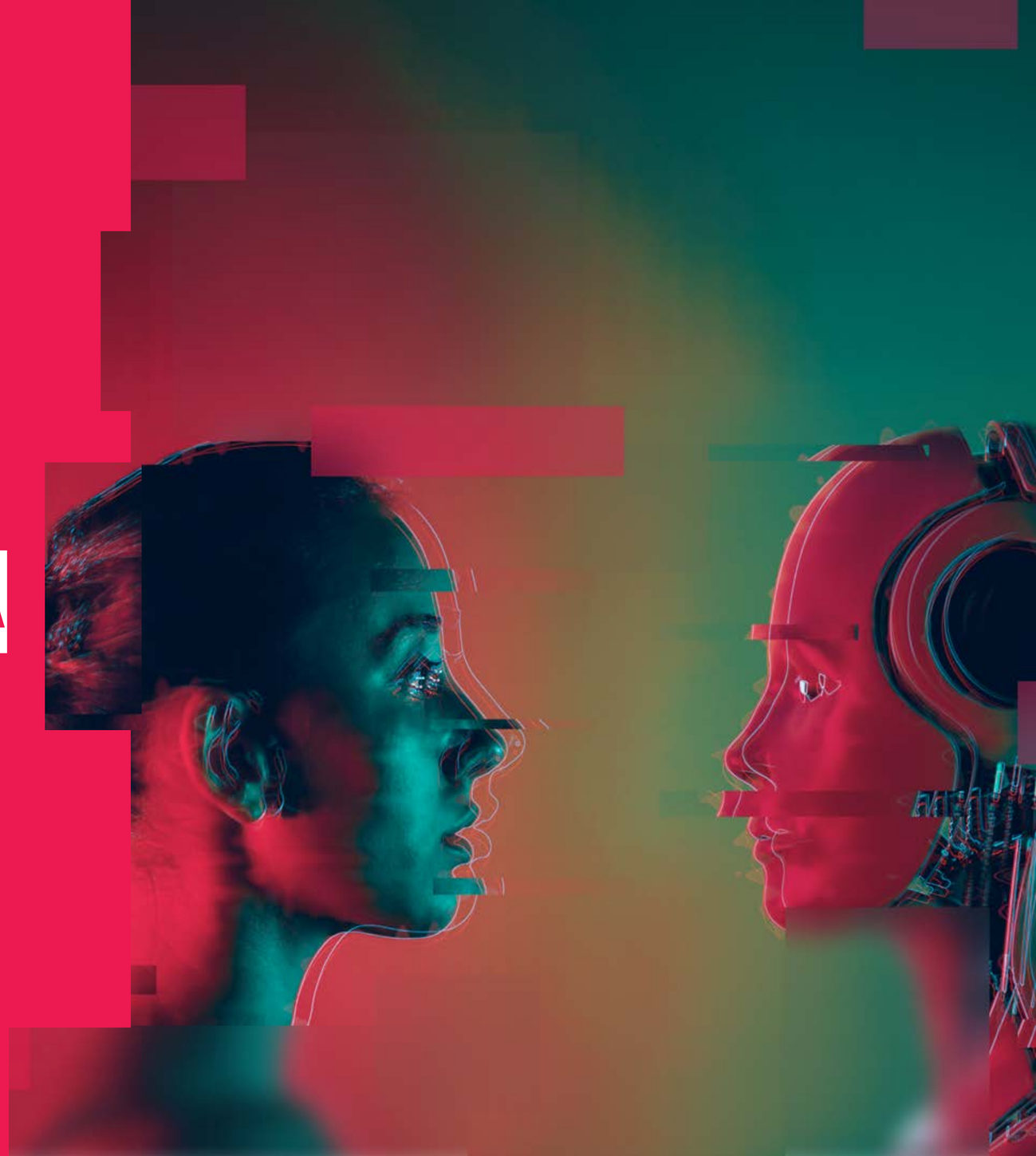
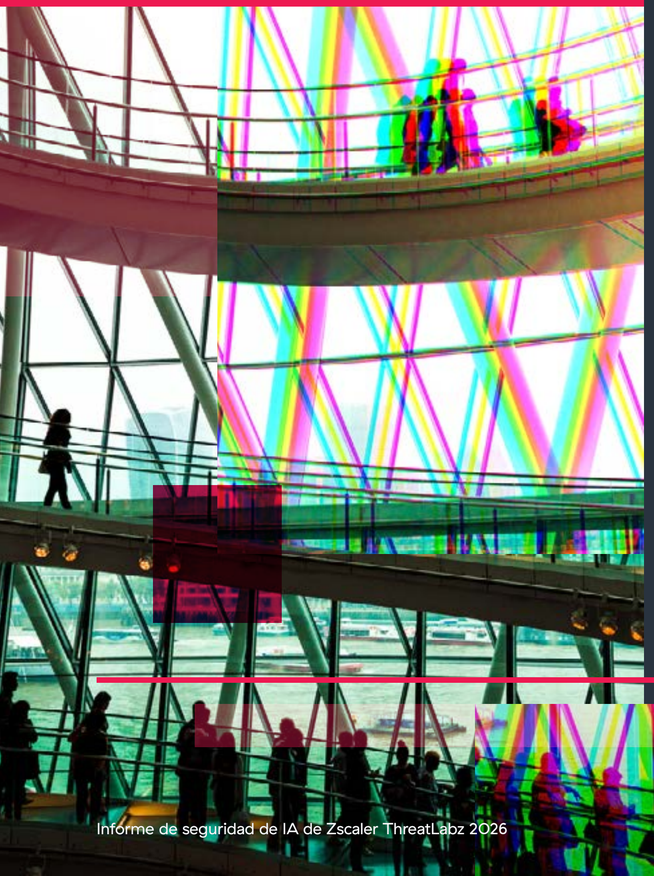
Informe de seguridad de IA 2026 de ThreatLabz



Índice



Resumen ejecutivo_

La realidad cotidiana de la IA en 2025 estuvo definida por la velocidad, la escala y el movimiento constante.

Las empresas ahora confían en la inteligencia artificial y el aprendizaje automático (iA/ML) en toda la empresa para avanzar más rápido, automatizar decisiones y aumentar la productividad. La IA apoya el desarrollo, las comunicaciones, la investigación y las operaciones a un ritmo que no habría parecido realista hace apenas unos años. Pero esta aceleración también conlleva cada vez más contrapartidas: los datos más confidenciales fluyen a través de más aplicaciones IA/ML, a menudo con menos visibilidad y menos protecciones.

Esa expansión de la huella de la IA ha ampliado la superficie de ataque empresarial y los actores de amenazas no tardaron en seguir esos pasos durante el año pasado. Las protecciones más bajas y un mayor realismo han hecho que los ataques sean más rápidos y más convincentes, mientras que los primeros signos de un mal uso de la IA agencial y semiautónoma apuntaban a un cambio en el modo en que evolucionan las amenazas. Al mismo tiempo, las organizaciones se enfrentan a una creciente combinación de riesgos, desde inteligencia artificial oculta e integrada hasta alucinaciones y modelos privados no seguros.

¿Cómo pueden las empresas proteger entornos donde la IA está en contacto con todo, habilitar la innovación impulsada por IA y defenderse contra amenazas impulsadas por IA? (Todo ello sin ralentizar la actividad empresarial, por supuesto).

El informe de seguridad de inteligencia artificial 2026 de ThreatLabz de Zscaler explora cómo las empresas están abordando este acto en búsqueda de equilibrio. El informe se basa en el análisis de 989.300 millones AI/ML transacciones observadas en Zscaler Zero Trust Exchange™ desde enero de 2025 hasta diciembre de 2025, lo que proporciona una visión fundamentada de cómo se utiliza realmente (y se restringe) la IA en entornos globales.

Los datos muestran una aceleración continua. La actividad de IA/ML empresarial aumentó un 83.3 % interanual, mientras que los volúmenes de transferencia de datos aumentaron un 92,6 %, alcanzando más de 18 000 terabytes (TB). A esta escala, la IA se comporta menos como un conjunto de herramientas discretas y más como una infraestructura siempre activa, que mueve y transforma continuamente los datos empresariales. Sin embargo, el acceso está lejos de estar libre de restricciones. Las organizaciones bloquearon el 39 % de las transacciones de IA/ML, lo que refleja preocupaciones persistentes en torno a la exposición de datos, la privacidad y la aplicación de políticas.

Los patrones de uso también revelan dónde se cruzan el valor y el riesgo. Las aplicaciones de inteligencia artificial en las que más confían los empleados, como Codeium, Grammarly y ChatGPT, se encuentran en el centro de cómo se realiza el trabajo, impulsan los niveles más altos de actividad y, al mismo tiempo, aparecen en primer plano en nuestros hallazgos de riesgo.

En 2026, proteger la IA implica mucho más que controlar las aplicaciones de IA/ML. Se trata de proteger el modo en que se descubre, construye, utiliza y gestiona la IA en toda la empresa. Las organizaciones necesitan visibilidad del uso y los riesgos de la IA, protecciones que fortalezcan los sistemas y datos de IA en tiempo real, y controles consistentes que aseguren el acceso, a la vez que impulsan la innovación. Este informe profundiza en las tendencias y realidades que configuran la seguridad de la IA y ofrece orientación a las empresas que buscan reducir el riesgo y adoptar la IA de forma segura.

Lo que esto significa para los líderes empresariales

- **La IA es ahora infraestructura empresarial.** Casi un billón de transacciones de IA indican operaciones continuas y siempre activas. La IA debe gestionarse con el mismo rigor que la nube, la identidad y los datos para facilitar una adopción segura y escalable.

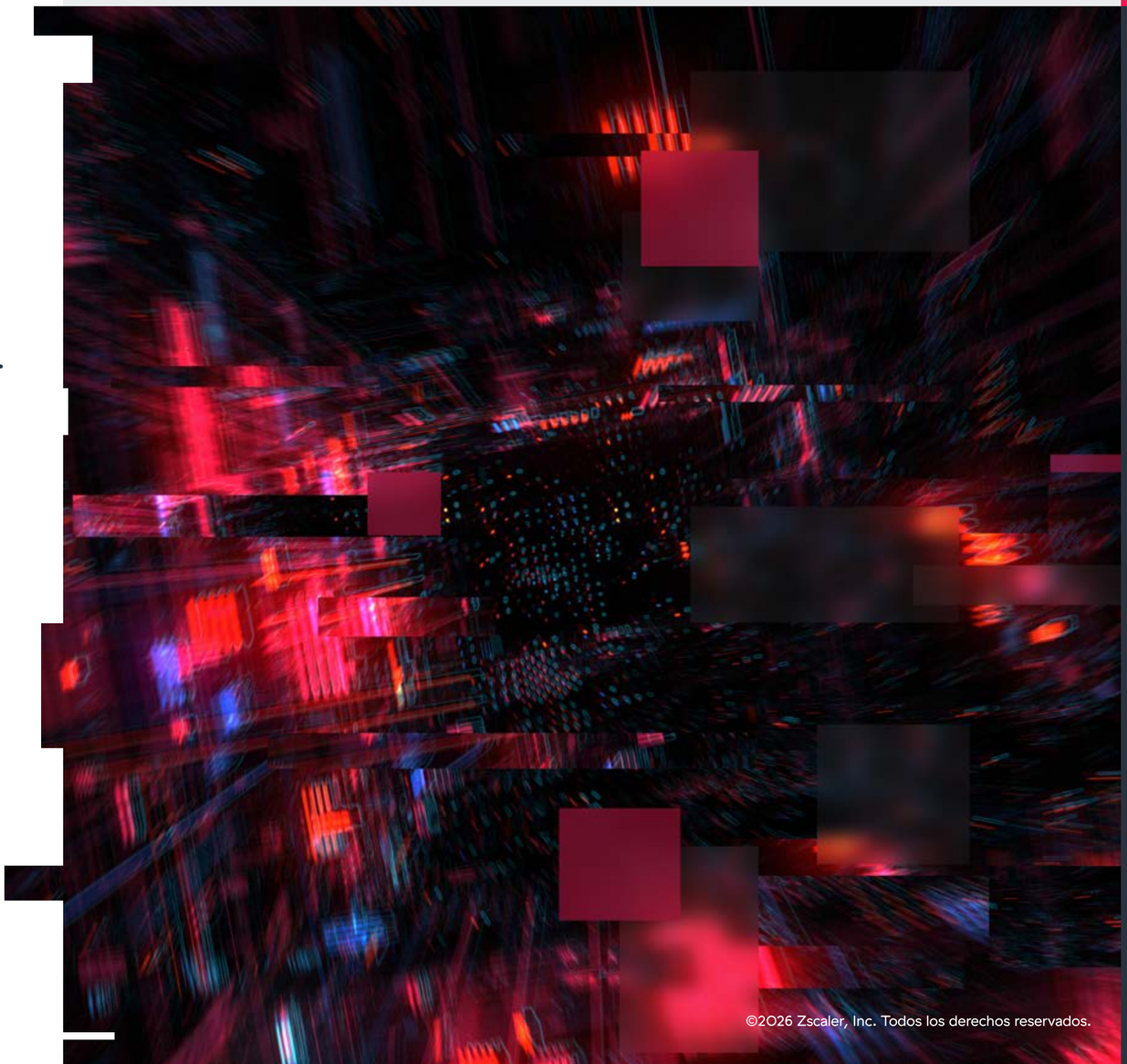
- **El riesgo de exposición de datos ahora escala con el volumen, no con la intención.** El movimiento de datos a escala de petabytes mediante flujos de trabajo de IA aumenta la exposición mediante la repetición y la velocidad, incluso cuando el uso está aprobado y alineado con la intención empresarial.

- **La IA aprobada es la principal superficie de riesgo.** Las herramientas de IA convencionales y autorizadas representan la mayor parte de la actividad de IA empresarial y las interacciones de datos. Si bien la IA en la sombra sigue siendo una preocupación clave, abordar las herramientas no autorizadas por sí solo no mitigará la totalidad de los riesgos y la exposición relacionados con la IA.

- **La seguridad limita la adopción de la IA.** Con el 39 % de las transacciones de IA bloqueadas, la aplicación de políticas está moldeando activamente su uso. Esto refleja la gobernanza en acción, no la resistencia a la IA, ya que los líderes buscan un equilibrio entre la velocidad de innovación y la tolerancia al riesgo.

- **Los modelos de seguridad tradicionales no están alineados con los flujos de trabajo de IA.** Los controles diseñados para la actividad humana y los datos estáticos no pueden seguir el ritmo de las interacciones de IA de alta frecuencia impulsadas por máquinas.

- **La ventaja competitiva favorecerá a las organizaciones que puedan gestionar la IA a gran escala.** Las empresas que permitan un uso amplio de la IA con controles sólidos e integrados avanzarán con mayor rapidez que aquellas obligadas a restringir completamente su uso debido a riesgos no gestionados.



Principales_ hallazgos

ThreatLabz analizó **989,3 mil millones de transacciones de IA y ML** en la nube Zscaler entre enero y diciembre de 2025.
Los hallazgos clave que figuran a continuación se basan en datos que abarcan diferentes períodos de tiempo* para un análisis comparativo.

El uso de IA empresarial continúa su fuerte trayectoria ascendente. La actividad de IA/ML aumentó un 83 % interanual, alcanzando casi un billón de transacciones en un ecosistema de más de 3400 aplicaciones.

Las empresas envían volúmenes de datos cada vez mayores a las herramientas de IA. Se transfirieron un total de 18 033 TB de datos a aplicaciones de IA/ML, un aumento interanual del 93 %.

Las altas tasas de bloqueo indican una gestión de riesgos continua. Las empresas bloquearon el 39 % de las transacciones de IA/ML totales, lo que subraya las continuas preocupaciones sobre la exposición de datos, la privacidad y la alineación de políticas a medida que el uso de IA va en aumento.

La IA empresarial es vulnerable a posibles ataques.
Los expertos en red teaming de Zscaler descubrieron que la mayoría de los sistemas de IA empresarial pueden ser vulnerados en tan solo 16 minutos y descubrieron fallos críticos en el 100 % de los sistemas analizados.

* Periodos de recogida de datos:

- Análisis anual e interanual: enero–diciembre de 2025, con comparaciones interanuales respecto del mismo período de 2024.
- Datos sobre violaciones de la DLP y datos a nivel de país: junio de 2025 a diciembre de 2025.

OpenAI se consolida como el principal proveedor de LLM. OpenAI representó la gran mayoría de las transacciones empresariales impulsadas por LLM (tres veces más que Codeium), lo que lo consolida como el LLM de facto en la actualidad.

ChatGPT representa la gran mayoría de las infracciones de DLP. En todas las aplicaciones IA/ML analizadas, ChatGPT generó 410 millones de violaciones de políticas de prevención de pérdida de datos (DLP), lo que confirma los riesgos empresariales vinculados a los asistentes de IA de alto contexto.

Las aplicaciones de productividad integradas consolidan el uso de la IA empresarial. Grammarly se convirtió en la aplicación número uno por volumen de transacciones, lo que refleja la dependencia de la IA que opera directamente dentro de los procesos de comunicación y negocios.

Las finanzas, los seguros y la manufactura vuelven a liderar el uso de IA empresarial. Por tercer año consecutivo, estos sectores representaron la mayor proporción de tráfico de IA/ML (el 23 % y el 20 %, respectivamente) detrás de sus esfuerzos de modernización y sus pesados flujos de trabajo de documentación.

Estados Unidos siguió siendo la principal fuente de transacciones de IA/ML. La actividad se concentró en EE. UU., que representó el 38 % de las transacciones, seguido de India (14 %) y Canadá (5 %).

La adopción de la IA continúa ampliando la superficie de ataque empresarial. Un uso más amplio de la IA en los flujos de trabajo empresariales ha creado más vías para la exposición de datos y el acceso, lo que aumenta la probabilidad de filtraciones de datos, usos indebidos y ataques asistidos por IA. Esto refuerza la necesidad de una arquitectura de zero trust y controles de seguridad basados en IA.

Tendencias de uso de IA/ML

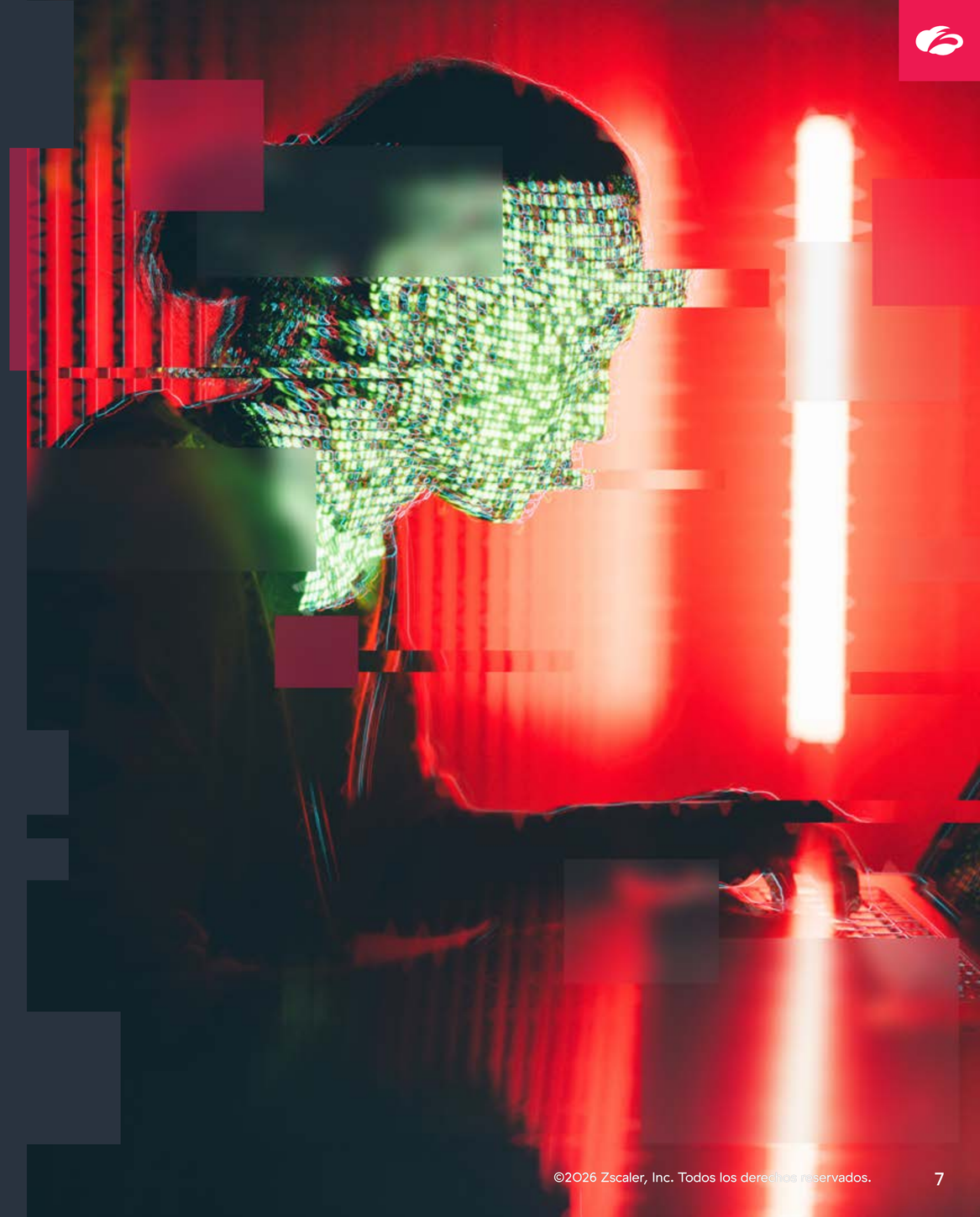
El uso empresarial de IA continuó su ascenso pronunciado y constante en 2025.

El análisis de ThreatLabz sobre las tendencias de uso de IA ahora incluye más de 3400 aplicaciones que impulsan transacciones de IA/ML, cuatro veces más que el año anterior. Si bien muchas de estas aplicaciones generan tráfico limitado, el mero crecimiento del ecosistema de aplicaciones es en sí mismo un indicador significativo. Refleja la rapidez con la que las capacidades de IA están proliferando entre proveedores, casos de uso y funciones comerciales, ampliando tanto las oportunidades como la exposición.

Para comprender cómo este crecimiento se traduce en el uso empresarial en el mundo real, ThreatLabz analizó la actividad de IA/ML en varias capas:

- **Transacciones de IA/ML en general** según la categoría de URL, incluidas tanto las actividades permitidas como las bloqueadas.
- **Clasificación de proveedores de LLM,** que identifica qué proveedores de modelos generan el mayor tráfico de IA/ML y los mayores flujos de IA empresarial.
- **Principales aplicaciones de IA/ML,** destacando las aplicaciones específicas que impulsan la actividad de IA empresarial y el volumen de tráfico.
- **Uso de IA departamental:** mapeo de aplicaciones de IA de gran volumen a departamentos empresariales comunes para comprender dónde se aplica la IA en el trabajo cotidiano.

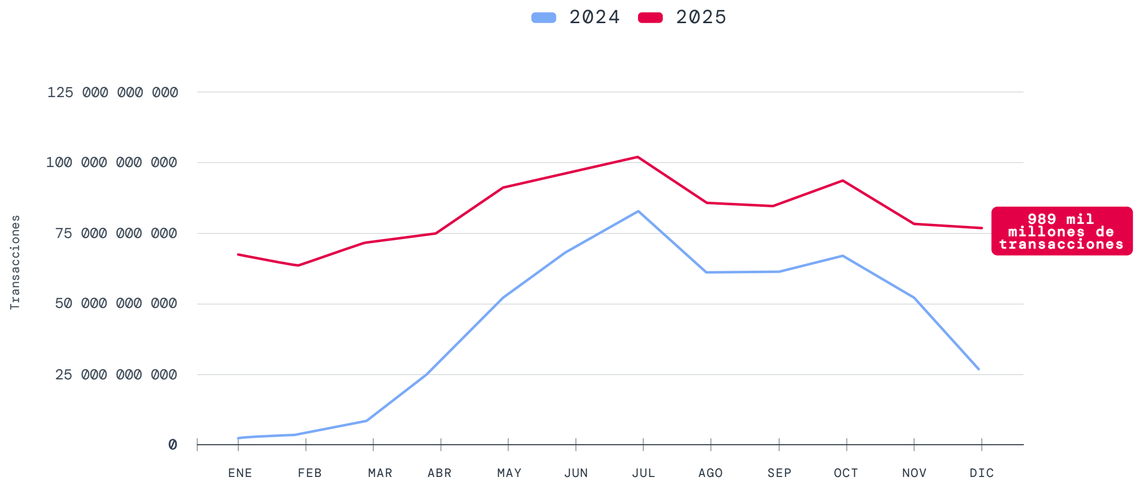
Con estas perspectivas, nuestro objetivo es proporcionar una visión integral de cómo se está adoptando realmente la IA en toda la empresa y dónde convergen el uso, la dependencia y los riesgos.



Crecimiento mundial en transacciones de IA/ML

Las transacciones de IA/ML se aproximaron al billón en 2025, totalizando 989,3 mil millones. Gran parte de este crecimiento está vinculado a aplicaciones de gran volumen como ChatGPT, Grammarly y Codeium.

TENDENCIAS DE USO DE IA POR VOLUMEN DE TRANSACCIONES



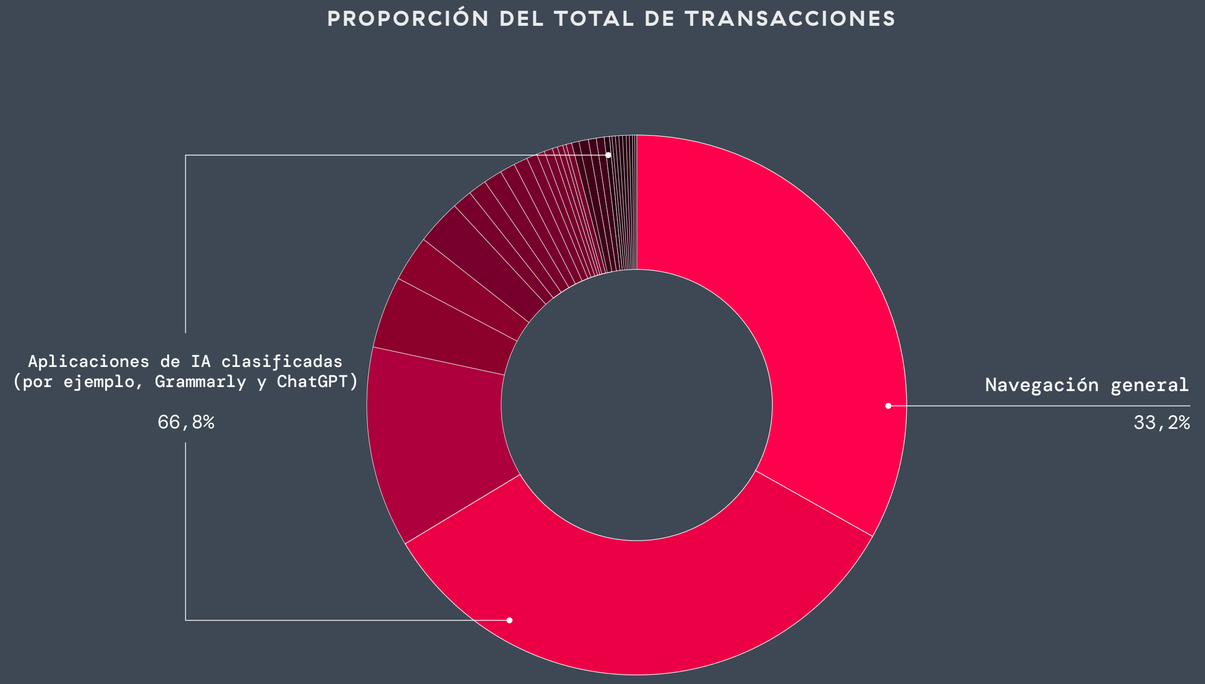
Cifra 1: Comparación interanual de transacciones de IA/ML (enero-diciembre de 2025)

CONCLUSIÓN CLAVE

La actividad de IA/ML aumentó un 83 % interanual en un ecosistema de más de 3400 aplicaciones.

Como en años anteriores, una parte del tráfico se clasifica en “Aplicaciones generales de IA”. Esto refleja transacciones de IA/ML que no se asignan a una aplicación conocida específica, pero que Zscaler identifica como relacionadas con IA mediante la categoización de URL activada por IA/ML, que analiza texto, imágenes y otras señales de contenido para reconocer la actividad relacionada con la IA. Las nuevas aplicaciones de IA surgen más rápido de lo que pueden clasificarse manualmente, lo que hace esencial detectar fuentes de tráfico de IA previamente desconocidas y someterlas a la aplicación de políticas de seguridad.

A menos que se indique lo contrario, el análisis posterior de este informe se centró exclusivamente en aplicaciones clasificadas. Este enfoque nos da visibilidad sobre la adopción de IA a través de aplicaciones de IA/ML.



Cifra 2: Distribución de transacciones de IA/ML en aplicaciones de IA generales y clasificadas

Principales proveedores, aplicaciones y departamentos de LLM

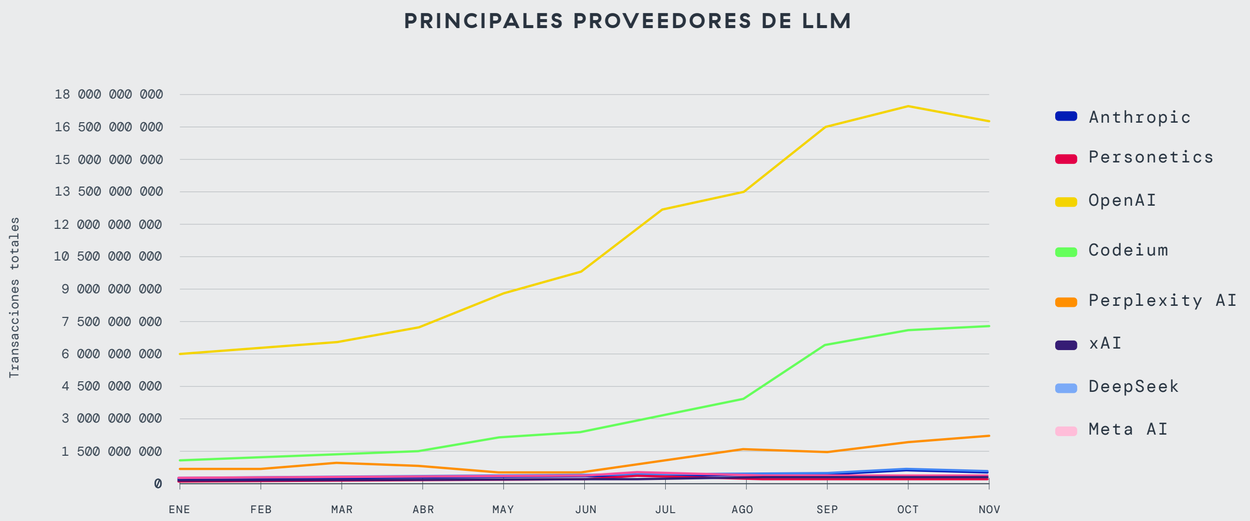
Analizar el uso de la IA empresarial a través de los proveedores de LLM ofrece una visión única de cómo funciona la IA a escala. Mientras los empleados interactúan diariamente con aplicaciones y funciones individuales, los patrones de transacciones muestran qué proveedores de modelos se encuentran consistentemente detrás de esas experiencias. La visibilidad a nivel de proveedor es una forma útil de comprender cómo está tomando forma la adopción de IA bajo la superficie.

Hallazgos clave de los proveedores de LLM

- **OpenAI** fue el líder indiscutible entre los proveedores de LLM en 2025, con 131 000 millones de transacciones, más del triple del volumen de su competidor más cercano. El lanzamiento de GPT-5 en agosto amplió su adopción en la co-dificación, el razonamiento multimodal y la ejecución de tareas complejas. Las opciones ampliadas de la API empresarial de OpenAI, que incluyen mayor privacidad y aislamiento de modelos, también reforzaron su papel como backend para copilotos y funciones SaaS basadas en IA.

- **Codeium** (rebautizada como Windsurf en 2025) se convirtió en la segunda fuente más importante de tráfico LLM empresarial (42 000 millones de transacciones). Su adopción probablemente se debió a sus modelos propietarios centrados en la codificación, frecuentes en los procesos de desarrollo de software y en los entornos de ingeniería. Esto refleja el análisis departamental que se presenta a continuación, donde el departamento de ingeniería destaca como el usuario más activo de IA.

- **Perplexity** ocupó el tercer puesto por volumen de transacciones el año pasado (12 mil millones de transacciones). Además de la búsqueda basada en IA, también opera LLM propios que impulsan su motor de respuestas. En consecuencia, el uso empresarial refleja una creciente dependencia de la investigación y la síntesis de conocimientos asistida por IA.



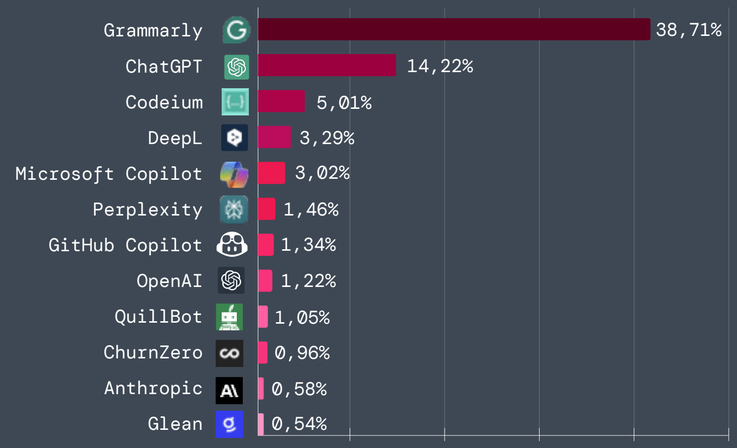
Cifra 3: Tendencias de transacciones de proveedores de LLM a lo largo de 2025

El volumen de transacciones se sigue concentrando principalmente en un conjunto de aplicaciones ampliamente adoptadas que participan directamente en el flujo de trabajo: investigación, edición, redacción, codificación, traducción y colaboración.

Hallazgos clave de la aplicación

- **Grammarly** emergió como la aplicación de IA/ML más activa en entornos empresariales (38,7 % del total de transacciones), superando a ChatGPT en volumen total de transacciones. Con funciones que abarcan desde resúmenes hasta reescritura avanzada y guía de tono, es fácil entender por qué Grammarly es fundamental en los flujos de trabajo cotidianos de contenido empresarial.
- **ChatGPT** siguió siendo un asistente de funciones generales predominante (14,2 %) y se utiliza ampliamente en roles de investigación, redacción y análisis, lo que lo convierte en un punto de contacto común para los datos empresariales.
- **Codeium** entró en el top cinco (5 %), mostrando cómo la IA se ha convertido en una parte habitual del trabajo de desarrollo de software, donde el código fuente y la lógica propietaria se procesan de forma rutinaria.
- **DeepL** siguió experimentando una fuerte adopción en organizaciones de todo el mundo (3,3%), apoyando la comunicación multilingüe en todo el contenido crítico para la actividad empresarial.
- **Microsoft Copilot** completó el top cinco (3 %), gracias a su profunda integración con Microsoft 365 y su papel en la automatización de las tareas de productividad cotidianas.

PRINCIPALES APLICACIONES DE IA



Cifra 4: Porcentaje del total transacciones de IA/ML impulsadas por aplicaciones de IA líderes

Nota: Zscaler Zero Trust Exchange hace el seguimiento de las transacciones de ChatGPT independientemente de otras transacciones de OpenAI en general.

LOS 20 MEJORES APLICACIONES DE IA/ML POR VOLUMEN DE TRANSACCIONES

Aplicación	Transacciones totales
Grammarly	327 311 080 013
ChatGPT	120 227 890 252
Codeium	42 337 652 986
DeepL	27 847 680 087
Microsoft Copilot	25 503 137 940
Perplexity	12 386 054 978
GitHub Copilot	11 348 420 722
OpenAI	10 352 420 115
QuillBot	8 913 115 535
ChurnZero	8 153 526 358
Anthropic	4 922 983 385
Glean	4 542 501 122
GliaCloud	3 249 239 347
Claude	2 850 954 278
Google Gemini	2 604 461 019
SundaySky	2 483 835 170
Yellow Messenger	1 734 555 650
Cresta	1 585 454 178
Poe	1 483 703 558

Más allá de qué aplicaciones de IA dominan el uso general, la siguiente capa de análisis pasa de las herramientas a los equipos.

ThreatLabz mapeó el tráfico de IA/ML a través de un conjunto definido de departamentos empresariales comunes para comprender mejor cómo se utiliza la IA en la práctica. Esta vista se centra en las aplicaciones con un uso sustancial (al menos un millón de transacciones) y las asocia con el departamento en el que se utilizan con más frecuencia. Los porcentajes de participación que se muestran reflejan el uso relativo dentro de este conjunto limitado de departamentos y aplicaciones, en lugar del tráfico total de IA empresarial.

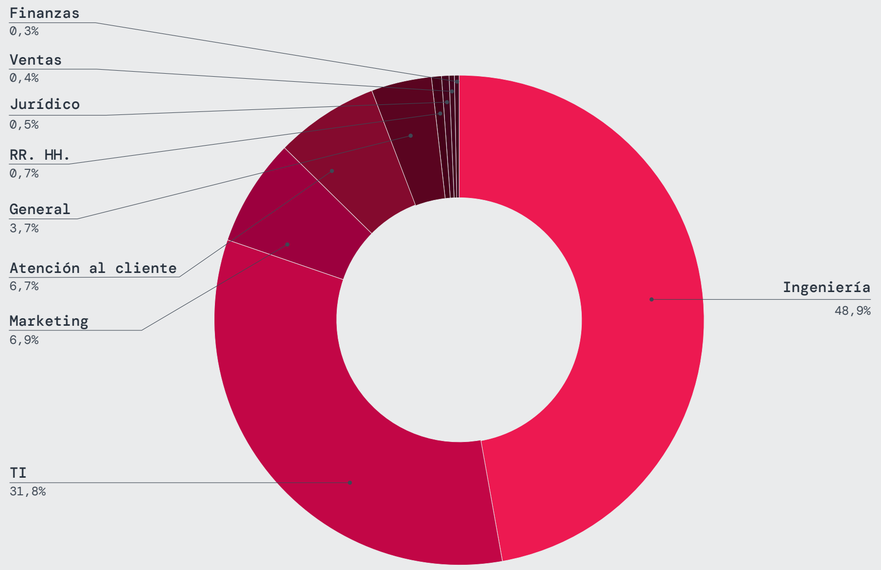
Hallazgos clave del departamento

- **La ingeniería** lideró el uso de IA empresarial, representando el 48,9 % de las transacciones de IA/ML dentro de esta vista delimitada. Los equipos de ingeniería, en particular, integran la IA en los ciclos de compilación diarios, donde incluso pequeñas mejoras de eficiencia se acumulan rápidamente entre versiones.

- **TI** le siguió de cerca como función dependiente de la IA, representando el 31,8 % de la actividad. El uso de la IA en TI suele favorecer la eficiencia operativa, incluyendo el soporte de sistemas, la resolución de problemas y la automatización de procesos internos.

- **Marketing** ocupa el tercer lugar en el uso de IA empresarial (6,9 %) en este análisis. La adopción en marketing se distribuye más entre flujos de trabajo centrados en el contenido y el diseño, lo que resulta en volúmenes de transacciones generales estables, pero inferiores, en comparación con los departamentos técnicos.

PARTICIPACIÓN DE TRANSACCIONES POR DEPARTAMENTO



Cifra 5: Cuota detransacciones de IA/ML por departamentos centrales de la empresa

Transacciones bloqueadas

Las organizaciones también apretaron las riendas de la IA empresarial en 2025. Las preocupaciones sobre la exposición de datos, la privacidad y el cumplimiento los llevaron a bloquear el 39,2 % del total de transacciones de IA/ML, reforzando la gobernanza de la IA como una parte estándar de las operaciones de seguridad cotidianas.

Las aplicaciones más afectadas por los controles de cumplimiento también estuvieron entre las aplicaciones de IA más utilizadas en la empresa. Grammarly representó la mayor parte de la actividad bloqueada: 171 200 millones de transacciones bloqueadas, lo que representó el 44,2 % de todas las transacciones de IA/ML bloqueadas. Las aplicaciones de IA de uso amplio también siguieron bajo vigilancia. ChatGPT y Microsoft Copilot fueron bloqueados con frecuencia, con 5700 millones y 4100 millones de transacciones bloqueadas respectivamente, ya que el acceso a datos no estructurados continúa aumentando el riesgo de que información empresarial confidencial se comparta de forma no intencional.

También se bloquearon con frecuencia los asistentes de codificación de IA, incluidos Codeium y Tabnine, para limitar la exposición de código propietario y artefactos de desarrollo. Las herramientas de transformación de lenguaje y contenido, como QuillBot y DeepL, afrontaron controles similares, lo que refleja esfuerzos más amplios para limitar el intercambio de contenido con modelos externos.

PRINCIPALES APLICACIONES DE IA BLOQUEADAS

1	Grammarly
2	GitHub Copilot
3	ChatGPT
4	Microsoft Copilot
5	QuillBot
6	Codeium
7	DeepL
8	Tabnine
9	Poe
10	Perplexity

Datos transferidos a aplicaciones de IA

El volumen de transacciones por sí solo no refleja completamente cómo las empresas utilizan la IA. Para agregar contexto, ThreatLabz también examinó la cantidad de datos transferidos entre entornos empresariales y aplicaciones de IA/ML.

Durante el último año, la transferencia de datos empresariales a aplicaciones de IA/ML siguieron en aumento y alcanzaron los 18 033 terabytes (TB), un aumento del 93 % interanual. Un subconjunto de las principales aplicaciones ampliamente adoptadas representó la mayor parte de este movimiento de datos. Grammarly siguió siendo la aplicación líder en esta medida, con 3615 TB de datos transferidos. Muy de cerca se situó ChatGPT (2021 TB), seguido por OpenAI (865 TB), DeepL (625 TB) y Codeium (387 TB), aplicaciones que abarcan casos de uso que normalmente manejan datos empresariales de alto valor.

A medida que la IA se integra más en el trabajo cotidiano, más datos empresariales se mueven a través de ella. Analizar tanto el tráfico como el volumen de datos ayuda a descubrir dónde está aumentando el uso de la IA, y dónde la seguridad y la supervisión son más importantes.

PORCENTAJE DE DATOS TRANSFERIDOS

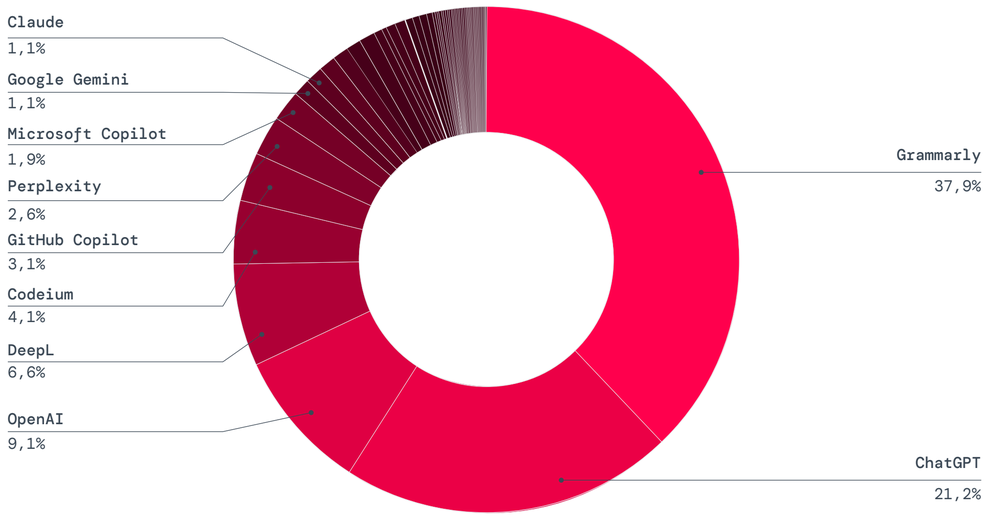
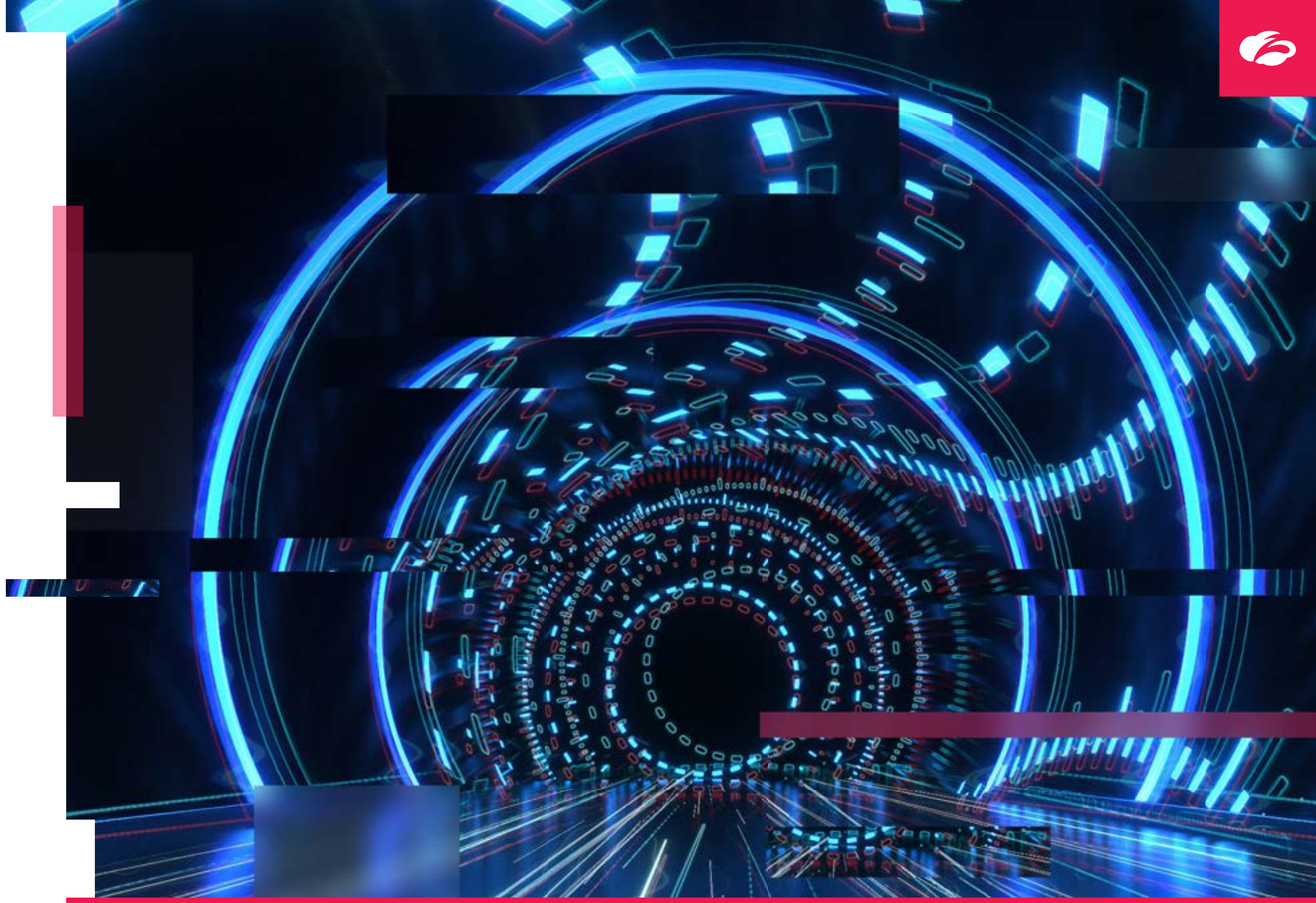


Figura 6: Principales aplicaciones de IA/ML por porcentaje del total de datos transferidos



HALLAZGO CLAVE

Se transfirieron un total de **18 033 TB de datos a aplicaciones de IA/ML**, lo que representa un aumento interanual del 93 %.

Pérdida de datos en aplicaciones de IA

La capacidad de la IA para acelerar el trabajo, desdela concepción de la idea hasta el resultado, en cuestión de minutos conlleva una desventaja importante: los datos confidenciales se pueden compartir con modelos externos en *segundos*. Además, con funciones de IA integradas en aplicaciones y servicios SaaS de uso habitual, el contenido suele transmitirse automáticamente, lo que aumenta la probabilidad de exposición inadvertida.

Prevenir la pérdida de datos en modelos externos se ha convertido en una de las prioridades de seguridad más importantes del año.

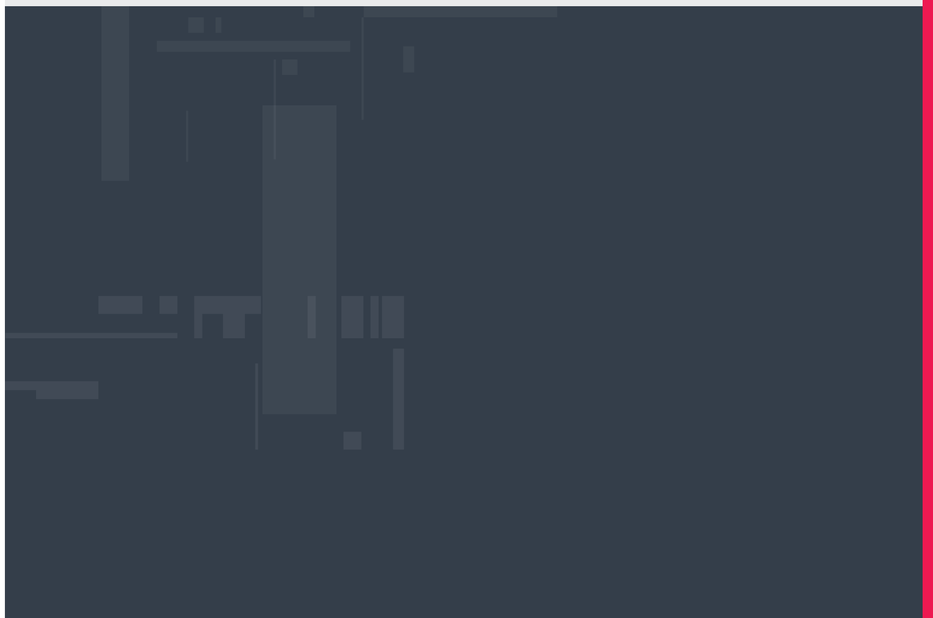
En la nube de Zscaler, las violaciones de las políticas de DLP relacio–nadas con la IA siguen siendo una de las señales más claras de este riesgo creciente. Estas violaciones ocurren cuando información confidencial, como registros financieros, información de identificación personal (PII), código fuente, datos de asistencia sanitaria y otro contenido regulado, intenta salir de la organización a través de una aplicación de IA y las políticas detienen dicha salida. Sin el DLP con reconocimiento de IA de Zscaler, esos datos habrían estado expuestos a modelos de terceros fuera del control de la empresa.

Las aplicaciones de IA más arriesgadas tienden a ser aquellas que los empleados usan sin pensar: asistentes de escritura, asistentes de codificación o funciones de IA integradas en suites de colaboración. Su conveniencia es exactamente lo que los hace más peligrosas: ven el mismo contenido confidencial que los empleados, a menudo en el momento en que se crea.

Las tendencias de violaciones muestran que las interacciones con la IA a menudo involucran parte de los datos más confidenciales de la empresa.

APLICACIONES DE IA/ML CON MÁS INFRACCIONES DE POLÍTICAS DLP

Aplicación	Recuento de infracciones de DLP
ChatGPT	410 181 006
Codeium	242 263 311
GitHub Copilot	31 223 009
Claude	14 417 246
Wordtune	5 161 758
DeepL	2 037 613
QuillBot	1 960 391
Microsoft Copilot	1 858 952
Perplexity	1 235 129
Google Gemini	841 374



Las infracciones de DLP de ChatGPT aumentaron un 99,3 % interanual. Las infracciones más comunes, específicas de ChatGPT, incluyeron la filtración de nombres y el uso de identificadores nacionales (posiblemente registros de clientes o datos de identidad).

Las violaciones de DLP empresarial vinculadas a **Codeium aumentaron un 100 %** interanual lo que sugiere un mayor riesgo de filtración de código fuente y lógica propietaria.

Lo que más destaca entre las principales violaciones de DLP de IA es el alcance global de la exposición. Los identificadores nacionales, los datos de pago, el código fuente y la información médica (cada uno de ellos regulado por estrictas regulaciones regionales) aparecen cada vez más en las interacciones de IA.

LAS 10 PRINCIPALES INFRACCIONES DE LAS POLÍTICAS DE DLP DE IA

1	Filtración de nombre
2	Número de seguridad social (EE. UU.)
3	Número de empresa (Japón)
4	Número del Servicio Nacional de Salud (Reino Unido)
5	Código fuente
6	Número de Medicare (Australia)
7	Número de identificación de proveedor nacional (EE. UU.)
8	Número de Seguro Social (Canadá)
9	Información médica
10	Información de tarjetas de crédito

Estas tendencias de DLP se corresponden con la misma dinámica de fallos observada al probar sistemas de IA en condiciones adversas reales: las averías críticas ocurren, a menudo, mediante interacciones comunes en lugar de ataques sofisticados. Descubra más en **¿Qué está fallando realmente en los sistemas de IA empresarial?**, a continuación.

Para aprender cómo mitigar la pérdida de datos de las aplicaciones de IA generativa, lea **Cómo las empresas están implementando a IA generativa lde manera segura** a continuación.

El auge de la IA integrada

No todo el uso de IA empresarial aparece en herramientas de IA generativa independientes. Cada vez ocurre más a través de IA incorporada: funciones integradas en aplicaciones cotidianas que no están clasificadas como aplicaciones de IA generativa, como resúmenes, recomendaciones o información automatizada que invoca IA únicamente en determinados momentos. Estas capacidades a menudo parecen actualizaciones naturales y esperadas de herramientas que los usuarios ya utilizan. Eso también hace que sea fácil pasar por alto el hecho de que la IA integrada interactúa asimismo con datos empresariales sin la misma visibilidad o las mismas barreras de protección que las aplicaciones de IA independientes, lo que la convierte en una dimensión más silenciosa pero cada vez más importante para asegurar la adopción de la IA. Como resultado, la IA incorporada representa una de las fuentes de riesgo de IA empresarial de más rápido crecimiento y menos visibles.

Este cambio de categoría es importante porque la IA incorporada está diseñada para aumentar la productividad al incorporar más contexto. El mismo principio de diseño también puede aumentar la exposición si la gobernanza y los controles no siguen el mismo ritmo. Los siguientes patrones de amenaza se asocian comúnmente con capacidades de IA integradas en aplicaciones empresariales.

Observaciones clave

USO COMPARTIDO EXCESIVO IMPULSADO POR PERMISOS HEREDADOS

La IA incorporada generalmente depende de controles de acceso y permisos de contenido existentes. Si una organización permite acceso amplio de manera predeterminada, pertenencias a grupo obsoletas o espacios de colaboración sobrecompartidos, la IA incorporada puede revelar involuntariamente información confidencial a usuarios que técnicamente tienen acceso pero no necesitan la información para su función. En la práctica, esto puede convertir una proliferación prolongada de permisos en una exposición de datos más rápida y visible.

MANIPULACIÓN INDIRECTA DE MENSAJES A TRAVÉS DE CONTENIDO EMPRESARIAL

La IA incorporada a menudo lee contenido empresarial, como correos electrónicos, tickets, documentación, registros de chat y archivos adjuntos, como parte del funcionamiento normal. Esto introduce un riesgo en el que las instrucciones ocultas o el contenido adverso pueden influir en cómo responde la IA, qué prioriza o cómo presenta la información. Cuando las funciones de IA están estrechamente integradas en los flujos de trabajo, el contenido en sí mismo puede convertirse en un canal de distribución para la manipulación.

EXPOSICIÓN DE MODELOS Y CONECTORES A LA CADENA DE SUMINISTRO

Las funciones de IA integradas con frecuencia dependen de múltiples componentes. Estos pueden incluir proveedores de modelos, capas de recuperación que extraen contenido de los sistemas empresariales y conectores que se integran en aplicaciones SaaS y repositorios de datos. Cada componente puede introducir nuevos límites de confianza y nuevos vectores de cambio. A medida que las funciones evolucionan, el perfil de riesgo puede cambiar a través de actualizaciones, cambios de configuración o integraciones recientemente habilitadas.

RIESGOS DE ACCIÓN Y AUTOMATIZACIÓN EN FLUJOS DE TRABAJO HABILITADOS POR IA

A medida que las funciones de la IA van más allá del resumen y la redacción y se encargan de la ejecución de tareas, la superficie de riesgo se expande. Si una capacidad de IA puede desencadenar acciones, recomendar cambios, generar código o completar registros, los errores o los resultados manipulados pueden convertirse en problemas operativos. Incluso sin ejecución de acciones directas, los resultados generados por la IA pueden influir en las decisiones y los flujos de trabajo posteriores de formas que son difíciles de auditar.

LAS VULNERABILIDADES DE IA INTEGRADAS EN EL MUNDO REAL PERMITEN UNA FÁCIL EXFILTRACIÓN DE DATOS

Dos ejemplos de exploits ampliamente difundidos en el ecosistema Copilot ilustran cómo la baja interacción del usuario aún puede resultar en un alto riesgo de IA integrada:

- **EchoLeak** se describe como una vulnerabilidad de estilo de inyección de aviso de clic cero en Microsoft 365 Copilot que podría permitir la exfiltración de datos a través de patrones normales de entrada de correo electrónico.
- **Reprompt** es un ataque de un solo clic reportado que utiliza avisos creados a través de parámetros de URL para provocar un comportamiento no deseado y una filtración de datos.

De cara al futuro, a medida que más proveedores de SaaS incluyan IA de forma predeterminada y amplíen las capacidades integradas, las empresas necesitarán ampliar la visibilidad, la gobernanza y la protección de datos de la IA a las aplicaciones y los flujos de trabajo donde la IA opera de forma implícita.

Uso de IA/ML por sector

La adopción de IA aumentó en todos los sectores en 2025, y todos los sectores contabilizados en la nube Zscaler mostraron aumentos interanuales en la actividad de IA/ML. Pero el ritmo y la madurez de la adopción varían mucho. En algunos sectores ya está realizando un trabajo real. En otros, todavía está en proceso.

Las organizaciones **financieras y de seguros** representan la mayor parte (23,3 %) del tráfico de IA/ML por segundo año consecutivo. Los bancos y las aseguradoras son pioneros en la adopción de la IA, dado que sus operaciones giran en torno a los datos, el análisis y la automatización. **El sector manufacturero** se mantuvo en segundo lugar con un 19,5 % de transacciones totales de IA/ML, lo que se puede atribuir a su inversión en automatización impulsada por IA, control de calidad, optimización de la cadena de suministro y más. **La tecnología y las comunicaciones,** y **la educación** registraron los mayores incrementos interanuales, como se refleja a continuación.

PROPORCIÓN DE TRANSACCIONES DE IA POR SECTOR VERTICAL

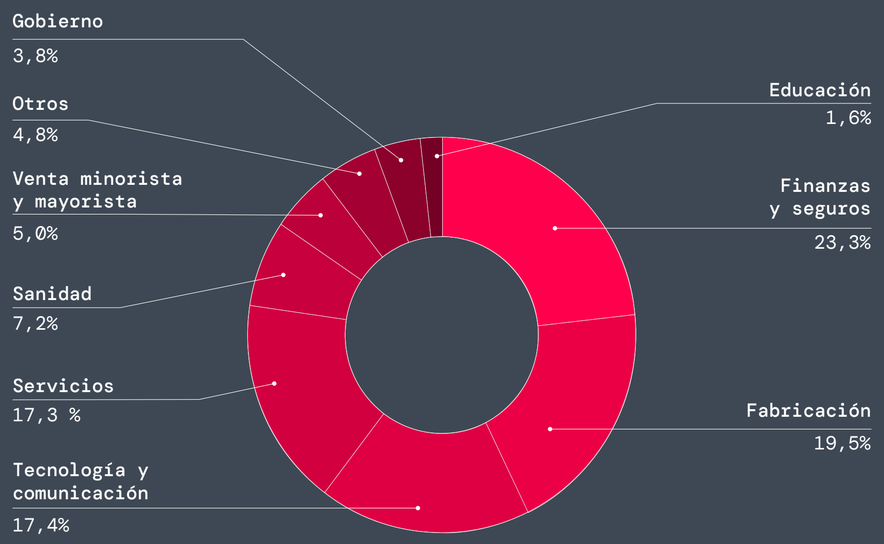


Figura 7: Sectores en los que se daan las mayores proporciones de transacciones de IA

PORCENTAJE DE TRANSACCIONES DE IA BLOQUEADAS POR SECTOR VERTICAL

Sector vertical	% de transacciones de IA bloqueadas
Finanzas y seguros	39,1 %
Fabricación	22,1 %
Servicios	13,5 %
Sanidad	8,5 %
Tecnología y comunicación	6,8 %
Gobierno	4,0 %
Otros	3,4 %
Venta minorista y mayorista	2,0 %
Educación	0,6 %

El uso de IA no se produce en el vacío; está influenciado por el riesgo específico del sector, las expectativas de cumplimiento y hasta qué punto han evolucionado los programas de seguridad.

Los patrones en el bloqueo de transacciones de IA/ML revelan cuán diferente es el modo en que los sectores equilibran la adopción de IA con la gestión de riesgos. Las finanzas y el sector de seguros no solo generaron la mayor parte de la actividad de IA, sino que también bloquearon aproximadamente el 40 % de esas transacciones. La alta tasa de bloqueo refleja algo más que simple cautela: es la realidad de operar en un entorno fuertemente regulado donde se esperan controles más estrictos sobre el uso de IA.

El sector manufacturero, el segundo más activo por volumen de transacciones de IA, bloqueó aproximadamente el 22 % de su tráfico de IA. Esto sugiere un punto medio pragmático, ya que los fabricantes implementan la IA extensamente, pero aún aplican una supervisión significativa para evitar el uso indebido y proteger contra la filtración de datos, especialmente en entorns IoT/OT.

SECTORES DESTACADOS

Las finanzas y los seguros siguen siendo el sector que más utiliza la IA: 230 000 millones de transacciones

Las finanzas y el sector de seguros fueron los que tuvieron una mayor actividad de IA en la nube de Zscaler en IA/ML. Representan casi una cuarta parte del uso empresarial total. Gran parte de este volumen proviene de herramientas de productividad cotidianas. Grammarly, ChatGPT y Microsoft Copilot fueron las aplicaciones de IA más utilizadas en bancos y aseguradoras por segundo año consecutivo. Los equipos de todas las organizaciones utilizan estas herramientas para resumir investigaciones, gestionar documentación de cumplimiento, detectar fraudes, agilizar las reclamaciones, respaldar la suscripción de seguros y realizar otras tareas esenciales. Estas tendencias se reflejaron en el impulso general del sector. Según la encuesta 2025 AI Adopter de Morgan Stanley,[1] la adopción de la IA en seguros aumentó del 48 % al 71 % para mediados de año, y del 66 % al 73 % en las empresas de servicios financieros.

Esta aceleración se vio reforzada por varias fuerzas del mercado de 2025. Los bancos se encuentran bajo la presión de costos y modernización, lo que los impulsa a poner en funcionamiento la IA más rápido que la mayoría de los otros sectores. Las compañías de seguros se enfrentan a una creciente gravedad de las reclamaciones y a una volatilidad impulsada por el clima, por lo que recurren a la IA para mejorar la precisión de los precios y los tiempos de respuesta.

Al mismo tiempo, el sector dista de estar despreocupado en cuanto al modo en que utiliza estas herramientas. Las finanzas y los seguros también bloquearon más del 39,1 % de transacciones de IA/ML en la nube Zscaler, una señal de una mayor sensibilidad al riesgo de pérdida de datos, el escrutinio regulatorio y la necesidad de gestionar de forma estricta las interacciones de los modelos con información financiera confidencial. Se mueven rápido, pero con los frenos a mano.

Las finanzas y os seguros seguirán definiendo cómo será la ambiciosa transformación de la IA en 2026.

[1] Business Insider, **3 partes del mercado donde la publicidad sobre IA se está convirtiendo en resultados reales, según Morgan Stanley**, 24 de julio de 2025.

©2026 Zscaler, Inc. Todos los derechos reservados.

SECTORES DESTACADOS

La tecnología registra el crecimiento más rápido en el uso de IA empresarial: más de un 202 % interanual

El sector tecnológico registró el mayor incremento interanual en transacciones de IA/ML en 2025 (202,3 %), superando a todos los demás sectores en la nube Zscaler. Si bien el sector tecnológico siempre ha sido un usuario activo de la IA (como usuario pionero y entusiasta de la IA generativa), el aumento de este año refleja con qué intensidad las empresas de software, los proveedores de la nube, las plataformas digitales y los equipos de ingeniería están integrando la IA tanto en sus productos como en sus flujos de trabajo internos.

Los asistentes de productividad líderes se utilizan intensamente en las organizaciones tecnológicas y potencian todo, desde la generación de código y la documentación técnica hasta el contenido de marketing. En consecuencia, Grammarly, Codeium, ChatGPT y Perplexity estuvieron entre las principales aplicaciones de IA utilizadas en el tráfico del sector tecnológico durante nuestro análisis.

Incluso con este rápido crecimiento, para muchas organizaciones tecnológicas, la IA está exponiendo brechas en la visibilidad y la aplicación de políticas. En respuesta, están invirtiendo más en supervisión y bloqueando aproximadamente el 7 % de las transacciones de IA (todavía una proporción relativamente pequeña en general, pero notablemente más alta que la de muchos otros sectores) a medida que perfeccionan los controles para respaldar una implementación segura.

SECTORES DESTACADOS

La educación muestra un crecimiento silencioso pero impresionante en la adopción de IA: más de un 184 % interanual

El sector de la educación representó únicamente una pequeña parte del total transacciones de IA/ML en la nube Zscaler en 2025, pero su tasa de crecimiento relata una historia diferente. La educación generó casi 16 mil millones de transacciones de IA/ML durante el año, registrando el segundo mayor incremento interanual en actividad de IA/ML con un 184,4 % que lo convierte en uno de los adoptantes de IA de más rápida crecimiento de todos los sectores.

Este aumento se alinea estrechamente con el uso cada vez mayor de la IA generativa en el aprendizaje y los flujos de trabajo en el aula. Los estudiantes y el personal utilizan ampliamente aplicaciones como ChatGPT y Microsoft Copilot para obtener ayuda con la escritura, la creación de contenido y la planificación de lecciones. Los administradores también están utilizando IA para agilizar tareas rutinarias, desde la redacción de comunicaciones hasta la mejora de los servicios para los estudiantes, lo que probablemente contribuye al aumento constante del volumen de transacciones.

Cabe destacar que este aumento se produjo con muy pocas complicaciones. Se bloqueó menos del 1 % de las transaccionew de IA/ML en educación, lo que sugiere que la mayor parte del uso está explícitamente permitido u ocurre en entornos donde la gobernanza y las barreras de protección aún están surgiendo. Esto hace que el sector educativo se muestre comprensiblemente reservado en comparación con sectores más grandes. Las escuelas y universidades deben abordar las preocupaciones sobre la privacidad de los datos y la integridad académica. Es probable que estos factores hayan mantenido el uso general de IA más bajo que en otros sectores, incluso cuando su adopción va en rápido aumento.

Aun así, un crecimiento de casi el triple en un solo año prepara el escenario para iniciativas de IA más estructuradas y responsables, y para su integración en el próximo año.

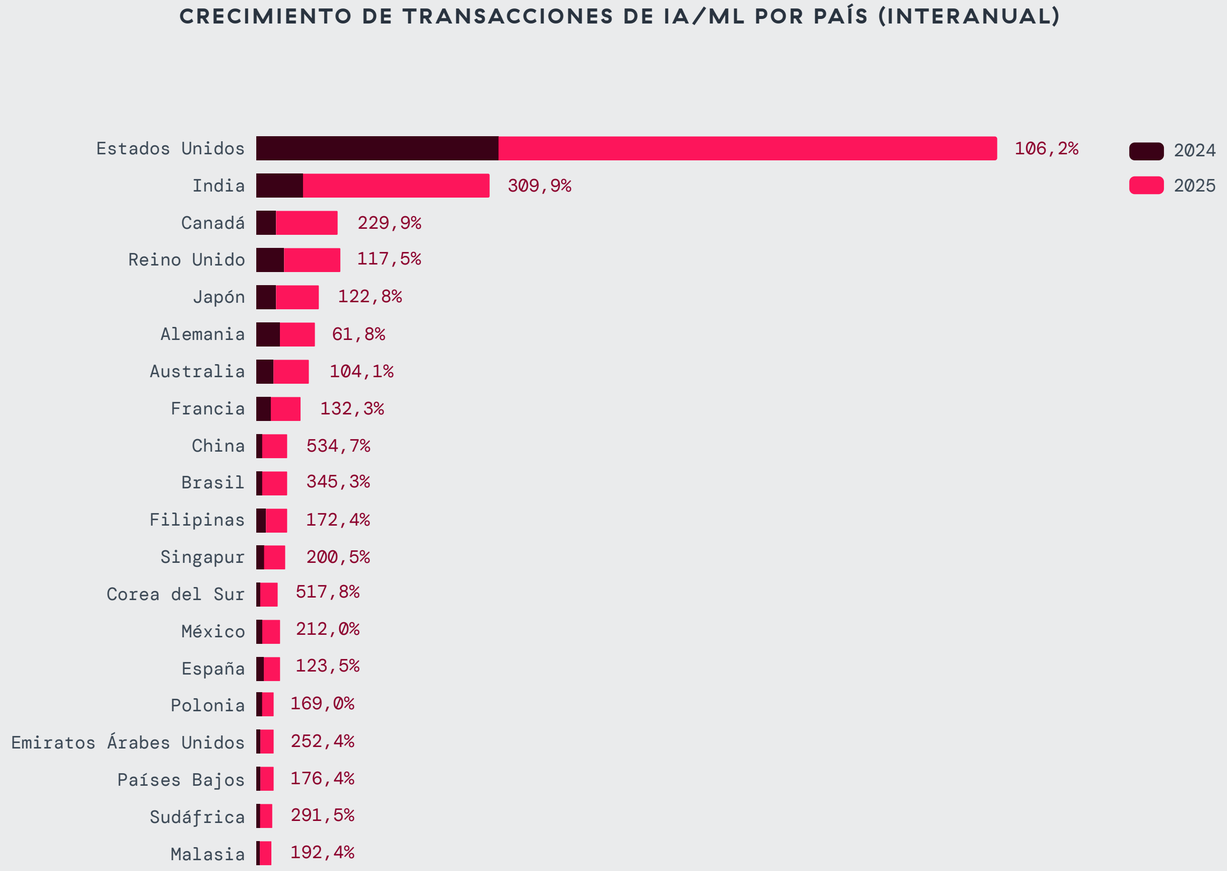
Uso de IA/ML por país

La distribución geográfica de la actividad de IA/ML se mantuvo prácticamente constante en 2025, con cambios sutiles en los márgenes. La IA está firmemente establecida en **Estados Unidos**—el epicentro del desarrollo e implementación de la IA empresarial— y el país sigue representando la mayor parte del volumen de tráfico de IA/ML Sin embargo, el uso de IA creció significativamente en varios mercados internacionales.

Aunque Estados Unidos siguió liderando en uso absoluto (218,9 mil millones transacciones de IA/ML, que representan el 37,6 % de la actividad global), la adopción de IA tuvo un crecimiento interanual más rápido en otros lugares. Esa aceleración global es más evidente en **India**, que fue la segunda mayor fuente de actividad de IA empresarial, alcanzando 82,3 mil millones de transacciones, un aumento interanual del 309,9 %. El crecimiento en India se alinea con los continuos esfuerzos de transformación digital respaldados por el gobierno en 2025, junto con una importante inversión pública y privada en infraestructura de IA y desarrollo de habilidades. Un personal laboral en expansión habilitado para la IA, combinada con arquitecturas de nube que permiten una implementación rápida y escalable de servicios de IA, probablemente contribuyó al crecimiento descomunal en el país con respecto a años anteriores.

Más allá de estos dos mercados principales, varios otros mercados maduros reforzaron la tendencia hacia una expansión constante de la IA, impulsada por las empresas. **Canadá** generó 27 200 millones de transacciones (más de un 229,9 % interanual), respaldado por la inversión federal en computación y programas de IA destinados a acelerar la adopción por parte de las empresas, especialmente en sectores regulados. El **Reino Unido** y **Japón** completaron los cinco primeros puestos, con aumentos del 117,5 % y el 122,8 %, respectivamente.

Esta amplia presencia en diferentes ámbitos geográficos refleja la transición de la IA hacia una capacidad empresarial estándar en todo el mundo. Los equipos de seguridad deben tener en cuenta esta presencia de uso más diseminada y garantizar una supervisión constante en todos los entornos geográficos.



Cifra 8: Crecimiento interanual en transacciones de IA/ML por país (los principales 20 según volumen de transacciones)

País	% compartir	Transacciones de IA/ML
Estados Unidos	37,6 %	219 B
India	14,1 %	82 B
Canadá	4,7 %	27 B
Reino Unido	4,3 %	25 B
Japón	3,2 %	19 B
Alemania	2,7 %	16 B
Australia	2,6 %	15 B
Francia	2,4 %	14 M
China	2,0 %	12 B
Brasil	1,8 %	11 B

Cifra 9: Mapa que muestra los 10 países principales según el volumen de transacciones de IA/ML (tabla a la derecha: totales de porcentaje de participación y de volumen de junio a diciembre de 2025)

Perspectivas de EMEA

La actividad de IA/ML en la región EMEA se mantuvo concentrada en un pequeño número de mercados europeos maduros. El Reino Unido, Alemania, Francia y España representaron casi la mitad de las transacciones regionales. Si bien el Reino Unido representa una parte menor de la actividad de IA en el mundo, tiene sisemáticamente una participación desproporcionadamente grande dentro de EMEA, liderando la región con un 20,3 % del tráfico de IAI/ML entre junio y diciembre de 2025.

Alemania le siguió con el 12,5 % de las transacciones de EMEA, impulsadas por la continua integración de IA en la fabricación, que generó más de 5500 millones de transacciones de IA/ML. Muy de cerca, Francia representó el 11 % de la actividad regional, sostenida por iniciativas gubernamentales como la estrategia Francia 2030, que incluye importantes compromisos de inversión en IA y sirve como sede de la Cumbre de Acción Internacional sobre IA.

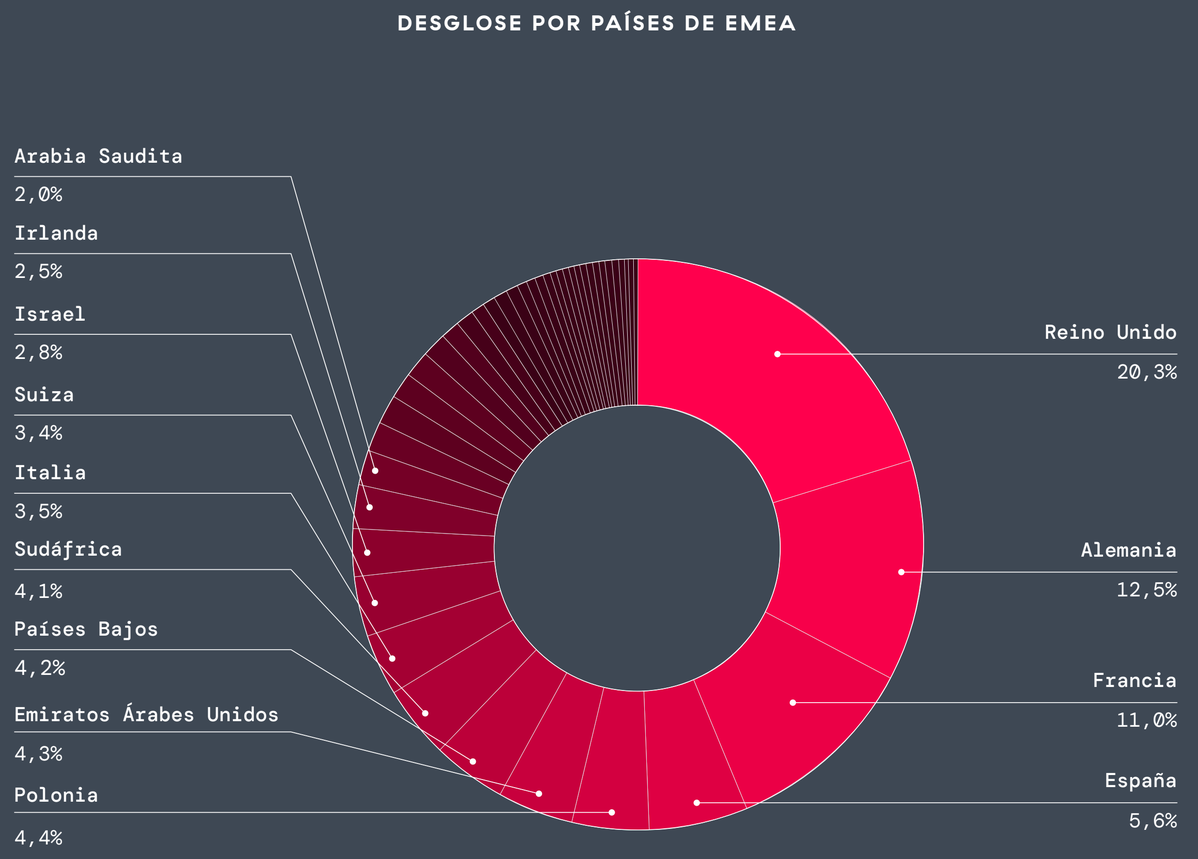


Figura 10: Porcentaje de transacciones de IA por país en la región EMEA

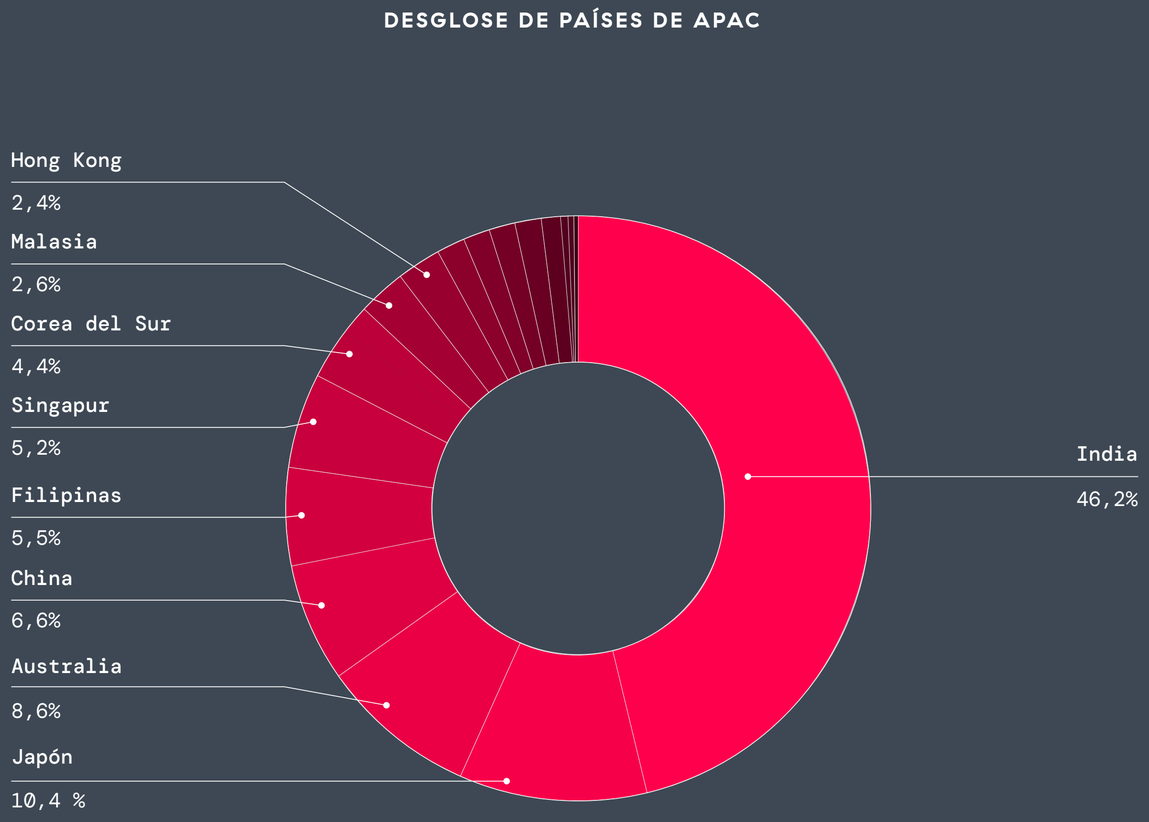


Figura 11: Porcentaje de transacciones de IA por país en la región APAC

INSTANTÁNEAS REGIONALES

Perspectivas de APAC

El uso de la IA en la región Asia-Pacífico (APAC) estuvo determinado por un marcado desequilibrio entre un único mercado de alto crecimiento y varias economías más establecidas. India, Japón y Australia recogieron la mayoría de las transacciones de IA/ML de la región. La India por sí sola acoge casi la mitad de toda la actividad (el 46,2 % del tráfico de IA/ML regional), impulsada en gran medida por el sector de Tecnología y comunicación (31 mil millones de transacciones).

Japón le siguió con el 10,4 % de las transacciones de APAC en el contexto de una política nacional sobre IA en evolución. El gobierno japonés aprobó una ley nacional de promoción de la IA que fomenta la adopción de IA a nivel empresarial e industrial mediante una orientación coordinada. Australia representó el 8,6 % de la actividad regional junto con un énfasis nacional continuo en la implementación responsable y segura de la IA.

Panorama de riesgos y amenazas de la IA empresarial

Como demuestra nuestra investigación, la IA está presente en cada capa de la actividad empresarial, desde las herramientas públicas de IA generativa hasta los LLM internos y las suites SaaS habilitadas para IA. Las organizaciones deben gestionar una superficie de ataque más amplia y compleja a medida que crece el uso. Los riesgos más importantes se dividen en las siguientes categorías:

Exposición de datos y filtración de información confidencial

Los sistemas de IA ven algunos de los datos más confidenciales de la empresa (código fuente, registros de clientes, detalles financieros y documentos legales), a menudo sin medidas de seguridad claras. Esta exposición generalmente surge del uso de IA en la sombra en herramientas públicas como ChatGPT, Grok y DeepSeek, así como de IA SaaS con demasiados permisos, como Microsoft Copilot, que genera datos debido a configuraciones incorrectas o etiquetas inexactas. Paralelamente, los canales de recuperación y generación aumentada (RAG) no controlados pueden extraer silenciosamente datos regulados hacia modelos privados. Una vez que se envía información confidencial a un sistema de IA, esta puede conservarse, reutilizarse o incluso exponerse mediante una manipulación inmediata o un comportamiento del modelo, lo que convierte el uso cotidiano de la IA en un riesgo real para los datos.

Falta de visibilidad sobre el uso de la IA y las indicaciones para los usuarios

Muchas organizaciones aún tienen dificultades para responder preguntas básicas sobre cómo se utiliza realmente la IA día a día. Los equipos de seguridad a menudo carecen de una visión clara de qué herramientas de IA utilizan los empleados, qué solicitudes envían y si los datos confidenciales están en peligro. Tampoco siempre es obvio qué equipos confían en la IA generativa para flujos de trabajo críticos. Cuando se revisan los mensajes, a menudo se aprecian intentos de inyección de mensajes, patrones de manipulación o comportamiento no conforme que eluden las barreras de protección con un mínimo esfuerzo. Pero la mayoría de las organizaciones no tienen las herramientas para observar esta actividad en tiempo real. Como resultado, la gobernanza de la IA tiende a ser reactiva y a entrar en acción solo cuando ya ha surgido un problema.

Calidad de datos, alucinaciones y manipulación de modelos

Con la IA integrada en las operaciones comerciales cootidianas, los errores en sus resultados tienen consecuencias reales. En 2025, las organizaciones tuvieron que corregir alucinaciones donde las directrices generadas por IA parecían reales, pero resultaron ser erróneas. Los sistemas respaldados por RAG también han producido resultados sesgados debido a entradas distorsionadas o de baja calidad, especialmente en equipos centrados en el cumplimiento normativo. **Ejercicios de red teaming y pruebas en condiciones reales** han demostrado cómo los atacantes pueden contaminar los canales de recuperación insertando contenido manipulado en las fuentes que los sistemas de IA ingieren o explotando debilidades de afianzamiento y precisión mediante una sutil variación de las indicaciones. Las alucinaciones, la variación implícita y los fallos de afianzamiento minan constantemente la confianza en los resultados de la IA. Cuando estos fallos no se detectan, los resultados defectuosos pueden influir directamente en las decisiones y aumentar el riesgo.

Modelos de IA privados no mapeados y no seguros

Las empresas ahora implementan una combinación de modelos administrados y no administrados, y capacidades de IA integradas en plataformas como Salesforce, ServiceNow y Atlassian.

Sin embargo, muchas organizaciones aún carecen de:

- Un inventario completo de modelos y servicios
- Comprensión de qué datos toca cada modelo
- Validación de la seguridad del modelo, niveles de parches o estado de vulnerabilidad
- Gobernanza para repositorios de código fuente que alimentan flujos de trabajo de IA

Esta falta de mapeo se vuelve especialmente peligrosa cuando los modelos privados heredan las mismas debilidades de inyección rápida, envenenamiento de RAG y filtación de datos observadas en los sistemas públicos. Cuando se desconocen los modelos y sus flujos de datos, las organizaciones no pueden aplicar políticas ni evaluar los riesgos de manera significativa.

Privacidad, cumplimiento y variabilidad del proveedor

Los proveedores de IA adoptan diferentes enfoques para gestionar los datos empresariales. Las indicaciones se pueden almacenar, reutilizar para capacitación o registrar de maneras que no siempre son claras. Los controles de acceso y el linaje de modelos varían ampliamente de un proveedor a otro. Esta inconsistencia crea desafíos de cumplimiento en marcos como RGPD, HIPAA y PCI DSS. El riesgo se agrava a medida que las aplicaciones SaaS entregan funciones de IA predeterminadas que eluden los procesos de aprobación establecidos, lo que hace que las políticas empresariales no estén alineadas con las expectativas regulatorias.

Amenazas y vulnerabilidades del mundo real

Los principales riesgos de la adopción de IA empresarial siguieron manifestándose en formas reales en 2025. Preocupaciones como la exposición de datos, la visibilidad limitada del uso de IA, las alucinaciones y otras surgieron como amenazas de seguridad tangibles y vulnerabilidades operativas en los entornos empresariales. Los incidentes reales y los resultados de las pruebas demostraron que estos riesgos surgen de cómo se implementan los sistemas de IA, cómo se conectan a los datos y cómo se confía en ellos dentro de los flujos de trabajo diarios.

Algunos de los riesgos subyacentes más importantes se manifiestan en la ingeniería social habilitada por IA, la filtración de datos a través de aplicaciones y asistentes de IA y el mal uso temprano de sistemas de IA agentes y semiautónomos.

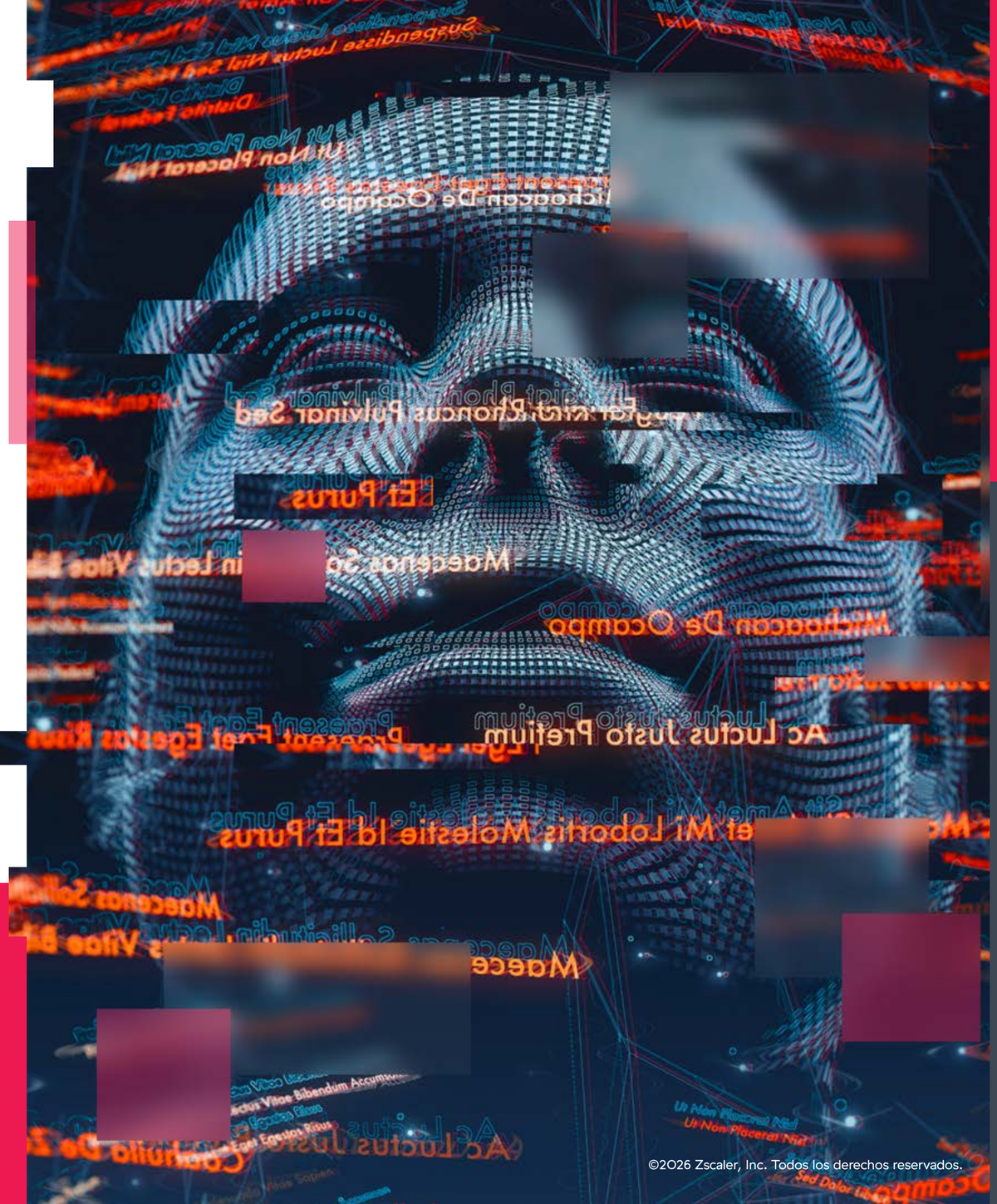
La ingeniería social basada en IA se intensificó a medida que los atacantes utilizaban la IA generativa para lograr suplantaciones de identidad más convincentes. El phishing de voz y vídeo deepfake («vishing») se convirtió en un problema documentado en 2025. En múltiples avisos, incluidas advertencias de las autoridades estadounidenses, se observó a actores de amenazas suplantando a funcionarios mediante voces y mensajes generados por IA.[2]

Los atacantes utilizan la IA para producir vídeos y voces deepfake convincentes, adaptados a roles y procesos de decisión específicos.

El año pasado también se presentó el primer informe creíble de una **campaña de ciberespionaje con IA agéntica**. Un grupo patrocinado por el estado chino automatizó entre el 80 % y el 90 % de la cadena de intrusión con IA agéntica, incluyendo el reconocimiento, la validación de exploits, la recolección de credenciales, el movimiento lateral y la exfiltración de datos. Los operadores humanos solo intervinieron para tomar decisiones de nivel más alto. Este incidente demostró cómo los agentes autónomos pueden ejecutar la estrategia de ataque tradicional, pero a la velocidad de una máquina, lo que alteró fundamentalmente la forma en que los defensores deben detectar y responder a las amenazas.

Más allá del abuso directo de los sistemas de IA, los atacantes comenzaron a incorporar IA en sus propios flujos de trabajo de desarrollo. En varias campañas observadas por ThreatLabz, el malware mostró características consistentes con la generación de código asistida por IA, lo que sugiere que la IA generativa se utiliza cada vez más en los ataques.

Los siguientes estudios de caso ponen en evidencia el riesgo de la IA: desde el engaño y la ejecución de ataques habilitados por la IA generativa hasta las pruebas de red team que revelan cómo funcionan los sistemas de IA empresariales en condiciones adversas reales.



[2] Investigación sobre ciberseguridad: **el FBI advierte que altos funcionarios estadounidenses están siendo suplantados mediante mensajes de texto y clonación de voz basada en inteligencia artificial**, 16 de mayo de 2025.

CASO PRÁCTICO

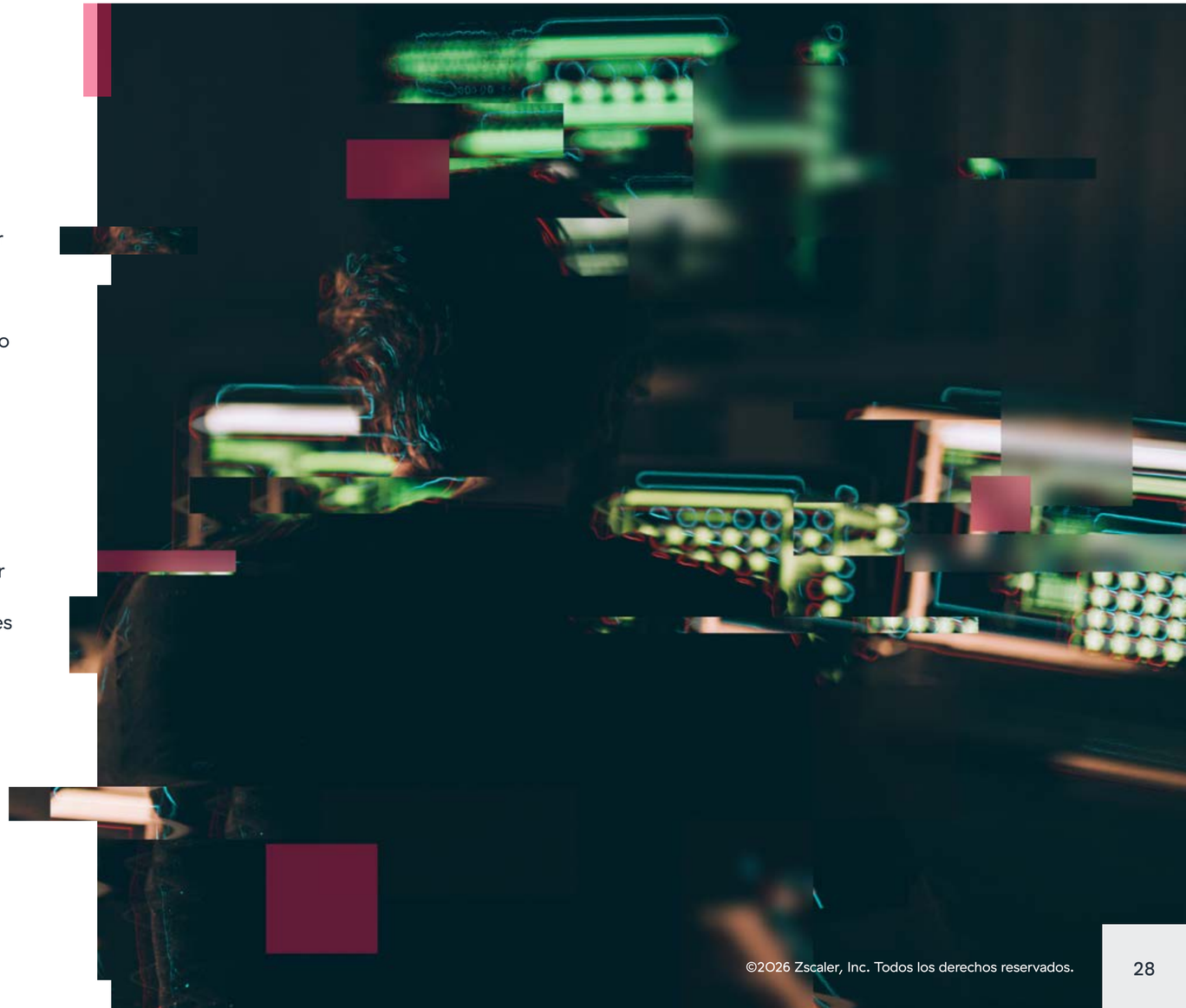
Malware mejorado con A generativa Ie ingeniería social en campañas vinculadas a la RPDC

Este estudio de caso destaca cómo IA generativa permite a los atacantes reforzar sus operaciones sin cambiar fundamentalmente sus objetivos o técnicas.

En la **campaña "Entrevista contagiosa"**, vinculada a actividades relacionadas con la República Popular Democrática de Corea y al programa más amplio de Trabajadores de TI de la RPDC, ThreatLabz observó cómo los actores de amenazas utilizan IA generativa como arma para industrializar la ingeniería social (creando y operativizando perfiles falsos convincentes). Al mismo tiempo, emplean en el desarrollo de malware codificación asistida por IA que dificulta la distinción entre el acceso de los atacantes entran y sus acciones posteriores de su actividad legítima, lo que eleva la necesidad de detección y respuesta.

Desarrollo de recursos e ingeniería social (Engaño en las entrevistas)

La campaña comienza con la fabricación de identidades digitales utilizando tecnología de IA generativa, creando guías de estudio completas, generando fotos de perfil profesionales, pero imposibles de rastrear y empleando herramientas de deepfake y manipulación de voz para enmascarar sus identidades durante entrevistas remotas. Este engaño está diseñado para eludir los procesos de investigación y asegurar posiciones técnicas confidenciales.



Los siguientes hallazgos subrayan en qué medida la fase de preparación de la entrevista de la operación depende de la IA.

GUÍAS DE ESTUDIO GENERADAS POR IA PARA DOMINAR LAS ENTREVISTAS

Los actores de amenazas producen manuales instructivos detallados utilizando IA generativa para prepararse para entrevistas técnicas.

Ejemplo: una sola “guía de estudio” consta de más de 70 páginas y cubre preguntas complejas en campos como ingeniería de back-end y desarrollo Web3.

Indicadores clave de la IA:

- Las respuestas en las guías incluyen frases distintivas, como “¡Ciertamente!” (figura 12).
- Elementos residuales del formato Markdown, que sugieren encarecidamente una acción directa de copiar y pegar desde la salida generada por el modelo de IA (figura 13).

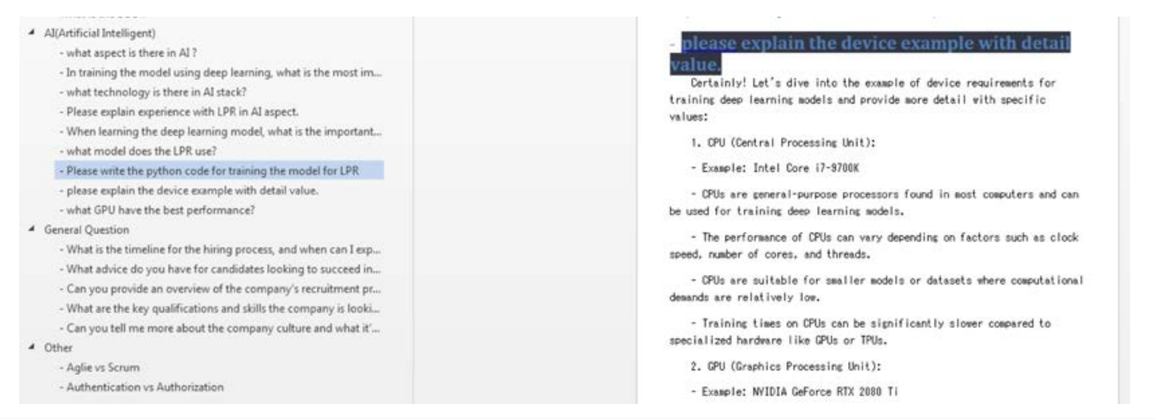
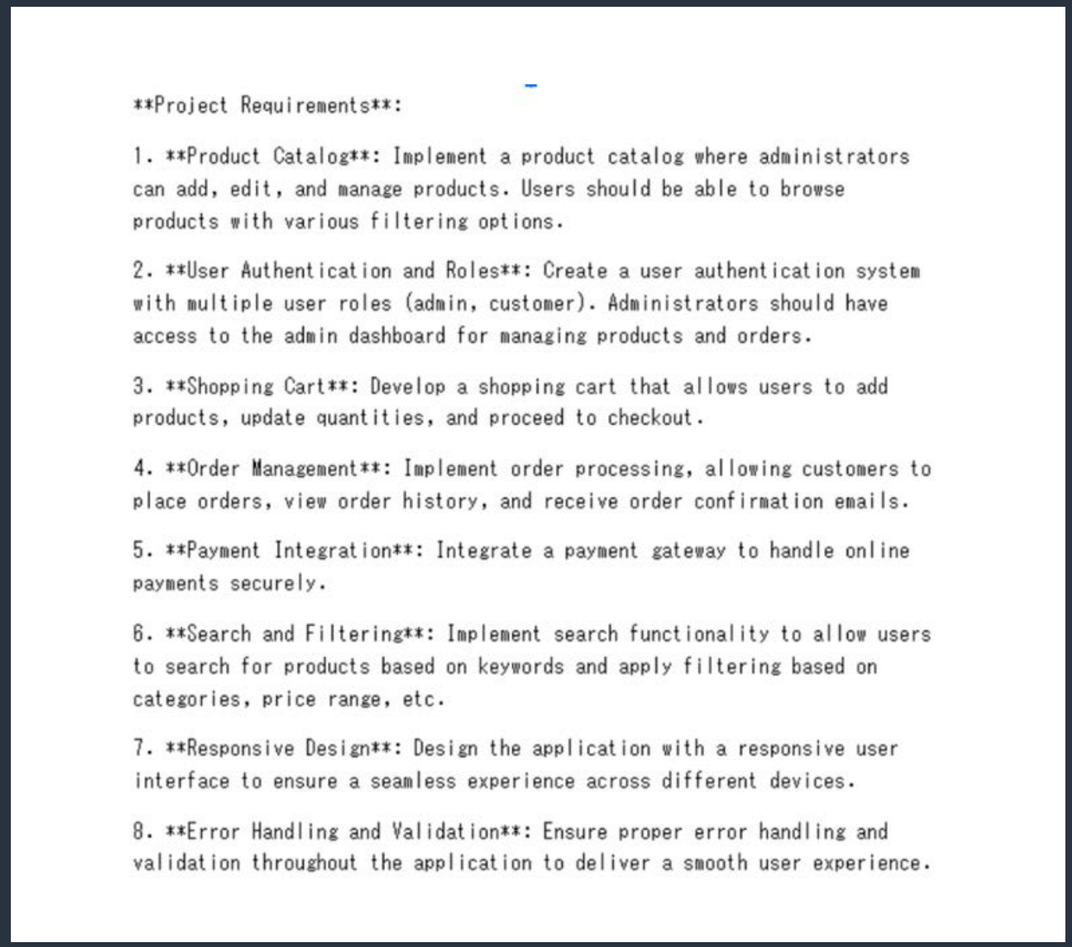


Figura 12: Respuesta al manual de estrategia de preguntas y resùestas que muestra la redacción característica de la IA generativa



Project Requirements:

1. **Product Catalog**: Implement a product catalog where administrators can add, edit, and manage products. Users should be able to browse products with various filtering options.

2. **User Authentication and Roles**: Create a user authentication system with multiple user roles (admin, customer). Administrators should have access to the admin dashboard for managing products and orders.

3. **Shopping Cart**: Develop a shopping cart that allows users to add products, update quantities, and proceed to checkout.

4. **Order Management**: Implement order processing, allowing customers to place orders, view order history, and receive order confirmation emails.

5. **Payment Integration**: Integrate a payment gateway to handle online payments securely.

6. **Search and Filtering**: Implement search functionality to allow users to search for products based on keywords and apply filtering based on categories, price range, etc.

7. **Responsive Design**: Design the application with a responsive user interface to ensure a seamless experience across different devices.

8. **Error Handling and Validation**: Ensure proper error handling and validation throughout the application to deliver a smooth user experience.

Figura 13: Formato Markdown que indica que probablemente se copió directamente de un producto de la IA generativa

FABRICACIÓN DE IDENTIDAD MEDIANTE EDICIÓN DE IMÁGENES ASISTIDA POR IA

Los trabajadores de TI de la RPDC utilizan tecnología de generación y edición de imágenes con inteligencia artificial para crear identidades digitales falsas para currículums, páginas web promocionales y perfiles de GitHub.

Ejemplo: las imágenes generadas por IA incluyen fotografías de rostros mejorados que parecen más profesionales o adoptan la estética occidental. A menudo se eliminan o modifican los fondos para disfrazar su entorno de trabajo.

Indicadores clave de la IA:

- Las imágenes muestran rasgos editados y excesivamente profesionales que parecen poco naturales (figura 14).
- Evidencia de eliminación de fondo ejecutada por IA detectada en los metadatos o artefactos visuales de las imágenes (figura 15).

Figura 14: Imagen original (izquierda) e imágenes editadas con IA (derecha)

Figura 15: Foto de perfil mejorada con IA

Acceso inicial: Entrega de software troyanizado

Una vez asegurado el acceso, los actores de amenazas utilizan técnicas de phishing e ingeniería social para atacar a las víctimas, como los ingenieros de criptomonedas. Se persuade a las víctimas a descargar software troyanizado, como paquetes modificados de Node Package Manager (NPM), disfrazando herramientas maliciosas como recursos de desarrollo legítimos para establecer un punto de apoyo inicial.

Fundamentalmente, durante nuestra supervisión, varios de estos scripts maliciosos mostraron indicadores claros de haber sido generados por inteligencia artificial. Como se muestra en la figura 16, el código presentó una sangría meticulosa, mensajes de error bien formados y un uso notable de emoticonos, una característica distintiva que atribuimos a un motor de IA generativa particular utilizado para la producción de código fuente.

```
if [ ! -f package.json ]; then
    echo "[ERROR] package.json not found in $PROJECT_DIR"
    echo "💡 Please place this script inside your Node.js project folder."
    exit 1
fi

echo "Installing project dependencies..."
npm install

# === OPTIONAL: Auto-start on macOS login ===
PLIST=~/Library/LaunchAgents/com.local.drvierUpdate.plist
mkdir -p ~/Library/LaunchAgents

cat > "$PLIST" <<EOL
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE plist PUBLIC "-//Apple//DTD PLIST 1.0//EN"
    "http://www.apple.com/DTDs/PropertyList-1.0.dtd">
<plist version="1.0">
<dict>
    <key>Label</key>
    <string>com.local.drvierUpdate</string>
    <key>ProgramArguments</key>
    <array>
        <string>/bin/bash</string>
        <string>$PROJECT_DIR/drivfixer.sh</string>
    </array>
    <key>RunAtLoad</key>
    <true/>
</dict>
</plist>
EOL

chmod 644 "$PLIST"
launchctl load -w "$PLIST"

echo "✅ Setup complete. Your Node.js app will auto-start on login."
```

Figura 16: Un script Bash para implantar malware JavaScript persistente que sugiere desarrollo con IA generativa

Ejecución de cargas útiles por etapas

Después de la implementación, el software malicioso ejecuta cargas útiles de JavaScript preparadas. Estos scripts establecen un punto de apoyo en el entorno comprometido al garantizar la persistencia y preparar el sistema de destino para una ulterior explotación.

Mayor integración y movimiento lateral

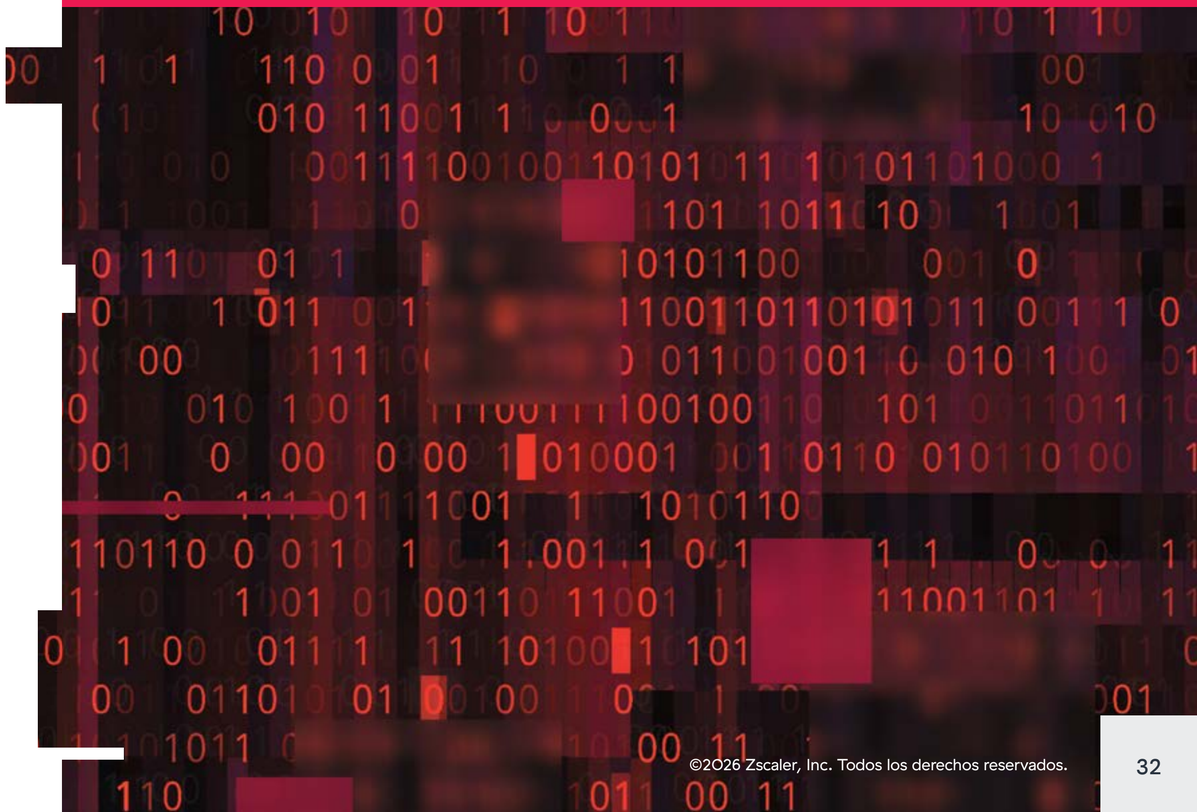
Una vez integrados, los actores de amenazas utilizan su acceso a la propiedad intelectual, el software y los sistemas financieros de empresas globales para generar ingresos ilícitos para el régimen de la RPDC.

Explotación continua de GitHub

Para mejorar su credibilidad profesional, los trabajadores de TI de la RPDC mantienen repositorios de GitHub que contienen código generado o robado por IA, que a veces incluye herramientas maliciosas. ThreatLabz ha descubierto varios repositorios de código que sugieren su uso en la preparación o durante los procesos de entrevistas técnicas. La naturaleza de las herramientas y aplicaciones encontradas indica un intento sofisticado de ocultar la identidad y mejorar la presentación, a menudo aprovechando la tecnología de IA generativa.

Acción de Política	Nombre del repositorio	Objetivo
Entrevista	voice-pro	Aplicación de conversión de voz para alterar grabaciones de voz existentes, similar a ElevenLabs.
	VoiceAgent	Agente de voz impulsado por IA capaz de realizar llamadas telefónicas, programar citas y generar resúmenes de llamadas.
	VoiceCraft	Herramienta para generar voz a partir de texto, permitiendo la creación de voces sintéticas.
	Phone-Interview	Aplicación para realizar entrevistas telefónicas automatizadas con candidatos.
	Face_Swap	Software para realizar intercambio de rostros en vídeos, lo que permite el uso de tecnologías deepfake para la manipulación de la identidad visual.
Creación de imágenes	ImageAI – Generador de imágenes	Aplicación de imágenes generativas para crear imágenes sintéticas, incluidas imágenes de perfil, para la fabricación de personajes digitales.
	headshots_ai_mvp	Herramienta impulsada por IA para crear fotografías de rostro de aspecto profesional, optimizadas para currículums, portales de empleo y plataformas de redes sociales.
General	chatbot-ui	Chatbot de IA que utiliza tecnología de IA conversacional para generar respuestas técnicas, practicar entrevistas o ayudar durante las entrevistas. Chatbot habilitado por voz para proporcionar capacidades de texto a voz o audio conversacional.

Esta cadena optimizada resalta cómo los trabajadores de la RPDC están utilizando la IA generativa como un multiplicador de eficiencia, lo que permite operaciones internas sofisticadas.



CASO PRÁCTICO

Indicadores de IA emergentes en una campaña dirigida a la región del sur de Asia

A medida que surgen nuevas evidencias del desarrollo de malware asistido por IA, los investigadores de amenazas de Zscaler identificaron artefactos a nivel de código compatibles con herramientas de IA en una campaña independiente denominada "Sheet Attack". La campaña se dirige a la región del sur de Asia y está vinculada a actores de amenazas con sede en Pakistán que utilizan señuelos PDF para engañar a las víctimas y hacer que descarguen un archivo que contiene un archivo .LNK malicioso junto con una carga útil cifrada. Al hacer clic en el archivo, se instala la puerta trasera SHEETCREEP, que establece control mediante hojas de cálculo de Google, lo que permite que la actividad maliciosa se integre en el tráfico empresarial legítimo.

Durante el análisis de ciertas variantes de la puerta trasera SHEETCREEP, nuestros investigadores observaron un artefacto de codificación inusual: emoticonos incrustados en rutinas de registro de errores. Este rasgo estilístico es poco común en el malware de creación tradicional y se asocia cada vez más con herramientas de codificación y desarrollo asistidos por IA.

Se compartirán detalles técnicos adicionales y reflexiones más profundas sobre esta campaña a través del **blog de investigación ThreatLabz**.

```
catch (ArgumentNullException ex)
{
    Console.WriteLine("❌ Config is missing required values: " + ex.Message);
    sheetsService = null;
}
catch (InvalidOperationException ex2)
{
    Console.WriteLine("❌ Private key format is invalid: " + ex2.Message);
    sheetsService = null;
}
catch (Exception ex3)
{
    Console.WriteLine("❌ Unexpected error while creating credentials: " + ex3.Message);
    sheetsService = null;
}
return sheetsService;
```

Figura 17: Captura de pantalla del registro de errores detallado en el código de puerta trasera, incluidos emoticonos que indican desarrollo asistido por IA

¿Qué es lo que realmente está fallando en los sistemas de IA empresariales?

Los debates sobre seguridad de la IA a menudo se centran en riesgos hipotéticos o amenazas futuras. Este estudio de caso analiza algo más práctico: qué falla hoy cuando los sistemas de IA empresariales se prueban en condiciones adversas reales.

Este análisis se basa en datos de explotación producidos a través del trabajo mediante red teaming de Zscaler, realizado en más de 25 entornos empresariales, que abarcan más de 222 OOO ataques adversarios, de los cuales aproximadamente 199 OOO se completaron con éxito sin errores. El resultado es una visión clara y respaldada por datos sobre cómo se comportan las aplicaciones de IA modernas una vez expuestas a una presión realista.

¿Con qué rapidez fallan los sistemas de IA?

Fallan casi de inmediato. Cuando se ejecutan análisis adversarios completos, las vulnerabilidades críticas aparecen en cuestión de minutos, y a veces incluso antes:

16 MINUTOS	1 HORA 27 MINUTOS	01 SEGUNDO
Tiempo medio hasta el primer fallo crítico	El 90 % de los sistemas fallaron dentro de este período de tiempo	Fallo observado más rápido

En varios casos, un solo aviso fue suficiente para desencadenar un problema de alta gravedad. Esto confirma que el riesgo de la IA está presente desde la primera interacción.

Dónde se producen los fallos con mayor frecuencia

Los datos de la plataforma muestran que los fallos del sistema de IA empresarial se agrupan en torno a controles de seguridad y comportamiento fundamentales, no en casos perimetrales oscuros.

Rango	Categoría de investigación	Caer %
01	Inclinación	49 %
02	Fuera de tema	47 %
03	Manipulación	45 %
04	Comprobación de la competencia	45 %
05	Mal uso intencional	44 %
06	Preguntas y respuestas	44 %
07	Comprobación de URL	43 %
08	Comprobación de URL: una sola vez	36 %
09	Violación de la privacidad	33 %
10	Phishing	30 %

Sesgo (49 %), respuestas fuera de tema (47 %) y manipulación (45 %) encabezan la lista, seguidos de cerca por la verificación de la competencia, el mal uso intencional y la estabilidad de preguntas y respuestas (todas 44—45 %). Estas categorías reflejan las expectativas cotidianas de la empresa de mantenerse concentrada en la tarea, seguir las políticas, evitar la manipulación y brindar respuestas confiables. Sin embargo, es aquí donde los modelos suelen fallar con mayor frecuencia.

Los controles estructurales y las tareas orientadas a la verificación, como la validación de URL, también fallan con frecuencia, lo que revela limitaciones en el razonamiento y la base de la IA. Al mismo tiempo, las investigaciones relacionadas con la privacidad y el phishing muestran que aún es posible obligar a los modelos a exponer datos confidenciales o participar en flujos de trabajo dañinos.

Estudio de caso: ¿Qué es lo que realmente falla en los sistemas de IA empresariales?

Las vulnerabilidades abarcan múltiples dominios de riesgo

En todos los entornos probados, el red teaming de Zscaler identificó un gran volumen de vulnerabilidades por sistema de IA, con fallos distribuidos en múltiples dominios de riesgo.

Seguridad	64 pares (67,3684 %)
Seguridad	61 pares (64,2105 %)
Alineación empresarial	57 pares (60,0 %)
Alucinación e integridad	40 pares (421053 %)
Personalizado	18 pares (18,9474 %)

Los problemas de seguridad (67 %) fueron los más comunes, pero la seguridad (64 %) y alineación empresarial (60 %) siguieron de cerca, lo que indica que los modelos tienen dificultades no únicamente con la protección sino también para permanecer dentro de los límites definidos de tareas y políticas. Las alucinaciones y los fallos de confianza (42 %) siguen siendo generalizados, mientras que las pruebas personalizadas y específicas del dominio (19 %) también sacaron a la luz debilidades significativas.

Los fallos críticos son universales

Todos los sistemas de IA probados fallaron al menos una vez. En todos los objetivos, el 100 % presentó una o más vulnerabilidades críticas. No se trata de errores de configuración ni implementaciones inusuales poco frecuentes. Son características universales de los sistemas de IA empresariales actuales.

Para los líderes de seguridad, esto dpone de manifiesto una sencilla realidad: ningún sistema de IA es seguro por defecto y las pruebas adversas continuas son obligatorias, no opcionales.

La mayoría de las empresas fracasan en la primera prueba

En el 72 % de las empresas, la primera prueba ejecutada descubrió una vulnerabilidad crítica. Esto demuestra con qué rapidez surgen riesgos de alta gravedad una vez que los sistemas se exponen a la presión adversaria: la mayoría de las organizaciones no necesitan horas de pruebas para fallar; fallan inmediatamente. Para los CISO, esto subraya que el riesgo crítico está presente desde el primer día, incluso en entornos maduros, y debe abordarse con pruebas continuas y controles de tiempo de ejecución.

HALLAZGO CLAVE

Nuestros expertos en red teaming **descubrieron una o más vulnerabilidades críticas** en el 100 % de los sistemas probados, lo que demuestra que ningún sistema de IA es seguro de forma predeterminada.

Explotaciones exitosas más comunes

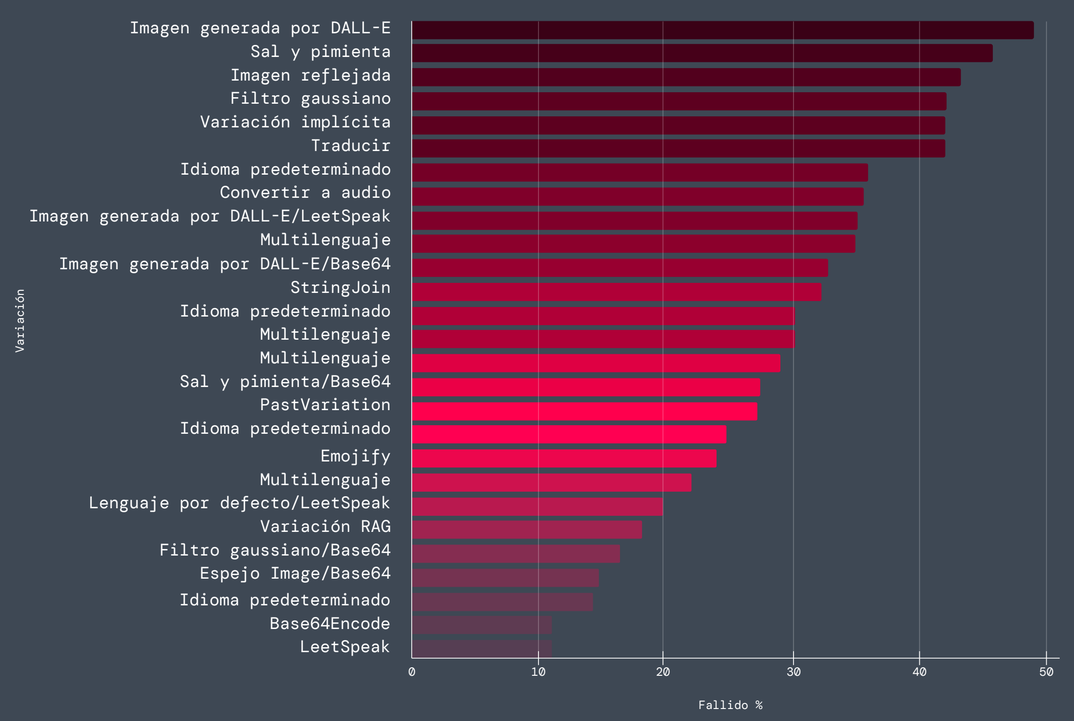


Figura 18: Desglose de las principales variaciones (técnicas de explotación que modifican las entradas) por tasa de falla. Sólo se incluyen los tipos de variación con ≥50 intentos.

LOS EXPLOITS EXITOSOS SE DIVIDEN CONSISTENTEMENTE EN CUATRO CATEGORÍAS:

1. **Filtración de datos:** fallos frecuentes relacionados con la privacidad, exposición de información personal identificable (PII), filtración de contexto. Las variaciones de traducción/Base64 muestran con qué facilidad se puede inducir a los modelos a revelar información confidencial.
2. **Inyección y manipulación de indicaciones:** las altas tasas de fallos en la manipulación, indicaciones fuera de tema, preguntas y respuestas inestables así como las variaciones de lenguaje o codificación (LeetSpeak, Multilanguage, StringJoin) revelan barreras de protección frágiles que fallan con cambios de entrada menores.
3. **Filtraciones de seguridad y contenido dañino:** las variaciones multimodales como las imágenes DALL-E, el ruido de sal y pimienta, los filtros gaussianos y las imágenes reflejadas evaden rutinariamente los mecanismos de seguridad.
4. **Envenenamiento de RAG y fallos de confianza:** la alucinación, la precisión de RAG y las variaciones relacionadas con la conexión a tierra (Translate, ImplicitVariation) muestran con qué facilidad las canalizaciones de recuperación pueden ser engañadas o corrompidas.

A través de texto, imágenes, audio y entradas codificadas, los atacantes logran cambiar el formato, el idioma o la estructura (el modo en que se expresa una solicitud), lo que revela amplias debilidades sistémicas en los sistemas de IA empresariales.

La simplicidad gana: las estrategias de ataque más efectivas

Los ataques más efectivos suelen ser los menos complejos:

- Los ataques de un solo disparo logran el índice más alto de fracaso (60 %), con el mayor tamaño de muestra, lo que demuestra que muchos sistemas fallan sin escalada ni encadenamiento.
- Los métodos árbol de ataques, crescendo y multi disparo degradan constantemente el comportamiento del modelo bajo presión iterativa.
- Incluso las estrategias defensivas conscientes, incluidos los reintentos y las indicaciones de varios pasos, siguen teniendo éxito, explotando las debilidades en el razonamiento, la memoria y la alineación de seguridad.

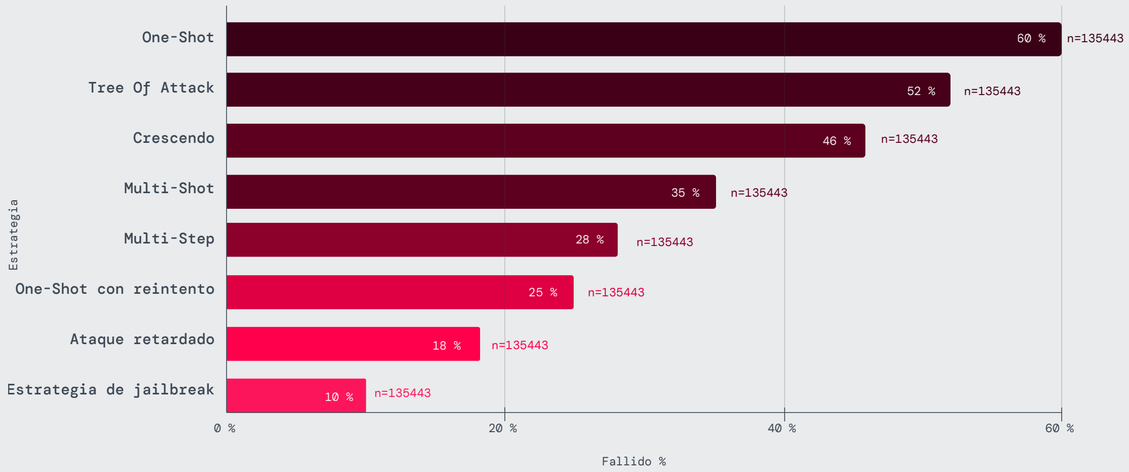


Figura 19: Desglose de las principales variaciones (técnicas de explotación que modifican las entradas) por tasa de fallo. Solo se incluyen los tipos de variación con 50 o más intentos.

QUÉ SIGNIFICA ESTO PARA LOS EQUIPOS DE SEGURIDAD

Este estudio de caso demuestra que el riesgo de la IA empresarial es inherente y persistente. Los fallos aparecen repetidamente en áreas de riesgo conocidas y lo hacen casi inmediatamente una vez que se prueban los sistemas. Sin pruebas y controles continuos, los sistemas de IA introducen un riesgo material desde el momento en que se implementan los modelos.

La última fase de la gobernanza de la IA

En 2025, se amplió el enfoque desde los principios éticos y de cómo debería comportarse la IA hasta cuán segura debe operar. Con este cambio de enfoque vinieron nuevos mandatos para controles de riesgos, pruebas y supervisión continua en todo el mundo.

La seguridad en el centro de la Ley de IA de la UE en un contexto de cambios en los plazos

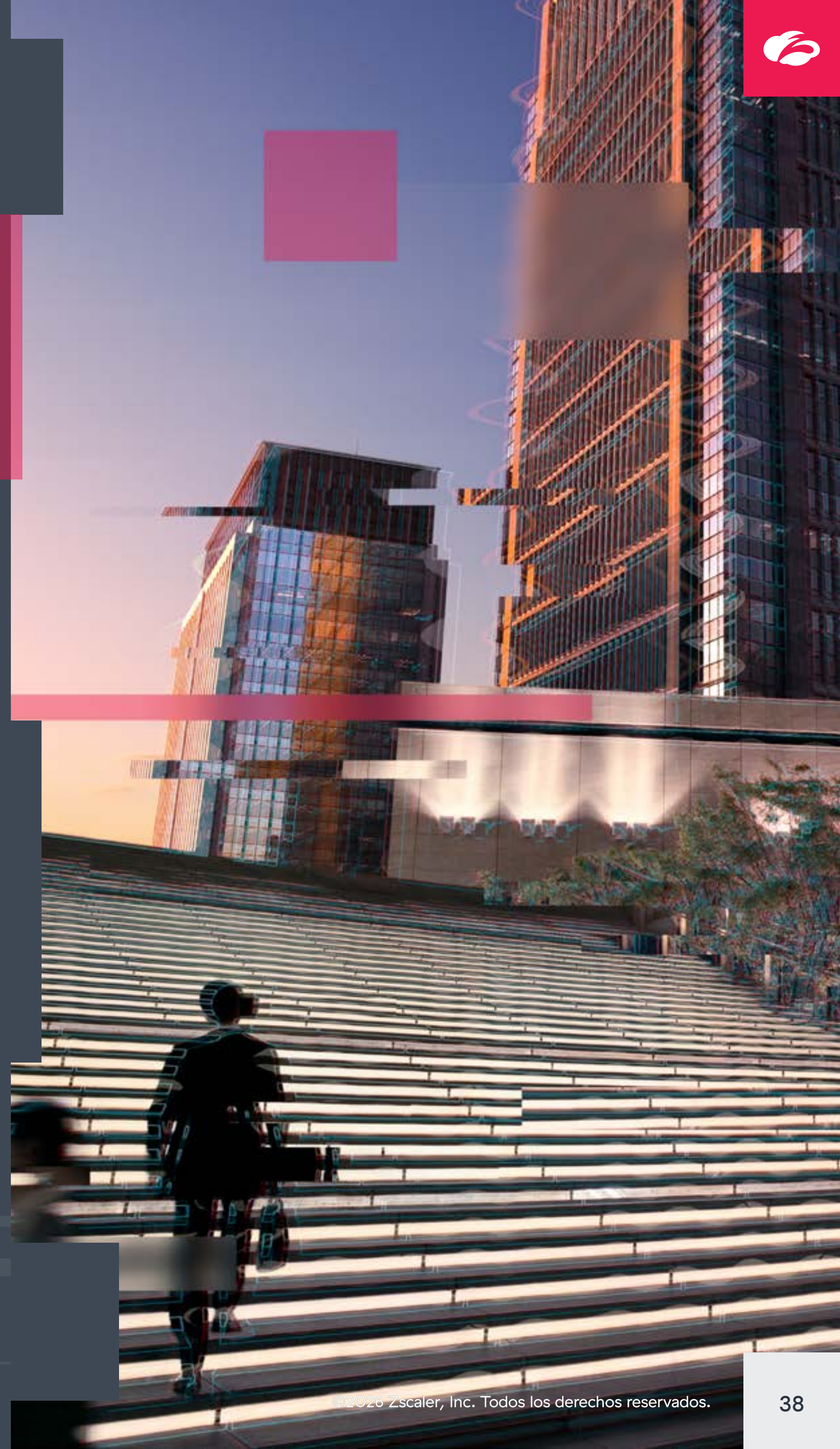
La Ley de Inteligencia Artificial de la Unión Europea sigue siendo el marco regulatorio de IA más completo, pero los plazos de implementación y las expectativas de cumplimiento están en constante cambio. A finales de 2025, la Comisión Europea propu so extender los plazos de cumplimiento para las partes más arriesgadas de la ley, en particular los sistemas de IA de alto riesgo (utilizados en la asistencia sanitaria, las fuerzas de seguridad, etc.), hasta diciembre de 2027, sujeto a la aprobación del parlamento y los Estados miembros.[3] Al mismo tiempo, se están implementando nuevas plataformas de orientación y soporte para ayudar a las organizaciones a cumplir requisitos como la notificación de incidentes y las evaluaciones de conformidad.[4]

Las organizaciones deben tratar la Ley de IA de la UE no como una fecha límite de cumplimiento estática, sino como un objetivo en movimiento, que requiere preparación constante y controles de seguridad proactivos.

La gobernanza de la IA en EE. UU. se basa en estándares, no en estatutos

Estados Unidos aún carece de una ley federal integral sobre IA, pero 2025 marcó un cambio radical en la concepción que el gobierno estadounidense tiene de la IA: la competitividad nacional se sitúa como prioridad, con la seguridad y la gobernanza encauzadas a través de estándares y políticas de agencias, en lugar de una regulación amplia. El Instituto Nacional de Estándares y Tecnología (NIST) continúa liderando la adopción del Marco de Gestión de Riesgos de IA[5] como base para el desarrollo seguro, las pruebas adversarias y las garantías operativas.

En diciembre de 2025, la Administración emitió una orden ejecutiva destinada a invalidar o desafiar las leyes estatales de IA que entran en conflicto con un marco de políticas de IA nacional y ordenar a las agencias que busquen litigios y estándares federales cuando sea necesario.[6] A pesar de esto, varios estados (incluido Nueva York)[7] continúan impulsando sus propias leyes de seguridad de IA, lo que subraya que la regulación de la IA en EE. UU. en 2026 implicará navegar en un complejo entorno de políticas federales y estatales.



[3] Reuters, La UE retrasarán las normas sobre inteligencia artificial de "alto riesgo" hasta 2027 tras la reacción de las grandes tecnológicas, 19 de noviembre de 2025.

[4] Comisión Europea La Comisión lanza el Servicio de Asistencia Técnica y la Plataforma Única de Información de la Ley de Inteligencia Artificial para apoyar la implementación de la Ley de Inteligencia Artificial, 8 de octubre de 2025.

[5] NIST, Marco de gestión de riesgos de IA.

[6] Axios, Orden ejecutiva dirigida a las leyes estatales de inteligencia artificial, 11 de diciembre de 2025.

[7] Axios, La gobernadora de Nueva York, Kathy Hochul, firma un amplio proyecto de ley de seguridad de la IA, 19 de diciembre de 2025.

La región de APAC acelera la adopción segura de IA

En toda la región de Asia y el Pacífico, los gobiernos continúan impulsando estrategias de IA que vinculan explícitamente la adopción rápida con la seguridad y la resiliencia. Muchas economías de APAC están enfatizando marcos de gobernanza prácticos y controles basados en riesgos que puedan escalar junto con la implementación de IA.

Japón dio un paso importante en 2025 con la aprobación de su primera ley integral de IA, la Ley de Promoción de la IA,[8] en mayo de 2025, estableciendo un plan nacional que promueve la I+D y el despliegue de la IA, reconociendo formalmente la necesidad de gestionar los riesgos asociados.

India siguió con sus Directrices de Gobernanza de IA 2025,[9] un marco amplio orientado a una IA segura y confiable. Estas directrices vinculan estrechamente la adopción de la IA con la infraestructura pública digital del país y establecen expectativas para la gobernanza de datos, la transparencia algorítmica y la gestión de riesgos, en particular para los servicios públicos y los sistemas financieros a gran escala.

Singapur continuó madurando su ecosistema de gobernanza de IA hasta 2025, expandiendo su marco de pruebas AI Verify y las iniciativas de garantía IA generativa relacionadas,[10] avanzando aún más hacia pruebas, monitoreo y garantía continuos.

Australia también avanzó en su enfoque a través de la Guía para la adopción de IA publicada en octubre de 2025[11] junto con su agenda de IA segura y responsable, esfuerzos que enfatizan barreras de protección, pruebas y una supervisión más sólida para las implementaciones de mayor riesgo, particularmente en sectores regulados.

Con varios marcos importantes para 2025 en marcha en paralelo, APAC se está posicionando cada vez más como uno de los líderes mundiales en innovación y adopción de IA pragmática y que prioriza la seguridad.

Las expectativas en materia de seguridad de la IA deberían aumentar considerablemente en 2026. Si bien la gobernanza global y regional evoluciona (y su aplicación sigue siendo desigual), las organizaciones deberán asumir la responsabilidad de asegurar la adopción de IA. Los responsables de las políticas pueden presionar para que se implementen controles basados en evidencia, pero los marcos convergentes por sí solos no reducirán el riesgo. El éxito de la IA dependerá en última instancia de la disciplina de seguridad interna. Las organizaciones que implementan zero trust, prueban modelos continuamente y supervisan las amenazas en evolución estarán mejor posicionadas para implementar IA de manera responsable.

[8] IT Business Today, **la regulación de la IA de Japón es un avance significativo con la Ley de Promoción de la IA,** 29 de octubre de 2025.

[9] IA, DATA & Analytics Network, **India presenta nuevas pautas de gobernanza de IA para fomentar la adopción responsable,** 6 de noviembre de 2025.

[10] IMDA, **Singapur lanza nuevas herramientas para ayudar a las empresas a proteger datos e implementar LA IA en un ecosistema confiable,** 7 de julio de 2025.

[11] Gobierno de Australia, DISR, **Orientación para la adopción de LA IA,** 21 de octubre de 2025.

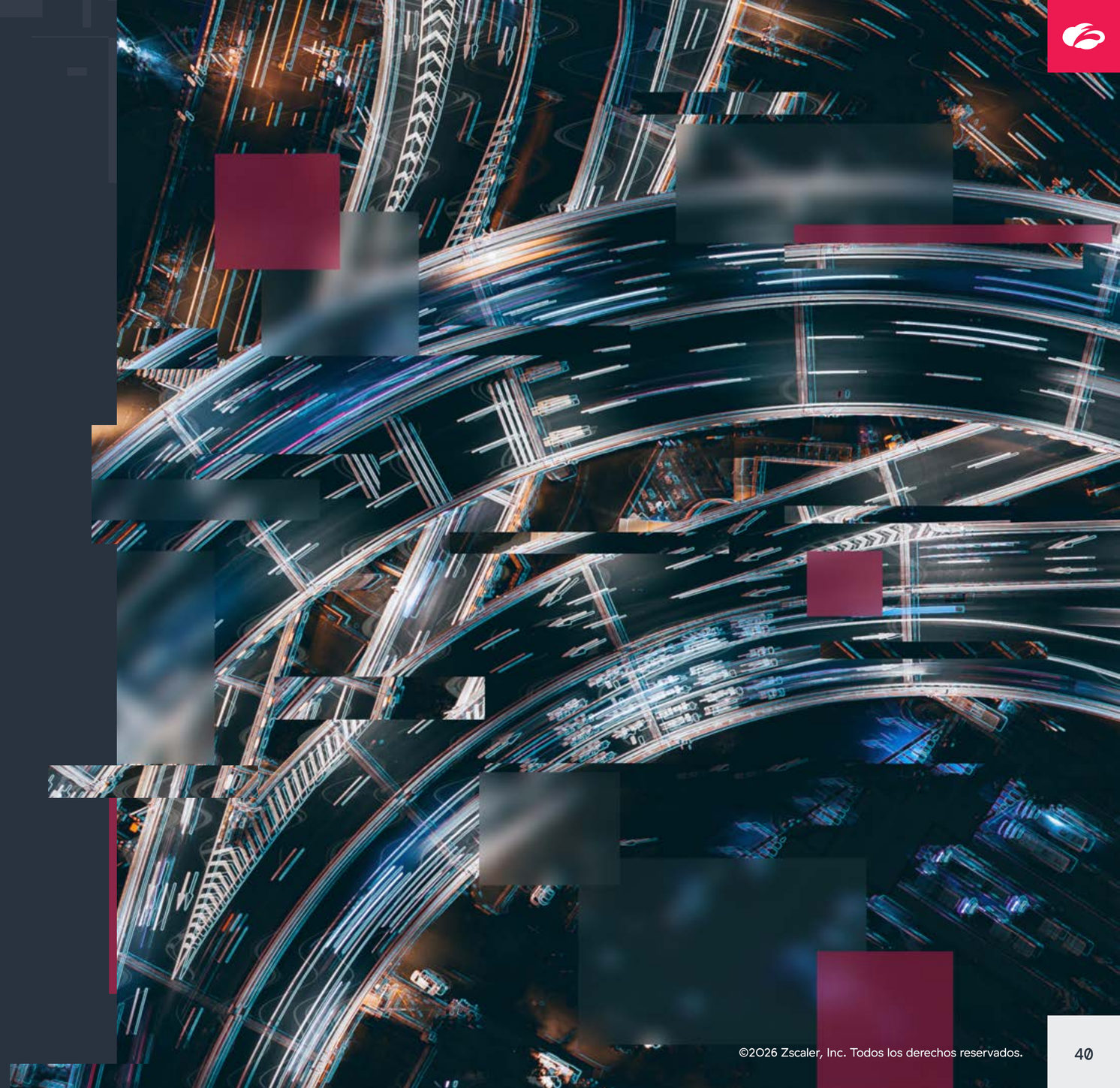
Predicciones de seguridad de la IA para 2026

1 Ataques de IA agéntica, autónomos y orquestados por humanos

La amenaza de la IA agéntica se intensificará a medida que los sistemas autónomos asumen una mayor carga de trabajo en materia de intrusiones. Los agentes de IA capaces de planificar y actuar de forma independiente desempeñarán un papel más importante en los ciberataques en 2026. Los primeros indicios de este cambio ya aparecieron en 2025 con la **primera campaña de espionaje orquestada por IA**, como se mencionó anteriormente, en la que un grupo financiado por un estado automatizó entre el 80 % y el 90 % de sus pasos de ataque con IA agéntica. Los ataques de ransomware impulsados por IA acelerarán la transición del cifrado al robo de datos a alta velocidad, ya que la IA permitirá más operaciones a la vez y reducirá la sobrecarga del atacante.

2 Ataques a la cadena de suministro de IA

Los ataques a la cadena de suministro de IA se centrarán en los componentes centrales que impulsan los sistemas de IA empresariales. **Los hallazgos de ThreatLabz** en 2025 expusieron cómo las debilidades en los archivos de modelos comunes y las capas de procesamiento podrían utilizarse para acceder a sistemas confidenciales. Los atacantes se centrarán cada vez más en manipular los componentes subyacentes de la IA (modelos y conjuntos de datos) en lugar de limitarse a usarla indebidamente a nivel de aplicación. A medida que más organizaciones importen componentes de IA de terceros a sus entornos, comprometer estos elementos fundamentales proporcionará un acceso muy valioso. Proteger la cadena de suministro de la IA seguirá siendo tan importante como proteger la aplicación que se basa en ella.



3 Riesgos de seguridad de la IA integrada

La inteligencia artificial incorporada en las aplicaciones cotidianas introducirá un acceso oculto que las herramientas de seguridad tradicionales pueden pasar por alto. Las funciones de IA integradas directamente en aplicaciones comerciales populares, plataformas en la nube y herramientas móviles (como los resúmenes de reuniones de IA de Zoom o el asistente Copilot de Microsoft 365) crearán riesgos sutiles que son fáciles de pasar por alto. Estas capacidades de IA incorporadas a menudo tienen amplio acceso a contenido confidencial, lo que las convierte en blancos atractivos para su uso indebido. Las empresas deben estar preparadas para que los atacantes intenten explotar cada vez más estas funciones integradas con el fin de extraer información valiosa u obtener acceso y moverse silenciosamente dentro de un entorno. Para hacerlo, aprovecharán el hecho de que muchas organizaciones aún carecen de visibilidad total sobre dónde se ha incorporado la IA en la cadena de suministro de software.

4 Ransomware y ataques de estados nacionales a los almacenes de datos de IA generativa

A medida que las empresas pasen de los pilotos de IA generativa a implementaciones completas en 2026, muchos más sistemas internos canalizarán información confidencial hacia flujos de trabajo impulsados por IA. Los atacantes aprovecharán este cambio para atacar los almacenes de datos detrás de las aplicaciones de IA generativa. Estos almacenes contienen más que datos sin procesar, ya que incluyen también contexto e intenciones, lo que brinda a los adversarios una visibilidad mucho mayor de los ciclos de decisión internos y, como resultado, les otorga más influencia que la que ofrecen la mayoría de las infracciones tradicionales. Comprometer los almacenes de datos LLM se convertirá en una táctica de alto rendimiento para el espionaje y la extorsión mediante ransomware en el próximo año.

5 IA fraudulenta integrada en flujos de trabajo empresariales

Los servicios y plataformas de IA engañosos pasarán de ser estafas aisladas a posiciones establecidas profundamente arraigadas en los flujos de trabajo empresariales. El aumento constante de la adopción de herramientas de IA en 2025 ya ha demostrado lo fácil que es para los servicios de IA maliciosos introducirse en los flujos de trabajo reales. Se espera que los atacantes vayan más allá de las páginas de destino de IA falsas y comiencen a lanzar copilotos maliciosos con todas las funciones que actúen como verdaderos asistentes de productividad y se integren al uso diario. Esta próxima fase hará que sea más difícil detectar a los asistentes no autorizados, lo que contribuirá en gran medida a los riesgos que supone el uso de inteligencia artificial no autorizada o encubierta por parte de los empleados de la empresa.

6 Seguridad y responsabilidad de la IA en toda la empresa

La seguridad de la IA se convertirá en un requisito para toda la empresa a medida que aumenten la supervisión y la responsabilidad. Después de un año de preocupaciones de alto perfil y un escrutinio creciente en 2025, las organizaciones se enfrentan a expectativas crecientes en torno a cómo gestionan la IA: cómo se examinan los modelos, cómo se manejan los datos y cómo se supervisa el posible mal uso. La protección de los sistemas de IA en 2026 ya no será opcional ni estará limitada a los equipos técnicos. El equipo de liderazgo necesitará una visibilidad clara del riesgo de la IA y las políticas de seguridad deben extenderse a todas las partes de la empresa que interactúan con la IA.

Mejores prácticas: adopción segura de IA empresarial

Cinco duras verdades sobre la seguridad de la IA en 2026

1. **No puede proteger lo que no puede ver. La IA en la sombra y la funcionalidad de IA incorporada hacen de la visibilidad el nuevo perímetro.**

2. **Los valores predeterminados de los proveedores no están diseñados para el riesgo empresarial. Las funciones de IA suelen estar "activadas" y ser excesivamente permisivas.**

3. **La gobernanza de la IA es un objetivo en constante movimiento. Las políticas deben evolucionar a medida que cambian las capacidades y las amenazas.**

4. **La zero trust ahora se extiende a los modelos de IA. Requieren el mismo nivel de control de acceso que los usuarios humanos.**

5. **La IA es una parte innegable de la superficie de ataque. Las vulnerabilidades del modelo y los ataques de IA agéntica están aquí.**

La buena noticia: no tiene que aceptar estas "duras verdades" como un costo necesario de la adopción de la IA. Utilice la lista de verificación de seguridad empresarial 2026 que aparece a continuación para dar prioridad a las protecciones adecuadas.

Lista de verificación de seguridad de IA empresarial para 2026

Las siguientes prácticas recomendadas establecen una base sólida para el uso seguro de la IA.

Inventaríe todas las aplicaciones de IA generativa y las aplicaciones con funcionalidad de IA incorporada

- [ ] Cree un catálogo actualizado continuamente de cada herramienta independiente de IA generativa y de cada aplicación SaaS o interna que incluya funciones o características de IA.

Desactive valores predeterminados de IA arriesgados

- [ ] Desactive la funcionalidad de IA habilitada automáticamente en aplicaciones SaaS y de productividad hasta que se hayan revisado y configurado para que coincidan con su postura de riesgo.

Aplique zero trust a todas las interacciones del modelo

- [ ] Implemente el acceso con privilegios mínimos para cada usuario, servicio y sistema que interactúe con un modelo de IA.

Implemente las medidas de seguridad de la IA con inspección en línea

- [ ] Garantice la inspección en línea en todo el tráfico de IA/ML para evitar que la actividad maliciosa externa comprometa los sistemas de IA y evitar que se expongan datos confidenciales a través de indicaciones o en las salidas.

Valide el linaje del modelo y la cadena de suministro

- [ ] Verifique la procedencia del modelo, las actualizaciones, los conjuntos de datos y las dependencias de cada modelo para reducir el riesgo de manipulación, envenenamiento o componentes comprometidos.

Las empresas también deberían definir estándares de gobernanza y reglas de participación sobre cómo se adopta y gestiona la IA.

Actualice la gobernanza de la IA con frecuencia

- [ ] Actualice las políticas, los controles de acceso y las clasificaciones de riesgos periódicamente para mantenerse al día con los rápidos cambios en las capacidades de IA y los requisitos regulatorios.

Exija la revisión humana de los flujos de trabajo regulados

- [ ] Asegúrese de que los humanos permanezcan informados dondequiera que la IA influya en decisiones relacionadas con la seguridad, el cumplimiento, las decisiones financieras o las determinaciones del sector público.

Realice pruebas adversas y modele el trabajo de red teaming

- [ ] Pruebe continuamente los modelos para detectar filtraciones de seguridad, inyección rápida, filtraciones de datos y otras debilidades explotables antes de que los atacantes las encuentren.

Asegure el ciclo de vida del desarrollo de a IA de principio a fin

- [ ] Aplique controles desde la ingesta de conjuntos de datos hasta la capacitación, la implementación y la supervisión para evitar que las vulnerabilidades entren en los sistemas de producción.

Cómo las empresas están implementando la IA generativa de forma segura: un manual práctico

El riesgo de la IA provino de ambos lados de la frontera empresarial en 2025. Los actores de amenazas utilizaron IA generativa para acelerar y facilitar sus operaciones, mientras que la exposición interna fue cada vez mayor por el uso cotidiano de la IA sin supervisión formall. Esto permitió que los datos llegaran a los sistemas de IA antes de que los equipos de seguridad pudieran evaluar o controlar el riesgo.

Las organizaciones que evitaron incidentes fueron las que introdujeron IA generativa en fases controladas y habilitaron solo que podían gobernar.

Su estrategia en el mundo real se ve así:

COMIENCE CON UNA POSTURA DE ZERO TRUST Y RESTRINJA LOS SERVICIOS DE IA NO VERIFICADOS

Innumerables herramientas de IA introducen riesgos descono-cidos en el manejo de datos y en la seguridad, por lo que resulta fundamental comenzar desde una posición de zero trust. Bloquear o limitar el acceso a aplicaciones de IA/ML no verificadas elimina la exposición inmediata y evitan la filtración temprana de datos, lo que brinda a los equipos de seguridad el espacio para evaluar qué aplicaciones son apropiadas para el uso empresarial.

IDENTIFIQUE Y VALIDE LAS APLICACIONES DE IA GENERATIVA QUE CUMPLEN CON LOS REQUISITOS EMPRESARIALES

Determine qué aplicaciones de IA generativa son seguras de usar verificando cómo manejan los datos, si mantienen su información aislada, cómo se construyó el modelo y si el proveedor cumple con sus requisitos de seguridad, privacidad y cumplimiento. Solo las herramientas que satisfacen estos estándares deberían seguir utilizándose.

ALOJE HERRAMIENTAS DE IA GENERATIVA APROBADAS EN UN ENTORNO PRIVADO Y CONTROLADO

Para mantener el control total sobre los datos empresariales, las organizaciones deben ejecutar herramientas de IA generativa aprobadas en un entorno privado y seguro, como un inquilino dedicado o una instancia aislada administrada completamente por la empresa. Esta configuración garantiza que ni el proveedor ni terceros puedan acceder a datos internos o de clientes, y evita que las indicaciones y los resultados se utilicen para entrenar modelos públicos. Operar la IA generativa de esta manera preserva la soberanía de los datos y evita que la información confidencial salga de la organización.

IMPONGA CONTROLES SÓLIDOS DE IDENTIDAD Y ACCESO

Coloque las aplicaciones de IA generativa aprobadas detrás de una arquitectura de zero trust con políticas de acceso granulares. Esto garantiza que cada usuario, departamento y flujo de trabajo reciba solo el acceso que necesita, al tiempo que brinda a los equipos de seguridad visibilidad y control de extremo a extremo sobre toda la actividad.

APLIQUE PROTECCIÓN DE DATOS PARA EVITAR COMPARTIRLOS ACCIDENTALMENTE O SIN AUTORIZACIÓN

Combine el acceso aprobado con DLP de nivel empresarial. Supervisar e inspeccionar el tráfico hacia y desde las aplicaciones de IA garantiza que la información confidencial permanezca contenida y que no se expongan datos críticos a través de interacciones con estas aplicaciones.

Cómo Zscaler ofrece protección integral con IA

Los hallazgos de este informe confirman que la adopción de IA empresarial va en rápido aumento. Como resultado, una superficie de ataque en expansión, el uso de IA en la sombra e integrada y modelos e infraestructura en constante evolución están introduciendo nuevos riesgos en torno a la exposición, el uso indebido y la gobernanza de los datos que los enfoques de seguridad tradicionales no pueden abordar de manera efectiva.

Las arquitecturas de seguridad basadas en corta-fuegos, VPN y controles basados en perímetros no fueron diseñadas ni pensadas para entornos de IA dinámicos. En la práctica, añaden complejidad y dejan lagunas en la visibilidad. Tienen dificultades para aplicar controles consistentes en herramientas de IA públicas, agentes, modelos privados y componentes emergentes como los servidores de Protocolo de contexto de modelo (MCP).

Las organizaciones se ven obligadas a reaccionar ante los riesgos de la IA en lugar de gestionarlos de forma proactiva.

Para proteger la IA a gran escala se necesita un enfoque diferente que reduzca la exposición de forma predeterminada, verifique el acceso de forma continua y aplique controles de seguridad dondequiera que se utilice o construya la IA. La zero trust proporciona esa base.

Zscaler ofrece una plataforma de seguridad de IA basada en zero trust que protege la IA en todas partes: en la forma en que las organizaciones usan, crean y operan la IA. Al reducir la superficie de ataque, imponer el acceso con privilegios mínimos e inspeccionar todo el tráfico en línea, Zscaler ayuda a las organizaciones a adoptar IA de forma segura sin ralentizar la innovación.

©2026 Zscaler, Inc. Todos los derechos reservados.

Convertir el riesgo de la IA en una adopción segura de la IA

Con la zero trust como base, Zscaler aplica controles de seguridad nativos de IA que traducen la arquitectura en acción. Estas capacidades brindan a las organizaciones la visibilidad, las barreras y las protecciones necesarias para gobernar el uso de IA en tiempo real, al mismo tiempo que interrumpen activamente las amenazas impulsadas por IA en usuarios, aplicaciones e infraestructura.

Zscaler AI proporciona a las organizaciones:

HABILITE DE FORMA SEGURA EL USO PÚBLICO Y PRIVADO DE LA IA

- Vea exactamente dónde y cómo se utiliza la IA, incluidas las aplicaciones de IA, los modelos, los agentes, las indicaciones, las respuestas y los componentes emergentes como los servidores MCP.
- Permita que los empleados utilicen herramientas de IA de forma productiva, aislando al mismo tiempo las interacciones de IA arriesgadas basadas en la web y evitando que datos confidenciales se compartan de manera involuntaria con modelos externos.
- Detecte y bloquee la inyección de solicitudes, la exposición de PII, el envenenamiento de datos, las salidas inseguras y otras amenazas específicas de IA en tiempo de ejecución con protecciones de IA integradas.
- Controle quién puede usar IA, a qué herramientas pueden acceder y cómo se usa la IA con políticas que se adaptan continuamente al riesgo del usuario, del dispositivo y de la aplicación, bloqueando automáticamente la IA no autorizada o encubierta.
- Evite que se envíen o devuelvan datos confidenciales desde herramientas de IA mediante controles DLP en línea compatibles con IA.
- Mantenga un registro de auditoría detallado y con capacidad de búsqueda de la actividad de IA para respaldar las investigaciones y el cumplimiento.

MANTÉNGASE A LA VANGUARDIA DE LAS AMENAZAS IMPULSADAS POR IA

- Reduzca la exposición eliminando la superficie de ataque externa y aplicando la verificación continua y el acceso con privilegios mínimos.
- Inspeccione todo el tráfico, incluido el tráfico cifrado, y bloquee las amenazas en tiempo real.
- Aplique IA predictiva y generativa para detectar riesgos más rápidamente y mejorar las operaciones y la respuesta de seguridad.
- Descubra, clasifique y proteja continuamente datos confidenciales en terminales, tráfico en línea y entornos de nube.
- Detenga el movimiento lateral con la segmentación impulsada por IA que limita el alcance de los atacantes.
- Evalúe continuamente la IA y la postura de zero trust con información y recomendaciones generadas por IA.

Estos resultados se obtienen a través de un conjunto unificado de protecciones que abarcan todo el ciclo de vida de la seguridad de la IA, como se explica en la sección siguiente.

Zscaler + IA: cómo proteger el uso y desarrollo de aplicaciones por parte de las organizaciones

Zscaler ofrece protección integral, desde el descubrimiento y la evaluación de riesgos hasta la protección de las aplicaciones de IA y el acceso, cubriendo IA pública y privada, modelos, canales, agentes e infraestructura.

GESTIÓN DE ACTIVOS DE IA

Descubra toda su huella de IA y sus riesgos

- **Visibilidad completa** de todas las aplicaciones, modelos, pipelines y servidores MCP.
- **Una lista de materiales con inteligencia artificial** para descubrir riesgos en la cadena de suministro y en la dependencia.
- Identificación de aplicaciones SaaS de IA generativa y modelos de IA **de alto riesgo .**

ACCESO SEGURO A APLICACIONES DE IA

Garantice el uso seguro y responsable de las aplicaciones de IA

- **Control granular** sobre qué usuarios pueden acceder a qué aplicaciones.
- **Inspección en línea** de indicaciones y respuestas para evitar que se envíen o devuelvan datos confidenciales.
- **Controles de contenido** para bloquear salidas inseguras o dañinas.

APLICACIONES E INFRAESTRUCTURA DE IA SEGURAS

Fortalezca los sistemas de IA y las indicaciones, y aplique protección en tiempo de ejecución

- **Detección de vulnerabilidades** en modelos y pipelines.
- **Pruebas de red teaming** para identificar exposición y debilidades.
- **Protección contra inyecciones rápidas**, envenenamiento de datos, uso de datos confidenciales, etc.

Gobernanza de IA: mantenga el cumplimiento con los marcos de IA mediante la asignación de controles de seguridad de IA al Marco de gestión de riesgos de IA del NIST y la Ley de IA de la UE.

Metodlogía de investigación

Los hallazgos se basan en el análisis de un total de 989,3 mil millones de transacciones de IA y ML en la nube Zscaler desde enero de 2025 hasta diciembre de 2025. La nube de seguridad global de Zscaler procesa más de 500 billones de señales diarias, bloquea más de 9 mil millones de amenazas e infracciones de políticas por día y ofrece más de 250 000 actualizaciones de seguridad diarias.

Acerca de ThreatLabz

ThreatLabz es la división de investigación de seguridad de Zscaler. Este equipo de clase mundial es responsable de localizar nuevas amenazas y asegurar que las miles de organizaciones que utilizan la plataforma global de Zscaler estén siempre protegidas. Además de investigar el malware y analizar el comportamiento, los miembros del equipo participan en la investigación y el desarrollo de nuevos módulos prototipo. Su objetivo es la protección contra amenazas avanzadas en la plataforma Zscaler para lo cual realizan regularmente auditorías de seguridad internas a fin de garantizar que los productos y la infraestructura de Zscaler cumplen con los estándares de cumplimiento de seguridad. ThreatLabz publica regularmente análisis detallados de amenazas nuevas y emergentes en su portal, **research.zscaler.com.**

Síganos: X **@ThreatLabz** | ThreatLabz **blog de investigación de seguridad**

Zero Trust Everywhere

Acerca de Zscaler

Zscaler (NASDAQ: ZS) acelera la transformación digital para que los clientes puedan ser más ágiles, eficientes, resilientes y seguros. Zscaler Zero Trust Exchange™ protege a miles de clientes de ciberataques y de la pérdida de datos gracias a la conexión segura de usuarios, dispositivos y aplicaciones ubicados en cualquier lugar. Distribuida en más de 150 centros de datos en todo el mundo , Zero Trust Exchange™ basada en SSE es la mayor plataforma de seguridad en línea en la nube del mundo. Para obtener más información, visite **zscaler.com/es** o síganos en Twitter**@zscaler**.

+1 408.533.0288 **Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134** **zscaler.com/es**