su infraestructura, se accedió a trazas de sus agentes y a telemetría de hasta el 1 % de sus usuarios.

Fuentes: Warp, aviso de incidente (26 de octubre de 2026).

Los registros, el viernes en la noche y tus clientes

La quinta pregunta es por los registros que permiten reconstruir qué hizo la IA, con qué datos y por orden de quién, y la sexta es qué pasa si la app falla un viernes en la noche, quién se entera, a quién avisa y en cuánto tiempo.

Frente a tus clientes cuenta lo que dicen los términos con los que tu app los atiende, que es donde se ve quién responde. Una app que nadie revisó igual tiene un responsable, que es la empresa que la puso a atender, con una pyme o con un banco detrás.

Fuentes: Centro Criptológico Nacional de España, principio de trazabilidad de la metodología CCN-STIC-2011 (29 de octubre de 2026).

La supervisión humana antes de ejecutar acciones críticas o irreversibles.

Principio de la metodología CCN-STIC-2011 del Centro Criptológico Nacional de España (29 de octubre de 2026)

Oscar Obando Chaves es Chief AI Officer y Legal AI Engineer ecuatoriano, con formación en derecho. Ayuda a las empresas a decidir cómo usan la IA y construye con ella sistemas para el negocio y para lo legal. Chief AI Officer y Senior AI Counsel en HEKA Law, ISO 42001 AI Management Leader e ISO 38507 AI Governance Leader. Publica análisis semanal sobre el impacto real de la IA.