Oscar Obando
CHAVES

Lo que pegaste en el chat

Qué pasa de verdad con la información que escribes en ChatGPT, Claude o Gemini, según el plan que pagas. El riesgo que casi nunca ocurre, el que ocurre todos los días, y el que ya tiene sentencias.

Oscar Obando Chaves

oscarobando.ai

Julio de 2026

EL PUNTO DE PARTIDA

Lo que ocurre con tus conversaciones depende del plan

La pregunta que se hace cualquier profesional después de pegar un contrato, un balance o el caso de un cliente en una conversación de inteligencia artificial merece una respuesta técnica y no un sermón, y esa respuesta empieza por un dato que mucha gente desconoce, porque lo que la empresa proveedora puede hacer con ese texto cambia de forma sustancial según el plan contratado.

En **OpenAI** los planes de consumo, o sea Free, Plus y Pro, se usan para entrenar los modelos de forma predeterminada, y la propia empresa lo declara en su centro de ayuda al explicar que ChatGPT mejora entrenando con las conversaciones de la gente salvo que el usuario lo desactive, con el interruptor ubicado en la sección de controles de datos de la configuración, mientras que los planes Team, Enterprise, Edu y la interfaz de programación quedan excluidos sin necesidad de trámite alguno.

En **Anthropic** la política cambió de sentido con los términos que entraron en vigor el **28 de septiembre de 2025**, porque antes las conversaciones de consumo se borraban a los treinta días y no alimentaban entrenamiento, y desde entonces los usuarios de Free, Pro y Max eligen entre permitir ese uso, con retención extendida a cinco años, o rechazarlo y conservar la ventana de treinta días, con Claude for Work, la interfaz de programación y las versiones de gobierno y educación fuera de ese régimen.

En **Google** la aplicación de consumo mantiene revisión humana de una muestra de conversaciones y retención prolongada, con un detalle que conviene subrayar porque el usuario no puede revertirlo, ya que las conversaciones revisadas por personas se conservan hasta **tres años** y permanecen aunque borres tu actividad, mientras Workspace y Vertex AI operan bajo control del administrador y sin usar los contenidos para entrenar los modelos públicos.

Fuentes. OpenAI, centro de ayuda sobre uso de datos para mejorar el rendimiento de los modelos. Anthropic, actualización de términos de consumo (vigente 28 de septiembre de 2025). Google, centro de privacidad de las aplicaciones Gemini.

EL MIEDO QUE CASI NUNCA OCURRE

Que tu secreto salga en la respuesta de un desconocido

El temor más extendido supone que un dato confidencial escrito en una conversación quedará aprendido por el modelo y aparecerá algún día en la respuesta que reciba otra persona, y la evidencia científica disponible permite ponerle una probabilidad realista a ese escenario, que resulta cercana a cero para un dato que aparece una sola vez.

El trabajo de referencia lo firmó **Nicholas Carlini** junto a un equipo de investigadores y se presentó en la conferencia ICLR de 2023, y establece que la memorización de un modelo crece con tres factores, el tamaño del modelo, la longitud del contexto usado para provocarla y sobre todo **el número de veces que un dato aparece duplicado** en los datos de entrenamiento, con la conclusión de que estos sistemas rara vez reproducen cadenas que se repiten pocas veces.

El estudio de Kandpal y su equipo llegó por otra vía al mismo lugar, porque encontró que los métodos para detectar memorización operan con una precisión cercana al azar frente a secuencias no duplicadas, y calculó que una secuencia presente diez veces se genera alrededor de mil veces más que una que aparece una sola vez, mientras el trabajo conocido como **The Secret Sharer** insertó datos señuelo en entrenamientos controlados y midió que un valor único queda por debajo del umbral que permitiría extraerlo.

Conviene entonces separar con precisión dos riesgos que la conversación pública mezcla todo el tiempo, porque una cosa es que el proveedor entrene con tus conversaciones, lo cual es una decisión de política que puedes desactivar o evitar contratando un plan empresarial, y otra muy distinta es que otro usuario reciba tu dato en una respuesta, escenario que sigue siendo técnicamente improbable incluso cuando el entrenamiento ocurre, y que además la deduplicación estándar de la industria reduce todavía más.

Fuentes. Carlini et al., "Quantifying Memorization Across Neural Language Models", ICLR 2023. Kandpal et al., "Deduplicating Training Data Mitigates Privacy Risks in Language Models" (2022). Carlini et al., "The Secret Sharer", USENIX Security 2019.

EL RIESGO QUE SÍ OCURRE

Tu contraseña vale más que la memoria del modelo

El vector de exposición que de verdad produce daño todos los días tiene poco de sofisticado y mucho de doméstico, porque consiste en que alguien entre a tu cuenta con tus credenciales robadas o en que tú mismo compartas por error lo que creías privado.

La medición más citada la publicó la firma **Group-IB** en junio de 2023, cuando identificó **101.134** dispositivos infectados con programas ladrones de credenciales que tenían guardadas las claves de acceso a ChatGPT, y esos accesos circulaban en mercados clandestinos, con Raccoon, Vidar y RedLine como las familias de programas maliciosos responsables, un episodio que la empresa aclaró que ocurrió en los equipos de los usuarios y no en sus propios sistemas.

El error humano aporta el resto de los casos conocidos, y el ejemplo clásico lo protagonizó **Samsung** en 2023, cuando en menos de veinte días sus ingenieros pegaron código fuente propietario, notas de reuniones y secuencias de prueba en tres incidentes distintos, lo que llevó a la compañía a prohibir estas plataformas en mayo de ese año, y el desenlace resulta instructivo porque en junio de 2026 la misma empresa desplegó ChatGPT Enterprise para sus empleados en Corea y para su división global de dispositivos, de modo que la solución llegó por la vía del plan adecuado.

El episodio del verano pasado completa el mapa de los descuidos, porque una función experimental permitía marcar una conversación compartida como localizable, y cerca de **4.500** conversaciones terminaron indexadas en el buscador con currículums, datos personales y código propietario a la vista, hasta que el responsable de seguridad de la información de OpenAI anunció el retiro de la función a inicios de agosto de 2025, en un caso donde ningún sistema fue vulnerado y donde todo el daño provino de una opción mal comprendida.

Fuentes. Group-IB, comunicado del 20 de junio de 2023. Forbes y Bloomberg sobre el caso Samsung (mayo de 2023) y anuncio de despliegue empresarial (junio de 2026). Engadget y declaraciones de Dane Stuckey, CISO de OpenAI (agosto de 2025).

EL RIESGO JURÍDICO

Los tribunales ya empezaron a responder

Para quien trabaja con información de clientes, el frente más serio ocurre en los tribunales, y durante los últimos meses aparecieron las primeras decisiones que responden preguntas que hasta hace poco se discutían en abstracto en congresos y seminarios.

Sobre el secreto comercial ya hay una respuesta judicial, porque en el caso **Trinidad contra OpenAI** resuelto este año en el distrito norte de California el juez Tigar desestimó una demanda por apropiación de secretos, y el razonamiento apuntó a que la demandante había desarrollado sus protocolos usando ChatGPT, con lo cual los divulgó de forma voluntaria a una empresa que no tenía obligación de mantenerlos confidenciales, de modo que el derecho sobre ellos quedó extinguido y el requisito de medidas razonables de protección quedó incumplido.

Sobre el privilegio profesional apareció la primera decisión de fondo en **Estados Unidos contra Heppner**, resuelta en febrero de este año en el distrito sur de Nueva York por el juez Jed Rakoff, quien sostuvo que los documentos generados por el acusado usando la versión pública de Claude carecían de protección por privilegio o por producto del trabajo, con la política de privacidad de la plataforma y la ausencia de expectativa razonable de confidencialidad entre sus fundamentos, aunque conviene registrar que otros tribunales han resuelto en sentido distinto y que la doctrina permanece en formación.

La confirmación más directa la dio el propio **Sam Altman** en julio de 2025, cuando explicó públicamente que hablar con ChatGPT carece del privilegio legal especial que ampara las conversaciones con un médico, un abogado o un terapeuta, y que la empresa podría verse obligada a entregar esas conversaciones ante un requerimiento judicial, una advertencia que la orden de preservación dictada en el litigio del New York Times volvió tangible al alcanzar incluso conversaciones borradas por los usuarios, con los planes Enterprise, Edu y los que operan sin retención de datos expresamente excluidos.

Fuentes. Trinidad v. OpenAI (N.D. Cal., 2026). United States v. Heppner (S.D.N.Y., febrero de 2026). Declaraciones de Sam Altman (julio de 2025). OpenAI, comunicado sobre las demandas de datos del New York Times. ABA, Opinión Formal 512 (29 de julio de 2024).

EL CUADRO COMPLETO

Qué protege cada plan

Reunidas en una tabla, las diferencias entre planes dejan de ser marketing y se vuelven decisiones concretas de gestión del riesgo.

Plan	Entrena con tus datos	Qué obtienes
Consumo sin ajustes (Free, Plus, Pro, Max)	**Sí, salvo que lo desactives**	Retención prolongada, revisión humana por abuso, conversaciones descubribles en juicio.
Consumo con entrenamiento desactivado	**No, hacia adelante**	Ventana de retención corta, sin efecto retroactivo sobre lo ya entrenado.
Team y Enterprise	**No, por contrato**	Retención configurable, controles de acceso, contrato de tratamiento de datos, exclusión del hold judicial.
Interfaz de programación con retención cero	**No**	Sin almacenamiento de solicitudes ni respuestas, sin revisión humana, disponible solo para clientes aprobados.

La tabla admite un resumen operativo simple, porque el salto de protección relevante ocurre al pasar del plan de consumo al plan empresarial, donde la prohibición de entrenar deja de ser una política revocable y pasa a ser una cláusula contractual, con la retención bajo control del cliente y con el contrato de tratamiento de datos que exige cualquier norma de protección de datos personales para procesar información de terceros.

Fuentes. Términos comerciales de OpenAI y de Anthropic. Documentación de retención cero de ambas empresas. OpenAI, comunicado sobre el alcance de la orden de preservación judicial.

EL VEREDICTO

Qué hacer si ya pegaste algo, y qué mirar de ahora en adelante

La conclusión práctica de todo lo anterior se puede formular con proporción, y consiste en que preocuparse por la información ya ingresada tiene sentido cuando compromete a un cliente o a un secreto, y carece de sentido cuando se trata del borrador de un correo, porque el daño esperado por regurgitación resulta mínimo y el daño jurídico depende por completo de qué escribiste.

Lo que todavía puedes hacer funciona hacia adelante y toma pocos minutos, porque puedes desactivar el uso para entrenamiento, borrar las conversaciones comprometidas, eliminar la memoria y las instrucciones personalizadas, revocar los enlaces que compartiste alguna vez y activar el segundo factor de autenticación, que es la medida que ataca el vector de riesgo más frecuente de todos.

Lo que ya no puedes revertir conviene aceptarlo con la misma claridad, porque una corrida de entrenamiento ya ejecutada no se deshace, la retención residual sigue su curso, los contenidos marcados por revisión de abuso permanecen archivados por períodos que llegan a varios años, y las conversaciones alcanzadas por una orden judicial de preservación quedan fuera del alcance de cualquier botón de borrado.

Queda una calibración que conviene hacer sin dramatismo, porque enviar esa misma información por correo electrónico, guardarla en un servicio de almacenamiento en la nube o dictársela a cualquier proveedor de programas empresariales presenta un perfil de riesgo comparable, con la misma descubribilidad judicial y la misma dependencia de la seguridad de tu cuenta, de manera que la pregunta útil deja de ser si usar esta tecnología y pasa a ser bajo qué plan, con qué contrato y con qué información. ¿Sabes hoy en qué plan está corriendo la conversación donde pegaste el último documento de un cliente?

Fuentes. Documentación oficial de OpenAI, Anthropic y Google sobre controles de datos y retención. Análisis propio sobre la calibración comparativa del riesgo.

Oscar Obando
CHAVES

“There's no special legal privilege for talking to ChatGPT like there is for talking to a doctor, lawyer, or therapist.”

— Sam Altman, director ejecutivo de OpenAI, julio de 2025

Oscar Obando Chaves
oscarobando.ai · linkedin.com/in/oscar-obando-chaves-237495133

Julio de 2026 · Quito, Ecuador