¿Usar IA es confidencial?

El laberinto de los depende en la confidencialidad en el uso de IA

Asociación de Internet MX.

Comité Jurídico y Relaciones con Gobierno AIMX

Aldo Ricardo Rodríguez

CEO de Lawgic AI

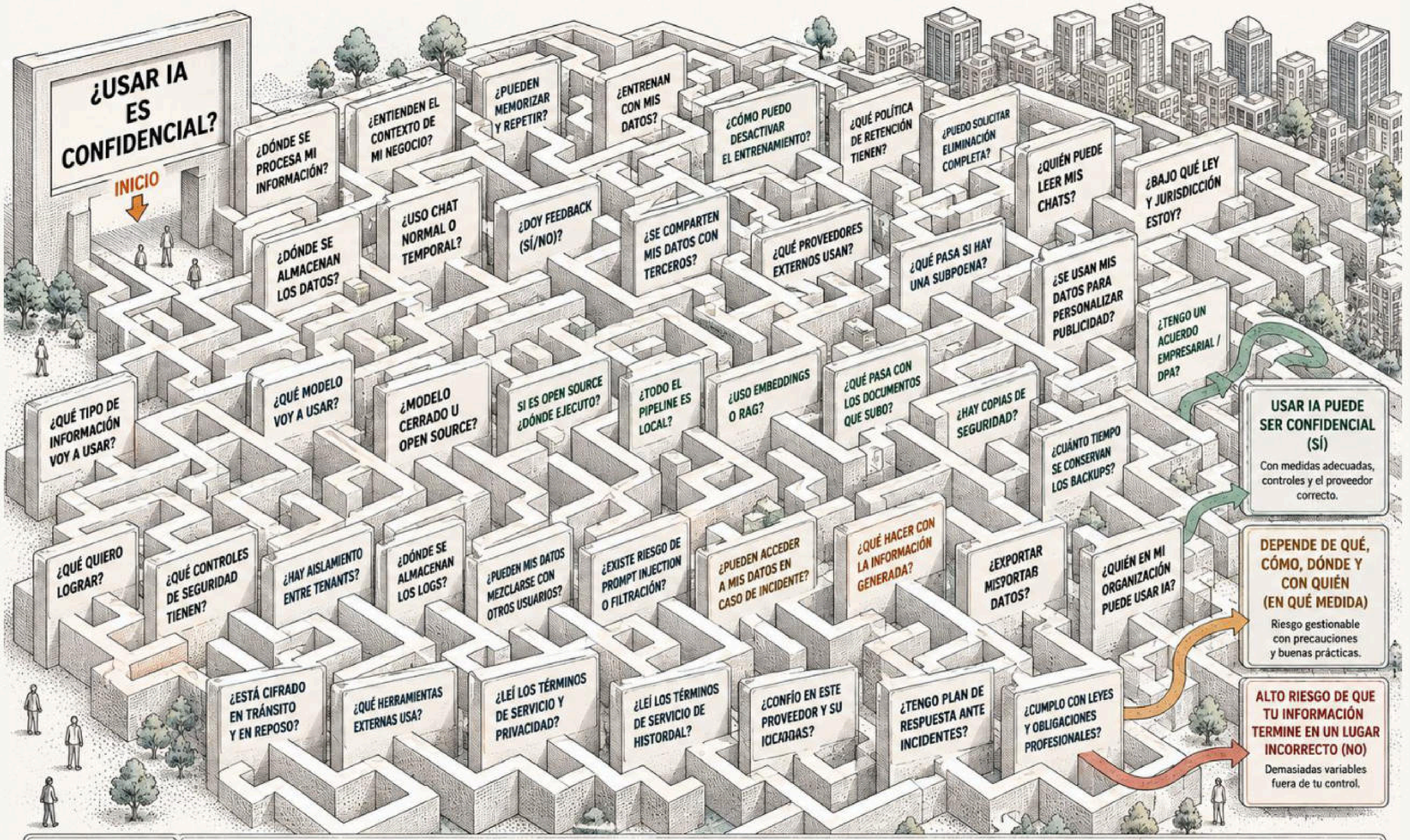
Agenda:

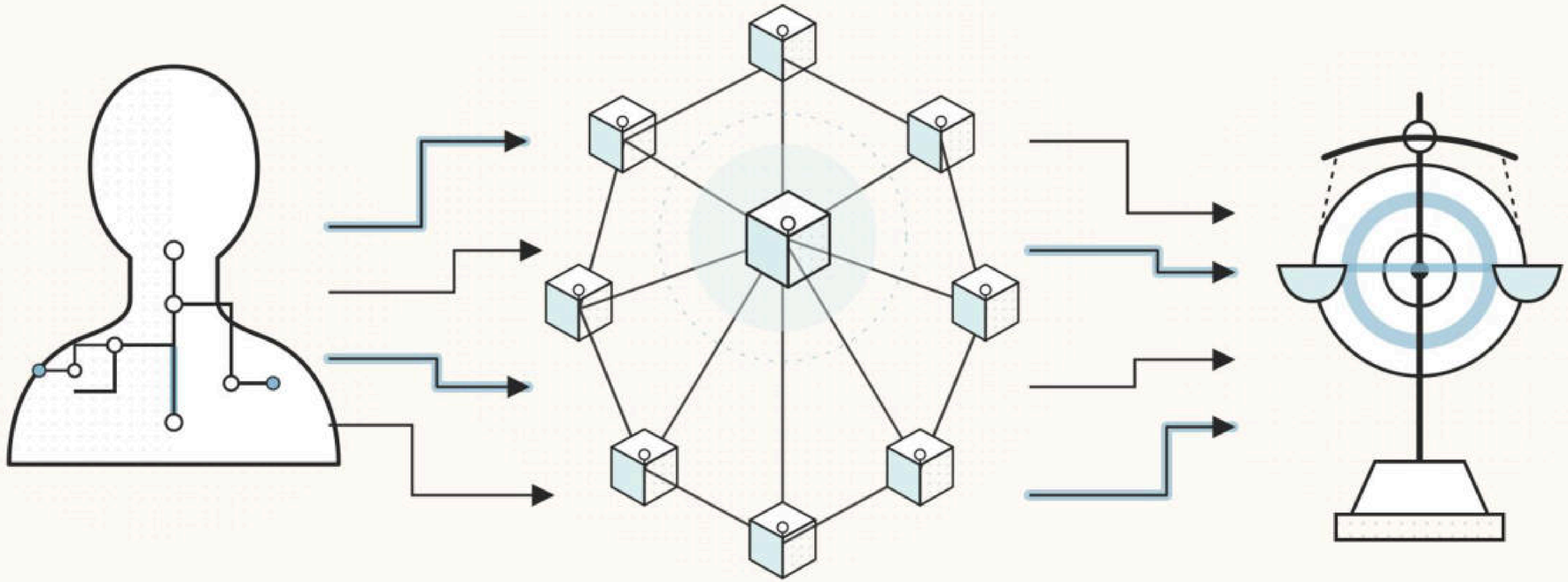
- Plática: "Confidencialidad en el uso de IA."

Jueves 13 de agosto 2026 | 10:00 hrs

Acerca de mi

- 🇺🇸 Vivo en SF
- 🌿 Fundador de Lawgic
- 🤖 Especialista en PI, UX/UI e I.A. Aplicada
- 👨🏻‍💻 Paso mucho tiempo en Hackathones
- 💻 Desarrollo Software Legal
- ⚙️ Participo en procesos regulatorios
- ⚖️ Posteo en Linkedin de IA y Derecho (En especial PI)
- 🛡️ Oficial de Cumplimiento en IA


¿USAR IA ES CONFIDENCIAL?
INICIO
¿QUÉ TIPO DE INFORMACIÓN VOY A USAR?
¿QUÉ QUIERO LOGRAR?
¿DÓNDE SE PROCESA MI INFORMACIÓN?
¿ENTIENDEN EL CONTEXTO DE MI NEGOCIO?
¿QUÉ MODELO VOY A USAR?
¿DÓNDE SE ALMACENAN LOS DATOS?
¿USO CHAT NORMAL O TEMPORAL?
¿MODELO CERRADO U OPEN SOURCE?
¿QUÉ CONTROLES DE SEGURIDAD TIENEN?
¿HAY AISLAMIENTO ENTRE TENANTS?
¿ESTÁ CIFRADO EN TRÁNSITO Y EN REPOSO?
¿QUÉ HERRAMIENTAS EXTERNAS USA?
¿PUEDEN MEMORIZAR Y REPETIR?
¿ENTRENAN CON MIS DATOS?
¿DOY FEEDBACK (SÍ/NO)?
SI ES OPEN SOURCE ¿DÓNDE EJECUTO?
¿DÓNDE SE ALMACENAN LOS LOGS?
¿PUEDEN MIS DATOS MEZCLARSE CON OTROS USUARIOS?
¿LEÍ LOS TÉRMINOS DE SERVICIO Y PRIVACIDAD?
¿CÓMO PUEDO DESACTIVAR EL ENTRENAMIENTO?
¿SE COMPARTEN MIS DATOS CON TERCEROS?
¿TODO EL PIPELINE ES LOCAL?
¿EXISTE RIESGO DE PROMPT INJECTION O FILTRACIÓN?
¿LEÍ LOS TÉRMINOS DE SERVICIO DE HISTORDAL?
¿QUÉ POLÍTICA DE RETENCIÓN TIENEN?
¿QUÉ PROVEEDORES EXTERNOS USAN?
¿USO EMBEDDINGS O RAG?
¿PUEDEN ACCEDER A MIS DATOS EN CASO DE INCIDENTE?
¿CONFÍO EN ESTE PROVEEDOR Y SU IOCADAS?
¿PUEDO SOLICITAR ELIMINACIÓN COMPLETA?
¿QUÉ PASA SI HAY UNA SUBPOENA?
¿QUÉ PASA CON LOS DOCUMENTOS QUE SUBO?
¿QUÉ HACER CON LA INFORMACIÓN GENERADA?
¿TENGO PLAN DE RESPUESTA ANTE INCIDENTES?
¿QUIÉN PUEDE LEER MIS CHATS?
¿SE USAN MIS DATOS PARA PERSONALIZAR PUBLICIDAD?
¿HAY COPIAS DE SEGURIDAD?
¿EXPORTAR MISPORTAB DATOS?
¿CUÁNTO TIEMPO SE CONSERVAN LOS BACKUPS?
¿CUMPLO CON LEYES Y OBLIGACIONES PROFESIONALES?
¿BAJO QUÉ LEY Y JURISDICCIÓN ESTOY?
¿TENGO UN ACUERDO EMPRESARIAL / DPA?
¿QUIÉN EN MI ORGANIZACIÓN PUEDE USAR IA?
USAR IA PUEDE SER CONFIDENCIAL (SÍ)
Con medidas adecuadas, controles y el proveedor correcto.
DEPENDE DE QUÉ, CÓMO, DÓNDE Y CON QUIÉN (EN QUÉ MEDIDA)
Riesgo gestionable con precauciones y buenas prácticas.
ALTO RIESGO DE QUE TU INFORMACIÓN TERMINE EN UN LUGAR INCORRECTO (NO)
Demasiadas variables fuera de tu control.
REGLAS DE ORO
MENOS ES MÁS: Comparte sólo lo necesario.
SI NO LO SABES, ASUME QUE NO ES CONFIDENCIAL.
REVISA, VALIDA Y VERIFICA SIEMPRE ANTES DE USAR.
USA HERRAMIENTAS Y PRÁCTICAS DE SEGURIDAD.
CAPACITA A TU EQUIPO Y ESTABLECE POLÍTICAS CLARAS.
LA TECNOLOGÍA CAMBIA, TUS DECISIONES TAMBIÉN.
NO HAY UNA RESPUESTA ÚNICA. HAY DECISIONES INFORMADAS.



La anatomía del precedente Heppner

Privilegio legal, privacidad y el mito de la inteligencia artificial.

Desmontando el mito urbano jurídico de Heppner.

LO QUE CIRCULA	LO QUE REALMENTE OCURRIÓ
Lo que circula en redes (Mito): El caso fue contra ChatGPT.	**Lo que realmente ocurrió (Hecho):** Utilizó Claude (Anthropic).
Lo que circula en redes (Mito): El juez pidió los chats a la empresa de IA.	**Lo que realmente ocurrió (Hecho):** El FBI encontró los materiales en los dispositivos incautados físicos del acusado.
Lo que circula en redes (Mito): Todo chat con IA es discoverable automáticamente.	**Lo que realmente ocurrió (Hecho):** El caso aplicó doctrinas tradicionales a hechos muy específicos; no creó una regla general absoluta.
Lo que circula en redes (Mito): El trabajo hecho con IA pierde work product siempre.	**Lo que realmente ocurrió (Hecho):** Se perdió porque el acusado actuó sin la dirección de sus abogados (counsel).

El contexto: Un target federal, un chatbot público y 31 documentos.

EL CASO

Asunto: United States v. Bradley Heppner, No. 25 Cr. 503 (JSR)

Tribunal: Southern District of New York (Juez Jed S. Rakoff).

Fecha de resolución: 17 de febrero de 2026.

Situación: Heppner, target de una investigación federal (securities fraud, wire fraud), usó la versión pública/consumer de Claude por iniciativa propia para organizar su defensa.

LOS INPUTS CONFIDENCIALES

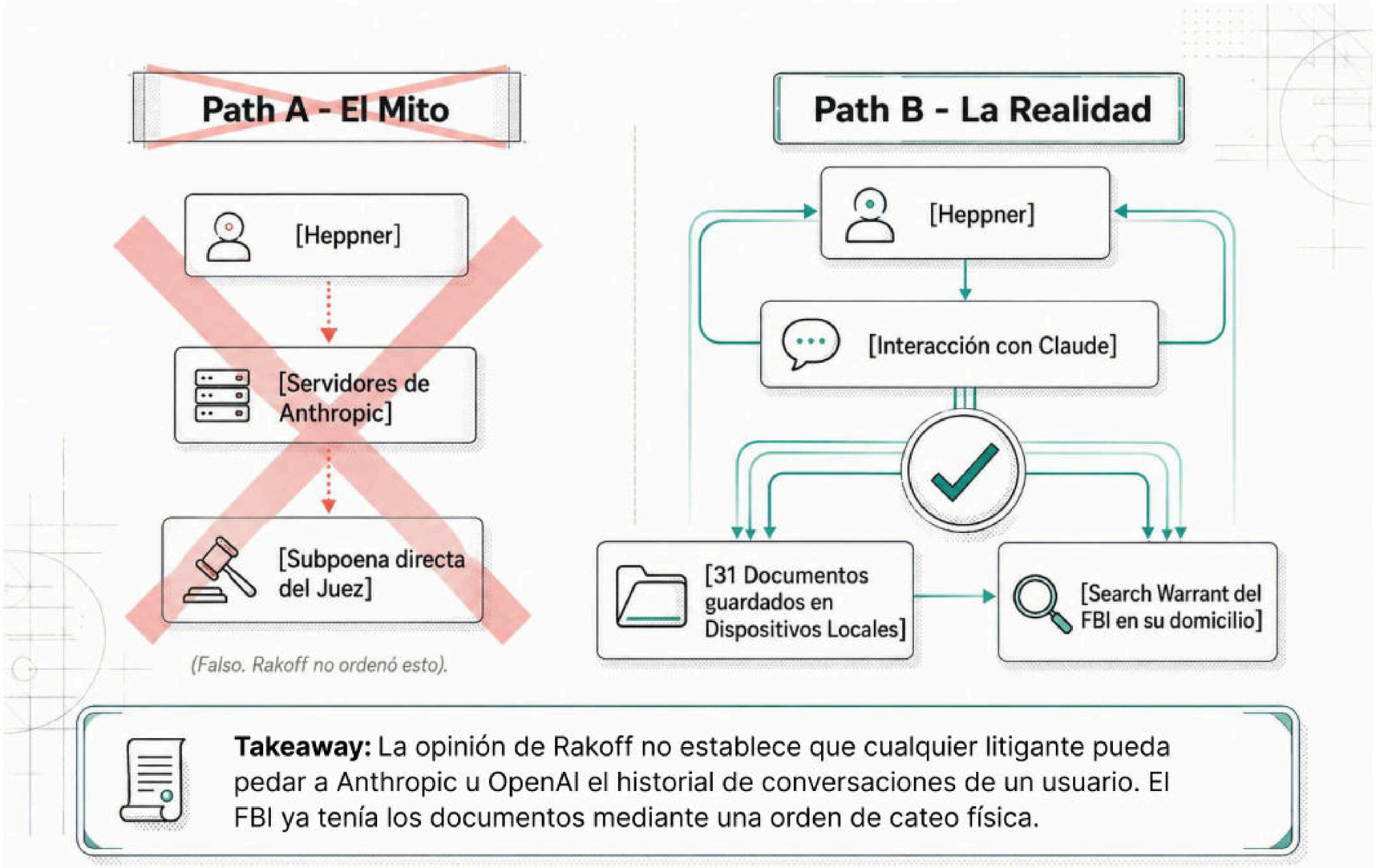
Heppner generó **31 documentos** alimentando a Claude con:

31

- Posibles cargos y estrategias de defensa.
- Hechos relacionados con su situación.

El error fatal: Información que había obtenido de conversaciones confidenciales previas con sus abogados.

La cadena de custodia: ¿Cómo llegaron los chats al gobierno?



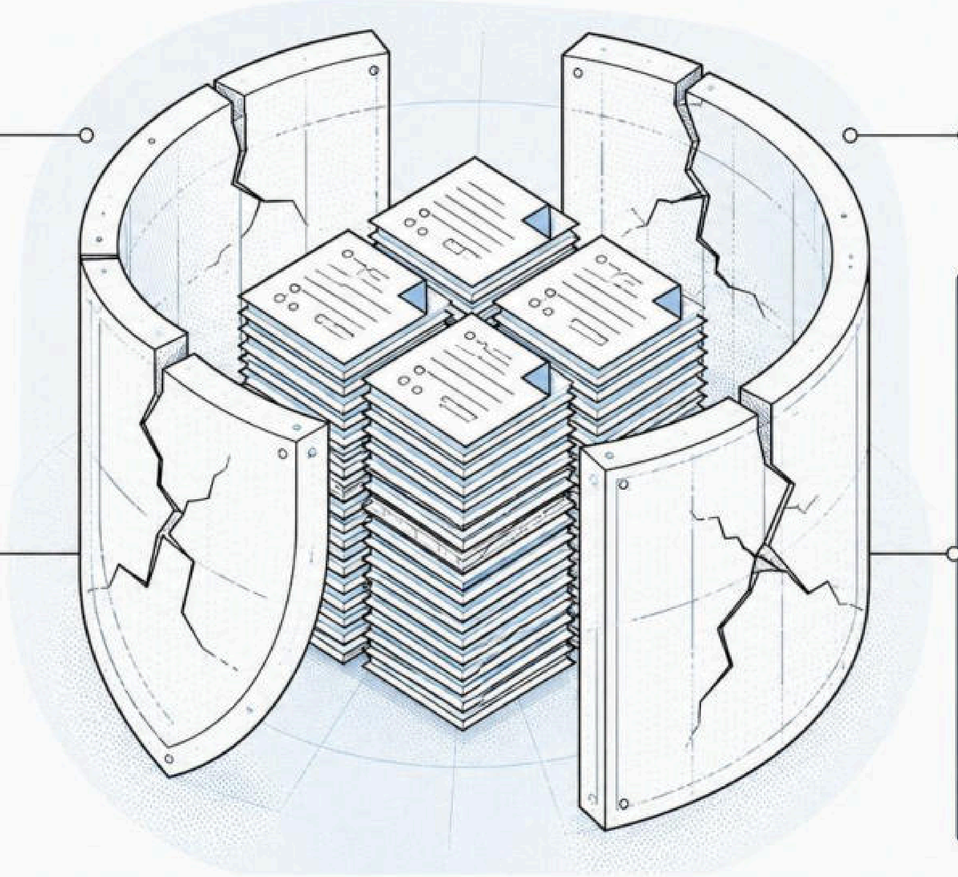
El choque de doctrinas: El intento de bloquear el discovery

Al encontrar los 31 documentos, la defensa los incluyó en su privilege log alegando dos escudos protectores distintos. El gobierno presentó una motion para destruirlos. El Juez Rakoff falló a favor del gobierno.

Escudo 1: Attorney-Client Privilege

Argumento de la defensa:
Usé a Claude para organizar mis pensamientos antes de compartirlos con mis abogados.

Resolución:
Rechazado



Escudo 2: Work-Product Doctrine

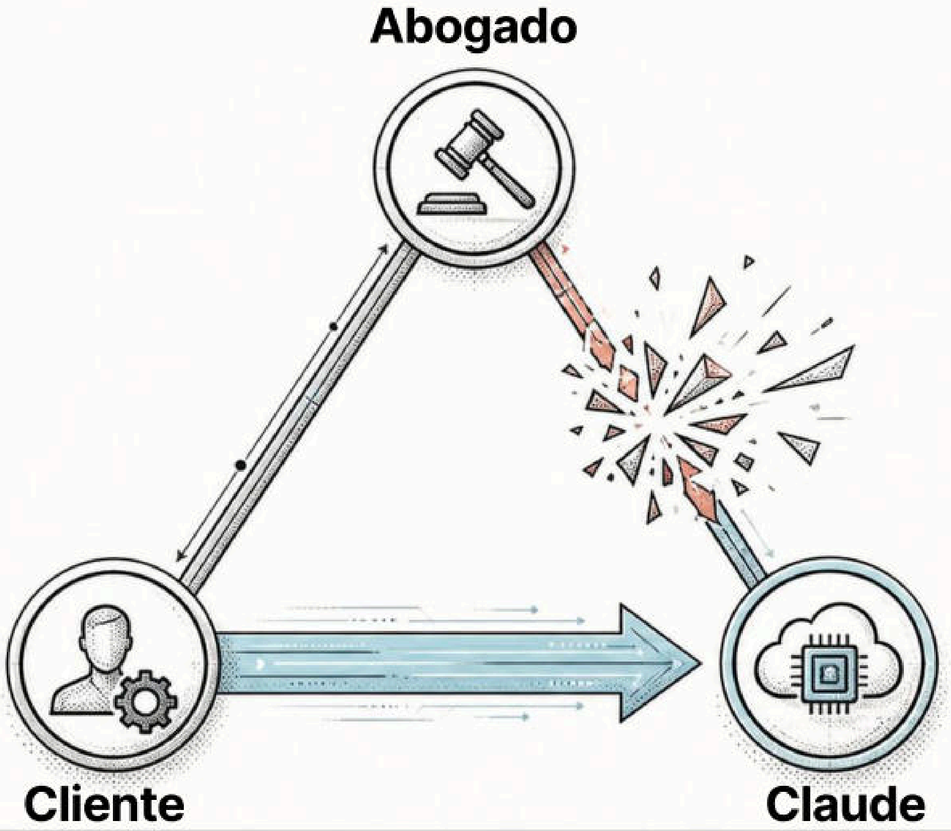
Argumento de la defensa:
Estos documentos fueron generados en anticipación al litigio.

Resolución:
Rechazado

El análisis del Privilege: Claude no es tu abogado.

El Estándar Tradicional:

comunicación confidencial con propósito de asesoría jurídica.

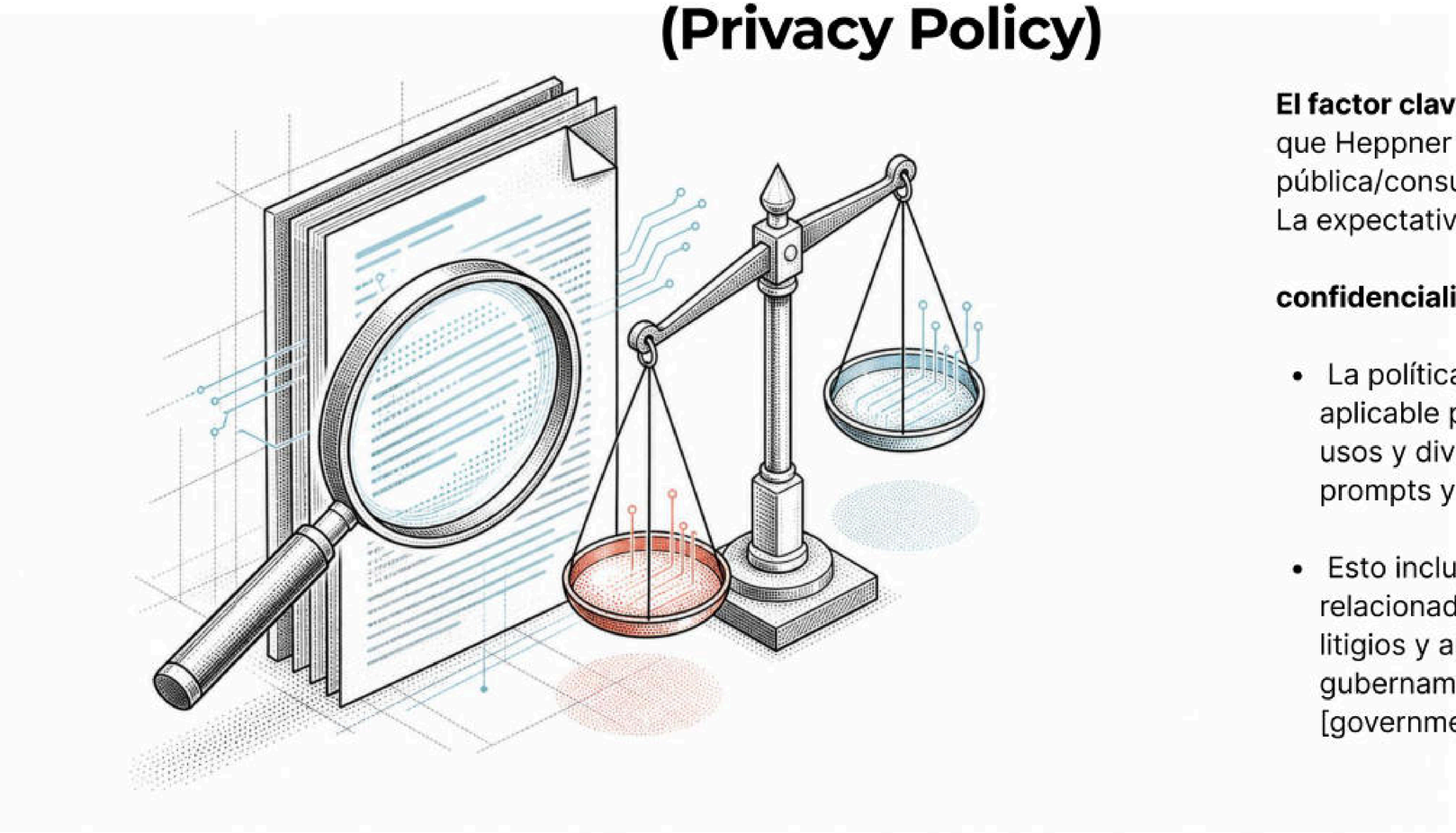


El Problema de Heppner:

- Una IA no es un licensed attorney.
- No tiene deberes fiduciarios profesionales.

Conclusión de Rakoff: Hablar con Claude sobre un problema jurídico, aunque preguntes exactamente lo mismo que le preguntarías a tu abogado, no crea attorney-client privilege.

El rol decisivo de los Términos de Servicio (Privacy Policy)



El factor clave: Rakoff observó que Heppner utilizó la versión pública/consumer de Claude. La expectativa de

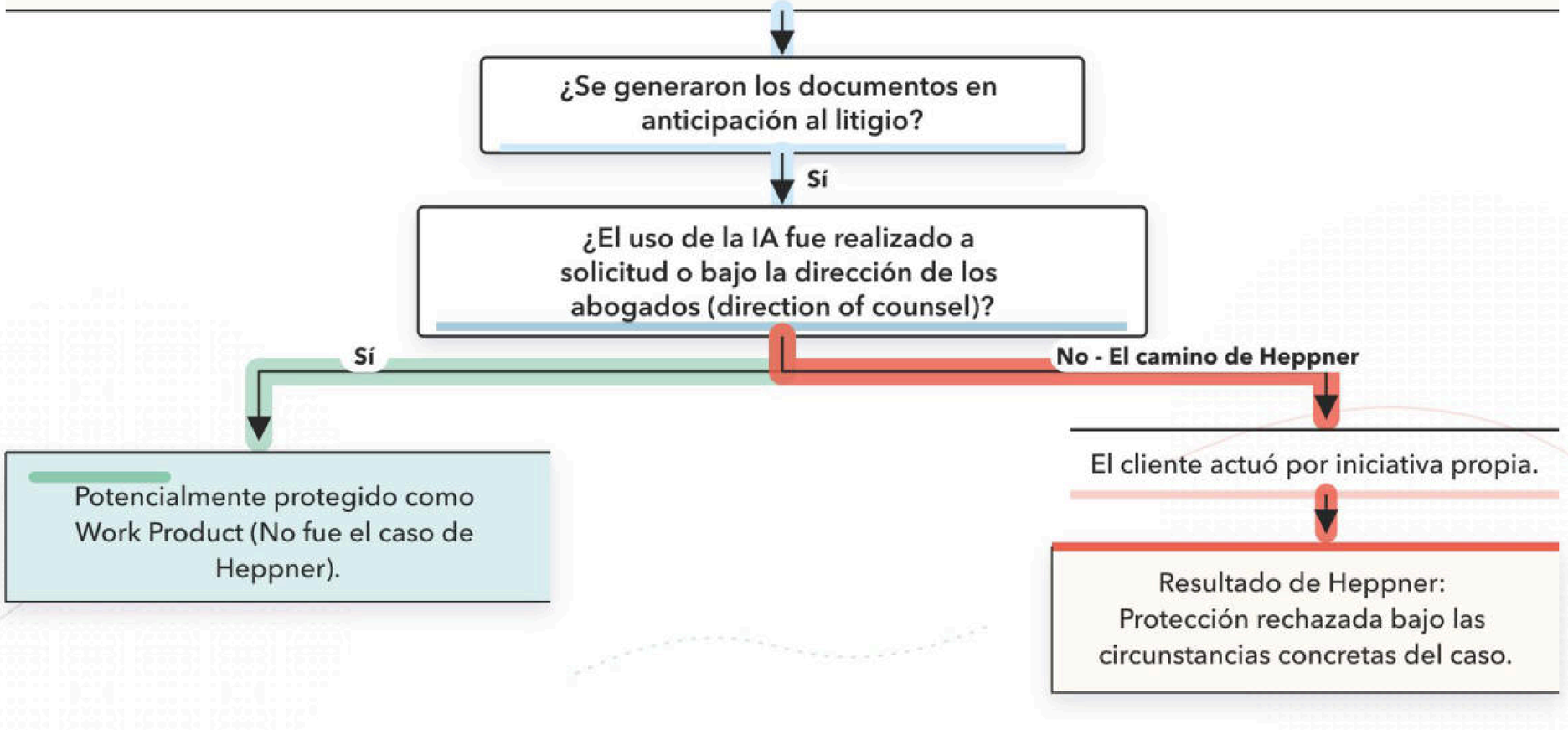
confidencialidad destruida:

- La política de privacidad aplicable permitía ciertos usos y divulgaciones de los prompts y outputs.
- Esto incluye circunstancias relacionadas con terceros, litigios y autoridades gubernamentales [government authorities].

El Precedente Real: Heppner NO establece que todo lo que dices a una IA es público. Establece que la información introducida voluntariamente en un servicio público sujeto a términos que permiten al proveedor procesarla o divulgarla no satisface el requisito jurídico de confidencialidad.

Diseccionando el Work Product: La regla de la "Dirección Legal"

El Mito: Un juez decidió que el trabajo realizado con IA nunca puede ser work product.
La Realidad: Rakoff rechazó la protección basándose en quién originó la acción.



La coincidencia increíble y la jurisprudencia dividida.

Contexto: El mismo 10 de febrero de 2026 en que Rakoff resolvió oralmente Heppner, un tribunal federal de Michigan decidió Warner v. Gilbarco, Inc.

HEPPNER (SDNY)

Herramienta
Claude (Anthropic)

Usuario
Target criminal
(Iniciativa propia)

Estatus jurídico de la IA
Equivalente a un tercero
(Third-party interlocutor)

Resultado
Protecciones denegadas
(Waived)

WARNER (ED MICHIGAN)

Herramienta
ChatGPT (OpenAI)

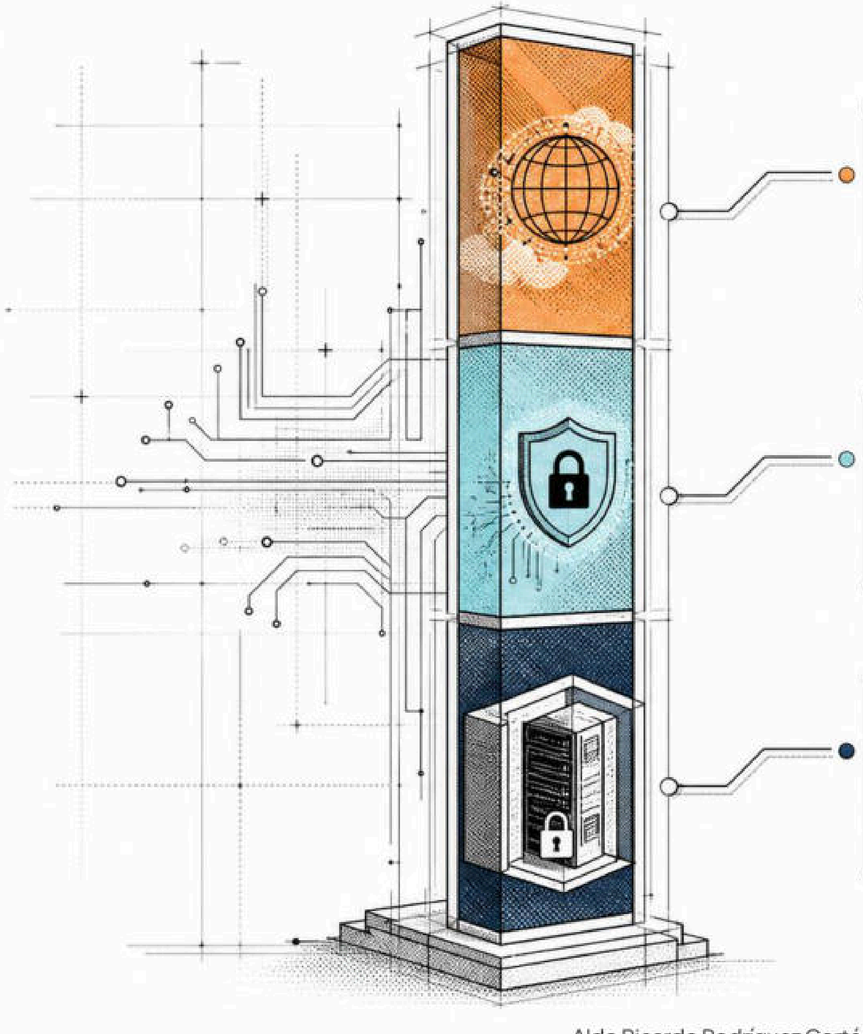
Usuario
Litigante Pro Se
(Preparando su litigio)

Estatus jurídico de la IA
Los programas generativos
son herramientas, no personas

Resultado
Interacciones protegidas
como Work Product

Conclusión de la ABA: La cuestión está lejísimos de resolverse a nivel appellate.

El verdadero legado: La arquitectura de privacidad es la arquitectura jurídica.



Nivel 1: Consumer AI (ChatGPT/Claude free/público)
Estructura: Datos sujetos a ToS públicos.
Impacto Legal: Alto riesgo de waiver (No hay expectativa de privacidad razonable).

Nivel 2: Enterprise AI / API Cerrada
Estructura: Confidencialidad contractual (contractual confidentiality), sin entrenamiento de modelos con los datos del cliente.
Impacto Legal: Riesgo Medio/Bajo (Potencialmente defendible como herramienta/agente).

Nivel 3: Self-Hosted / On-Premise LLM
Estructura: Operado estrictamente dentro del dominio del despacho, bajo dirección de counsel.
Impacto Legal: Altamente Protegido (Equivalente a infraestructura TI interna).

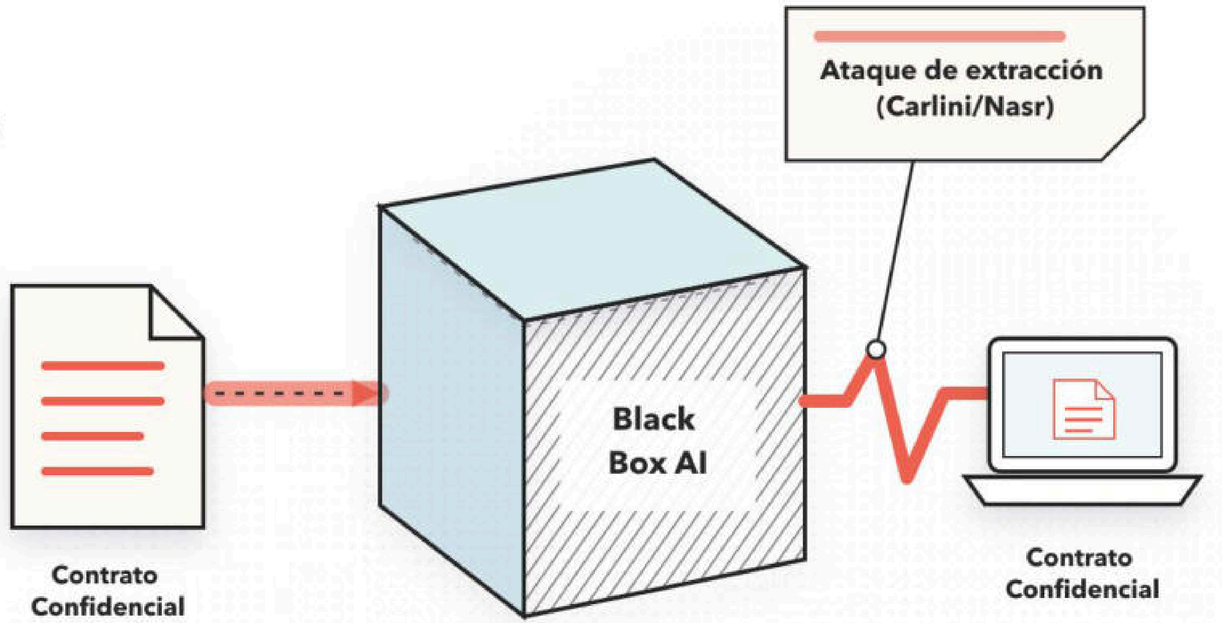
Si meto información a ChatGPT **esta será usada para entrenamiento**

¿Mi información puede terminar en las manos equivocadas?

El temor corporativo principal es la regurgitación de información confidencial.

Existe un miedo fundamentado en la industria: si introducimos un contrato confidencial en ChatGPT, el modelo podría 'memorizar' esos datos y regurgitarlos textualmente en la respuesta de otro usuario externo.

Este riesgo, demostrado en investigaciones como la de Carlini/Nasr, paraliza la adopción corporativa.

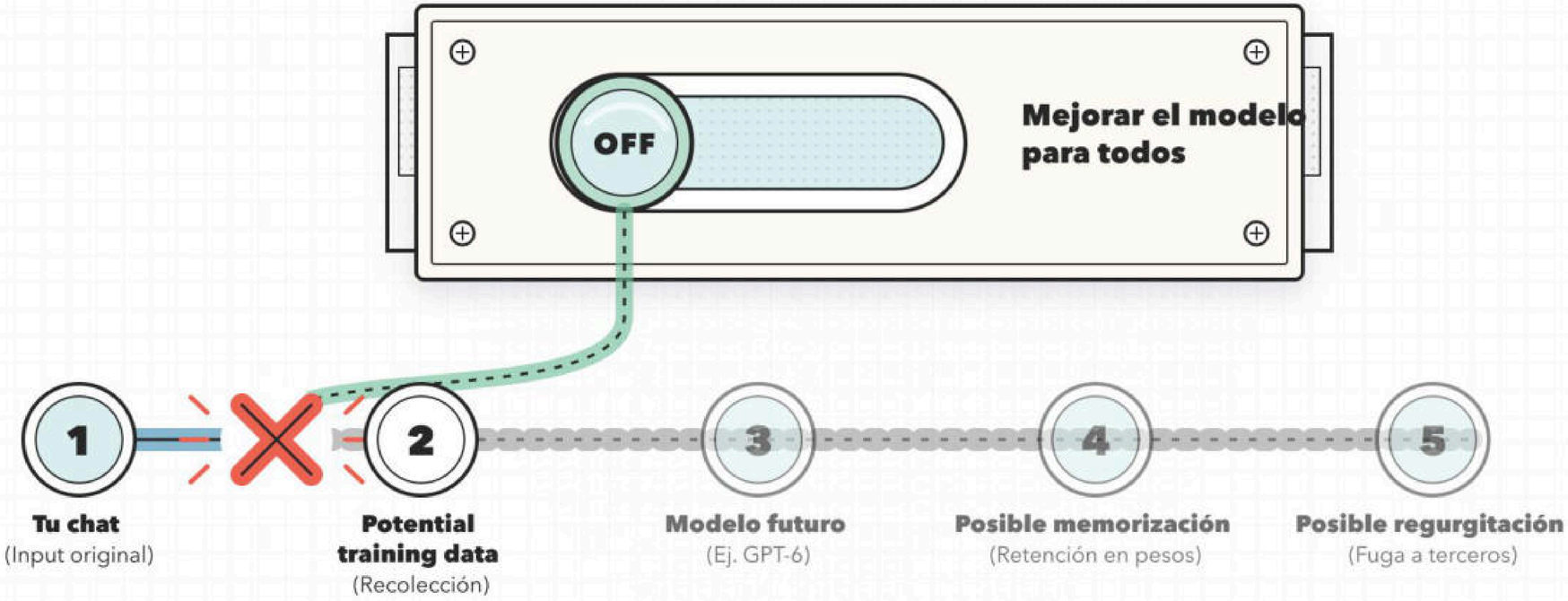


Una fuga de datos requiere que la información atraviese una secuencia exacta de cinco pasos.

Para que tus datos lleguen a un tercero, el flujo de información no puede interrumpirse.
Cada interacción personal con ChatGPT Free o Plus inicia este recorrido por defecto.

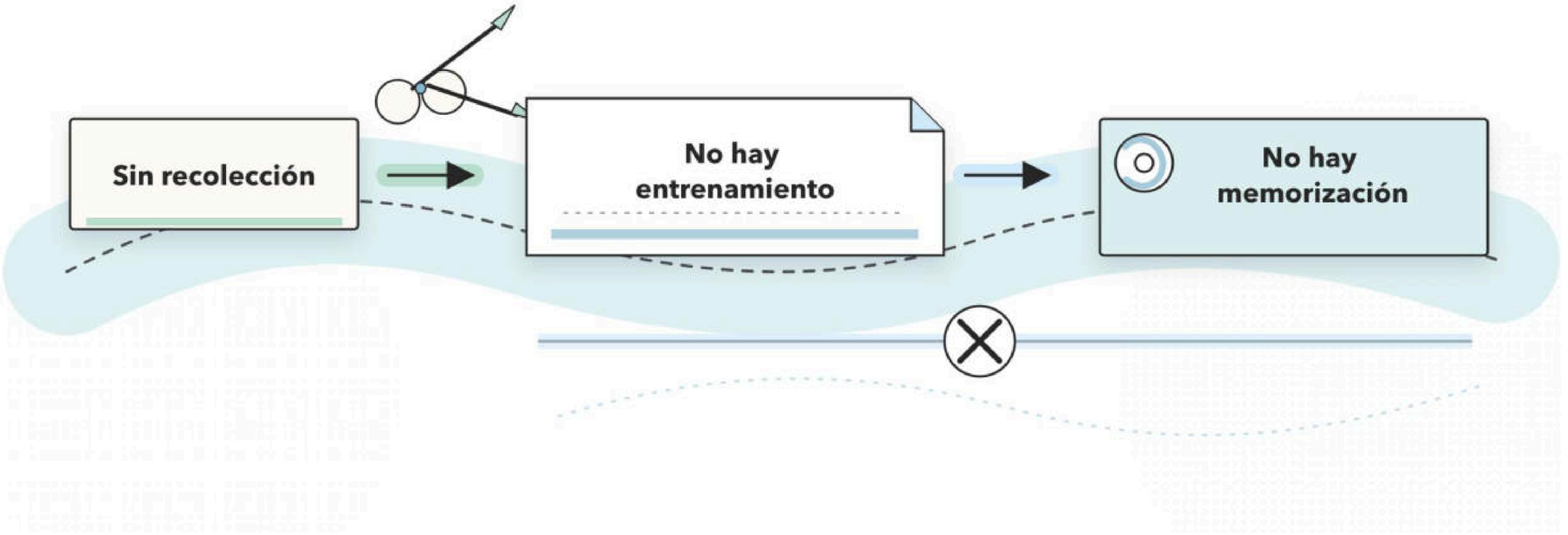
Desactivar la mejora del modelo actúa como un cortacircuitos definitivo.

Para las cuentas personales, hacer opt-out en la configuración de privacidad corta la cadena de raíz. OpenAI establece de forma inequívoca que, al apagar este interruptor, las conversaciones nuevas quedan excluidas del conjunto de datos de entrenamiento.



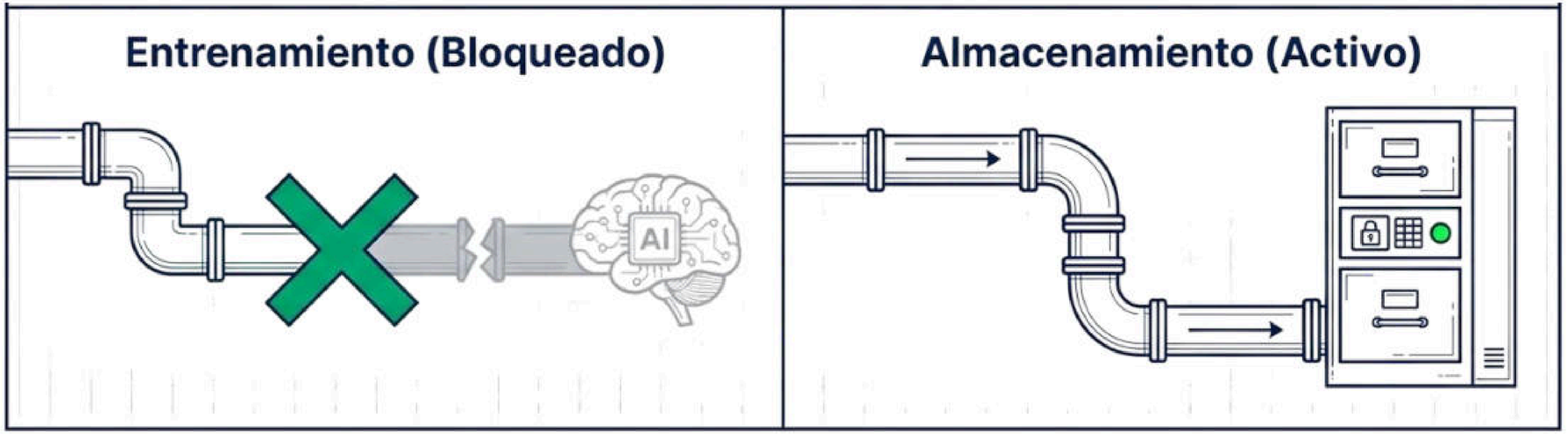
Cortar el flujo de entrada elimina matemáticamente el riesgo de regurgitación.

Ese mecanismo particular de fuga desaparece por completo respecto de las conversaciones nuevas. Es imposible que un modelo memorice y regurgite información que nunca fue inyectada en su base de entrenamiento.



Evitar el entrenamiento no impide que el chat se almacene en el historial

Ésta es la distinción más crítica: No training # no retention. Con el interruptor apagado, tus conversaciones continúan apareciendo en tu historial. El riesgo de entrenamiento se elimina, pero el flujo de almacenamiento se mantiene intacto.

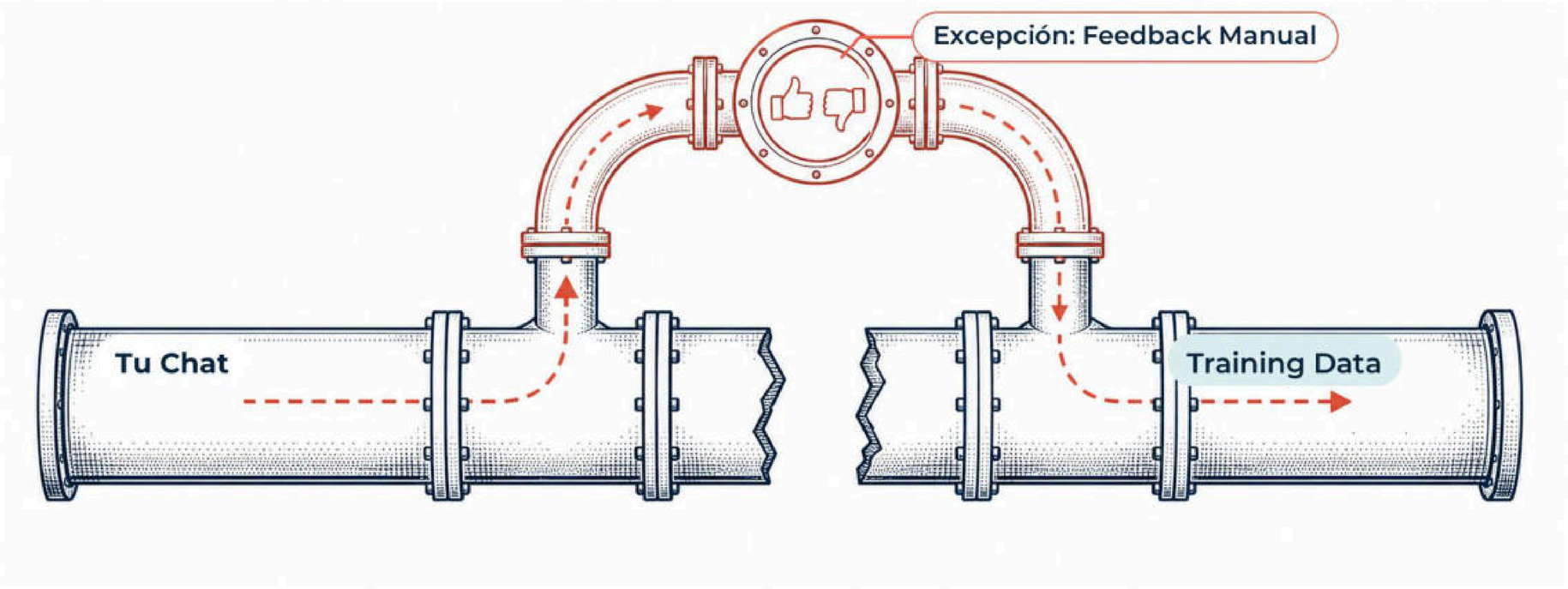


‘Temporary Chat’ proporciona un entorno efímero para consultas altamente sensibles.

Si el objetivo es minimizar también el almacenamiento, esta función extrema garantiza que las conversaciones no aparezcan en el historial, no creen memoria a largo plazo y no se utilicen para entrenamiento. Únicamente se conservan conservan en servidores seguros hasta 30 días estrictamente por motivos de seguridad y auditoría.

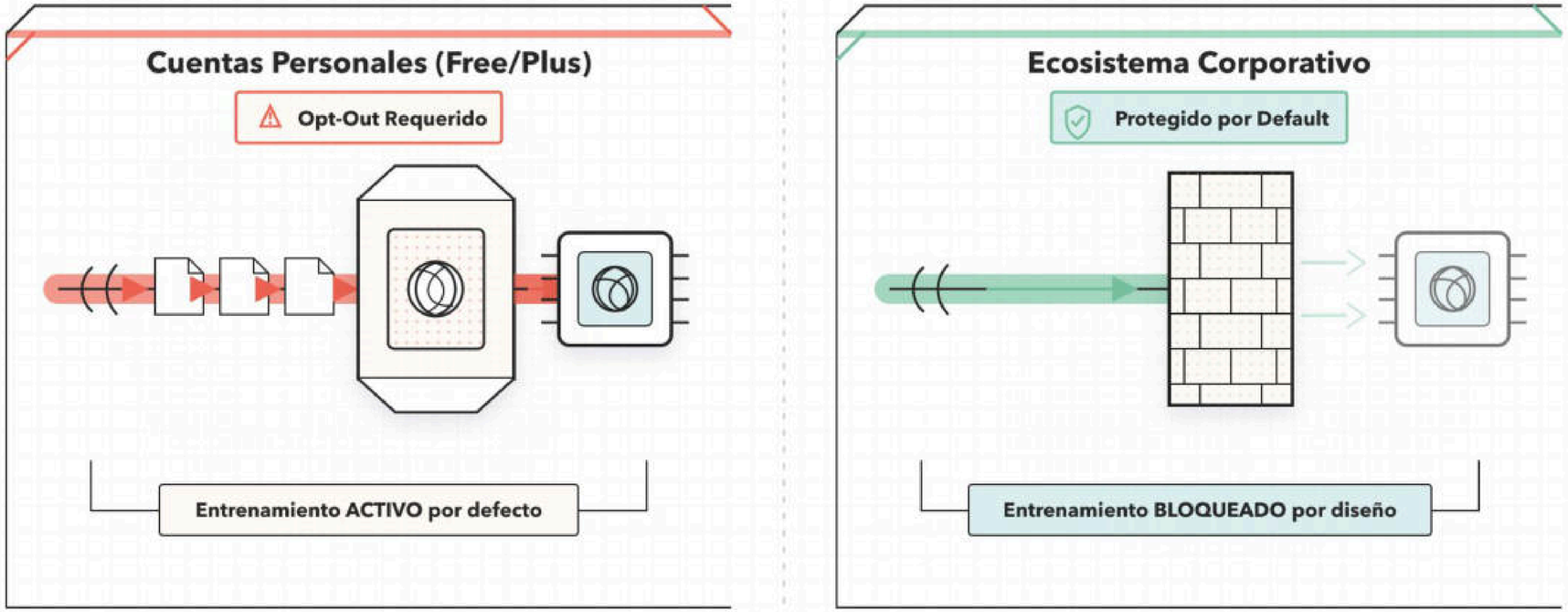
Calificar una respuesta reactiva el flujo de entrenamiento incluso con el modo privado activo.

Existe una excepción fácil de pasar por alto: el feedback manual. Si envías voluntariamente retroalimentación sobre una interacción, esa conversación específica puede ser extraída para entrenamiento. **Regla de oro: Nunca des feedback en prompts que contengan información corporativa confidencial.**



El ecosistema corporativo invierte la regla: la privacidad total es el estándar.

OpenAI opera bajo dos paradigmas opuestos. En las cuentas personales, el usuario debe defenderse (opt-out). En ChatGPT Business, Enterprise, Edu y la API, OpenAI no entrena por defecto con los inputs ni outputs. La organización tendría que solicitar explícitamente compartir esa data (opt-in).



Matriz de diagnóstico: Riesgo de memorización por nivel de servicio.

Un mapa exacto de cómo varían las políticas de datos según la licencia contratada.

Nivel de Servicio	¿Entrena por defecto?	Riesgo de memorización
ChatGPT Free / Plus / Pro	Sí	Existe
Personal + 'Improve model' OFF	No	Cortado
Temporary Chat	No	Cortado
ChatGPT Business / Enterprise	No	Cortado por default
ChatGPT Edu / OpenAI API	No	Cortado por default

El veredicto técnico y legal para las implementaciones corporativas.

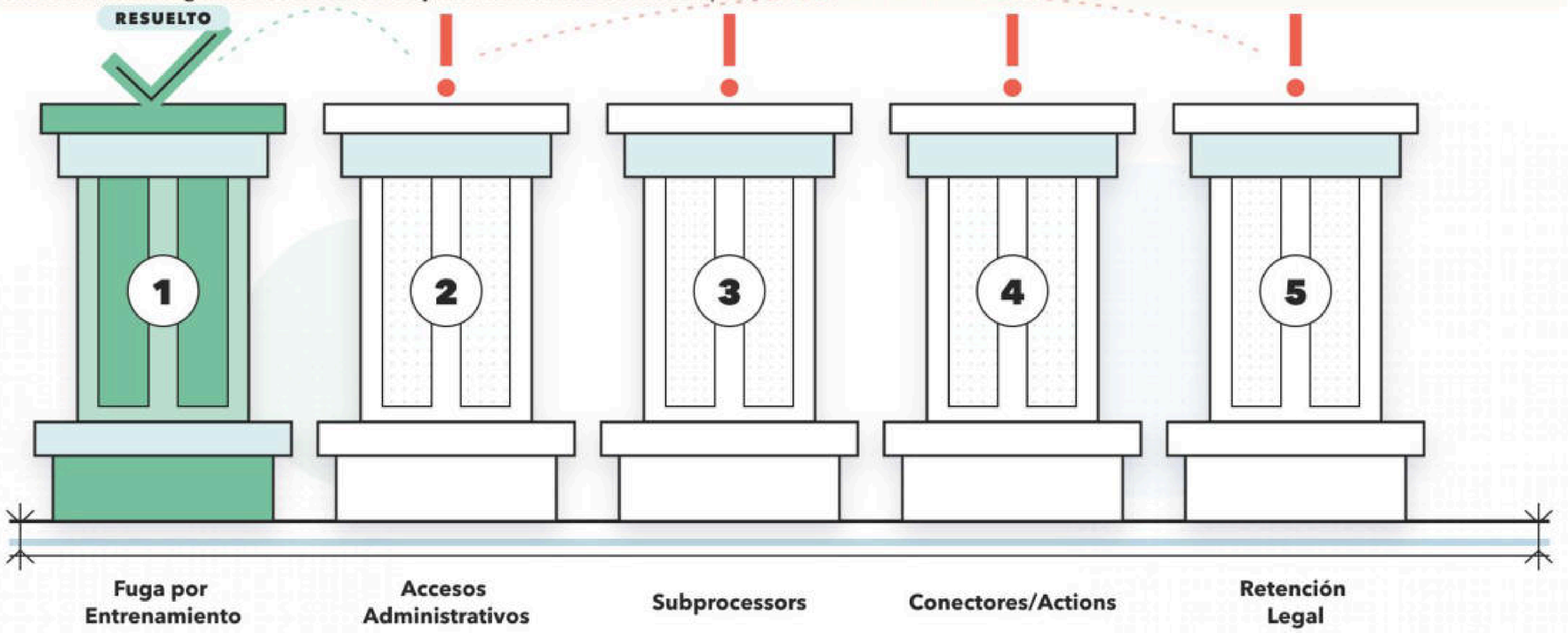
Bajo la política actual, los escenarios técnicos son diferentes. Los datos de Enterprise y API están aislados del motor de entrenamiento. El famoso riesgo de extracción de training data no es trasladable a los entornos corporativos oficiales de OpenAI.

> *Si introduzco un contrato confidencial en ChatGPT Enterprise, ¿puede acabar memorizado y aparecer en la respuesta de un tercero en GPT-6?*

NO.

Eliminar el riesgo de entrenamiento no equivale a un riesgo de privacidad cero.

Solucionar el riesgo de memorización es solo el primer paso. Las organizaciones maduras deben seguir gestionando vectores de riesgo tradicionales: el control de acceso, los logs de retención, las integraciones de terceros y las normativas de los subprocessors.

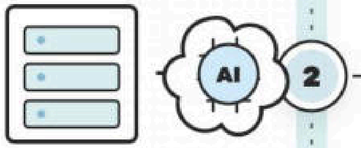


Resumen ejecutivo: Los tres pilares de una estrategia segura

1

1. El Opt-Out funciona.

Apagar el entrenamiento corta de raíz la cadena de memorización en cuentas personales.



2

2. No Training ≠ No Retention.

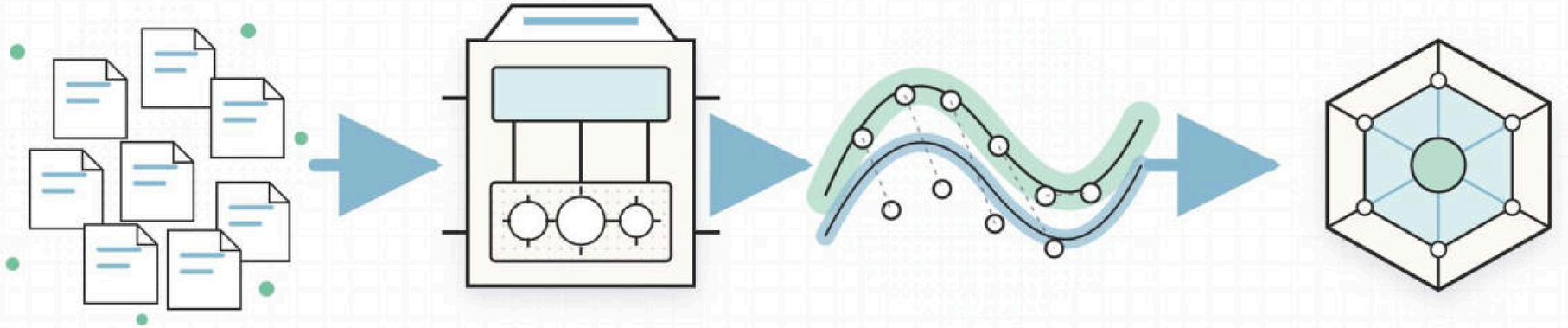
Evitar el entrenamiento no borra tu historial de los servidores; usa Temporary Chat si requieres impermanencia.

3

3. Enterprise es seguro por diseño.

Las versiones Business, Enterprise y API bloquean el entrenamiento de datos por defecto sin requerir acción del usuario.

La realidad algorítmica: Entrenar no es almacenar



Datos masivos (Millones de ejemplos)

Proceso de entrenamiento

Ajuste matemático de parámetros (Weights)

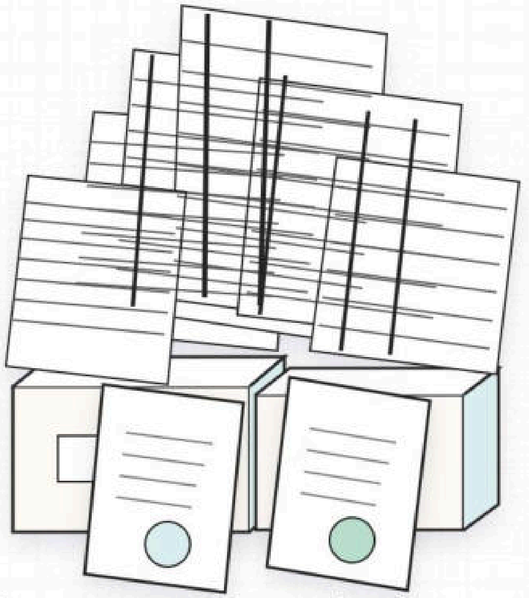
Modelo generativo.

Durante el entrenamiento, al modelo se le presentan cantidades enormes de texto para modificar estadísticamente sus parámetros (weights). El objetivo es predecir patrones, no conservar copias exactas.

Anthropic y **OpenAI** explican su proceso como el aprendizaje de relaciones y relaciones y patrones, no como la recuperación de documentos originales.

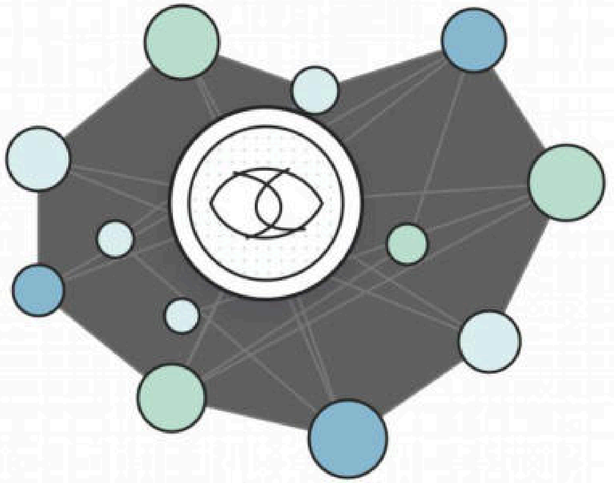
La analogía de los 10,000 contratos

Lo que no hace:



Conservar una copia mental perfectamente indexada de cada página para repetirla.

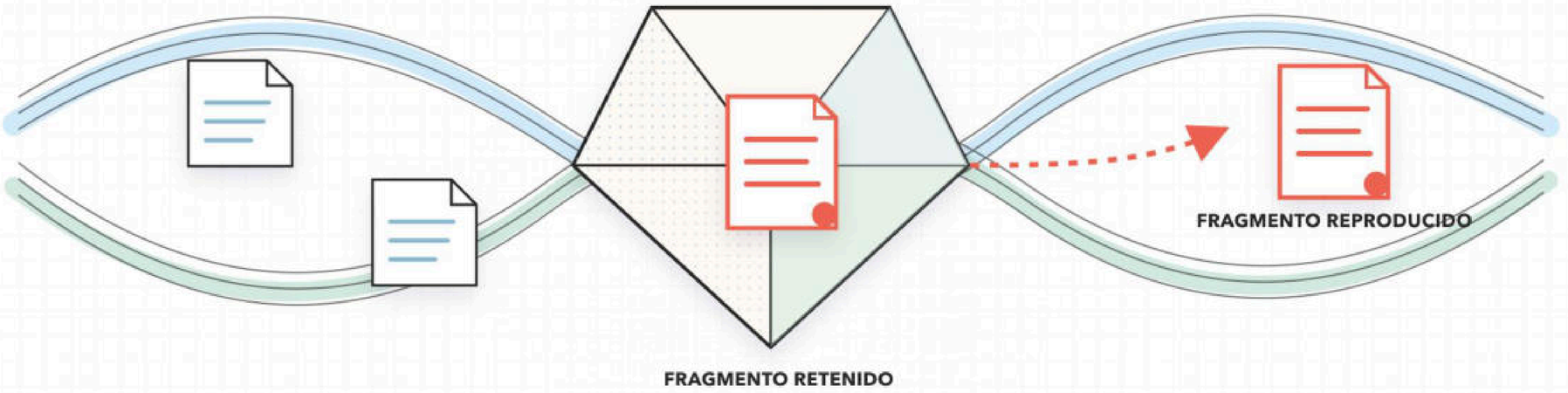
Lo que sí aprende (Patrones):



- Cómo se redacta un acuerdo de adquisición.
- Cómo estructurar fechas de cierre (closing dates).
- Lógica de precios y redacción jurídica.

El resultado final es **síntesis, no fotocopia.**

El mito corporativo: "La IA nunca recuerda"



"Como los modelos aprenden patrones, es técnicamente imposible que reproduzcan información de entrenamiento."

Esta afirmación corporativa simplifica demasiado el riesgo. La memorización de training data es un fenómeno real, documentado científicamente y reconocido por autoridades en ciberseguridad.

El nuevo estándar: Gobernanza de la Información

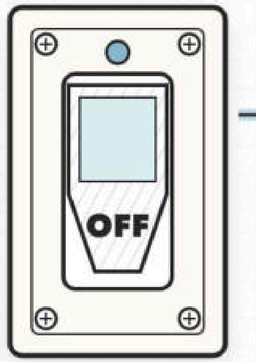
La posición seria no es "la IA no recuerda nada", ni tampoco "la IA publica todos tus secretos".

La pregunta ya no es "¿permitimos la IA?", sino "¿qué datos permitimos, en qué sistema y bajo qué configuración?".

Escenario	Usuario	Herramienta	Protección	Caso / Fundamento
Abogado desarrolla estrategia litigiosa usando IA	Profesional con licencia	Enterprise o comercial	**Sí — work product**	*Tremblay v. OpenAI (N.D. Cal., ago. 2024); Concord v. Anthropic (N.D. Cal., may. 2025)*
Profesional usa IA empresarial con garantías contractuales	Profesional con licencia	Enterprise	**Sí — profesional + contractual**	*Análisis de Debevoise & Plimpton sobre Heppner (feb. 2026): herramientas enterprise podrían tener resultado distinto*
Profesional usa IA comercial (consumo) para trabajo con clientes	Profesional con licencia	Comercial	**Parcial — work product probable, riesgo por ToS**	*Tremblay y Concord protegen prompts de abogados; ToS de consumo podrían debilitar la expectativa*
Cliente usa IA empresarial bajo instrucción de su abogado	Cliente / lego	Enterprise	**Sí — extensión abogado-cliente**	*Doctrina Kovel (United States v. Kovel, 2d Cir., 1961): terceros como extensión del abogado mantienen privilegio*
Cliente usa IA comercial bajo instrucción de su abogado	Cliente / lego	Comercial	**Debatible — dirección ayuda, ToS debilita**	*Sin precedente directo. Heppner distingue que la defensa «did not direct» al acusado; a contrario, la dirección podría cambiar el análisis*
Cliente usa IA comercial por cuenta propia para preparar consulta legal	Cliente / lego	Comercial	**No según Heppner**	*United States v. Heppner (S.D.N.Y., feb. 2026): fallo oral, sin opinión escrita. Debería ser privado por intencionalidad del usuario*
Usuario usa versión gratuita de IA para cualquier propósito	Cualquiera	Gratuita	**No — mínima expectativa**	*Heppner (por extensión). ToS de versiones gratuitas autorizan explícitamente uso de datos para entrenamiento sin reserva*

El Interruptor de Privacidad: Entrenamiento ≠ Monitoreo

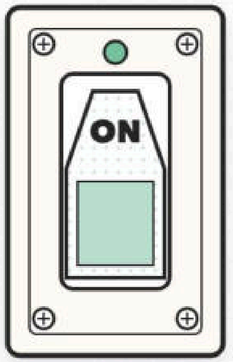
Apagar "Improve the model for everyone" detiene el uso de datos para IA, pero no desactiva la seguridad.



~~Chat → Training~~

(Mejora de IA)

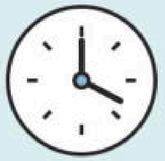
Bloqueado



~~Chat → Safety Monitoring~~

(Sistemas de seguridad)

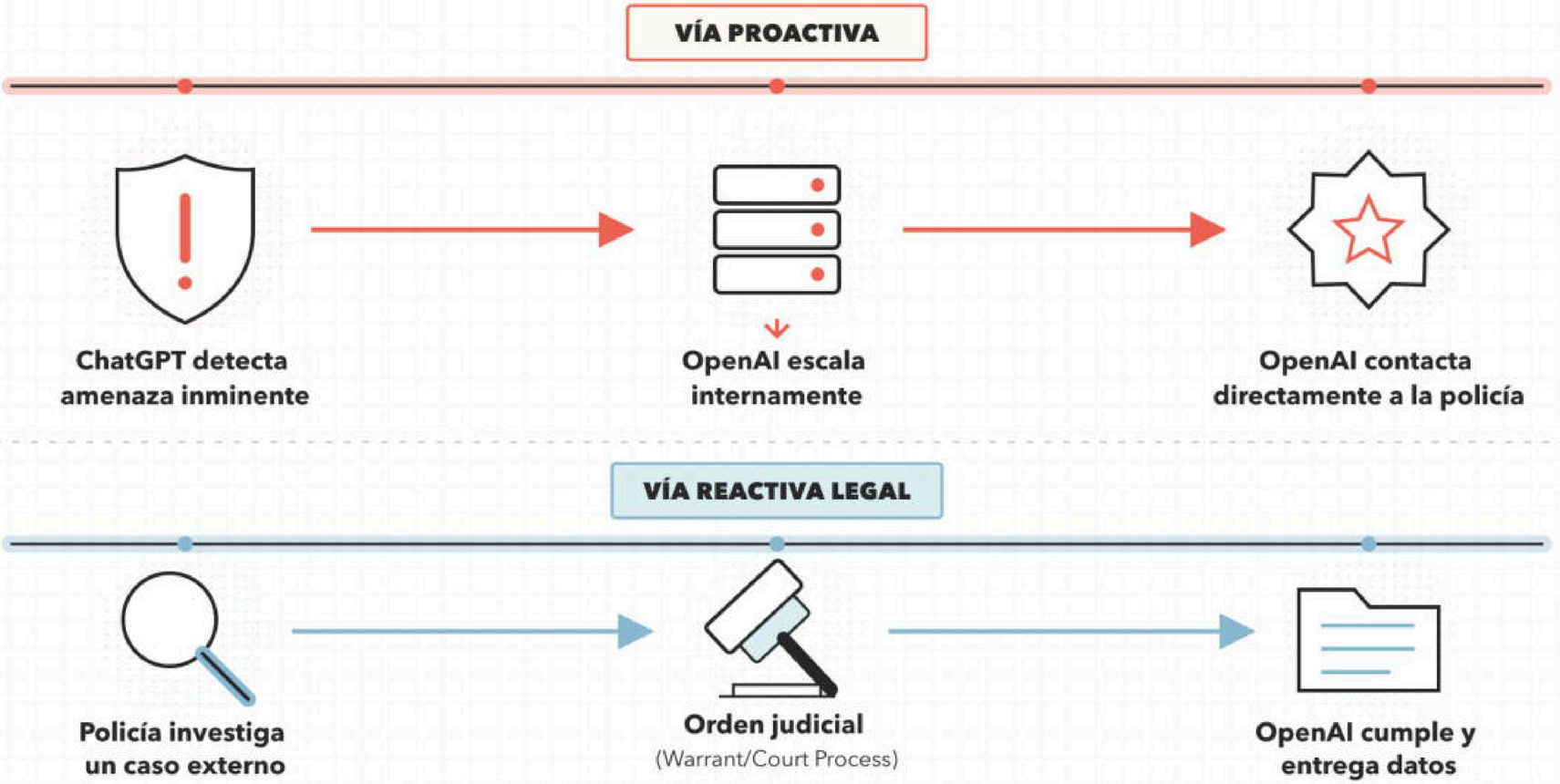
Activo



El caso de "Temporary Chat"

Aunque no se usa para training y se elimina de tu historial, OpenAI conserva**una copia** hasta por **30 días** exclusivamente por razones de seguridad.

El Puente Dual hacia las Autoridades



Ambos pueden ocurrir, pero jurídicamente son caminos completamente diferentes.

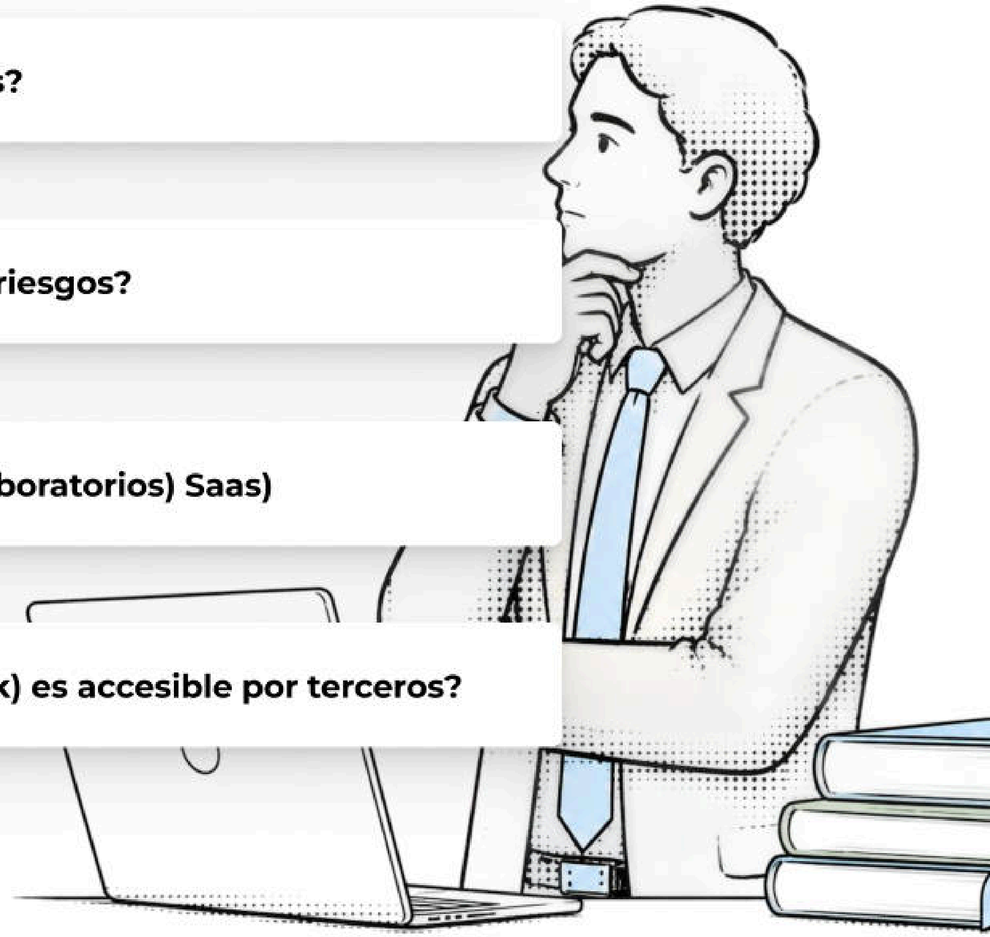
Preguntas típicas

¿Hay modelos más seguros que otros?

¿Correr modelo en local me quita todos los riesgos?

¿Qué pasa con proveedores de servicios de IA (no laboratorios) Saas)

¿La memoria de mi cuenta de (ChatGPT / Claude / Gemini / Grok) es accesible por terceros?




Entorno
Cloud / SaaS
Local / Self-Hosted
Tipo de Modelo
Cerrado
Open Weights
Riesgo Contractual
Ej: ChatGPT, Claude
N/A
(Generalmente imposible)
Falsa Sensación de Seguridad
Ej: Llama en Groq
Tu prompt sigue viajando a un tercero.
Privacidad Arquitectónica
Ej: Qwen/Llama en hardware propio
Open + Cloud te expone exactamente igual que un modelo cerrado.
El valor real está en el cuadrante Open + Local.

La Cascada de Disponibilidad

«La IA se queda con tu información confidencial» se convirtió en un dogma gremial no porque sea cierto, sino porque está en todas partes. Nadie fue a verificar la fuente.