DELL Technologies

Los 10 principales riesgos de ciberseguridad para la IA generativa y los LLM

Empezar

INTRODUCCIÓN

La inteligencia artificial (IA) está revolucionando la forma en que operan las organizaciones, y tanto la IA generativa como los modelos de lenguaje grandes (LLM) se están convirtiendo en cargas de trabajo fundamentales en los entornos empresariales modernos.

Al igual que cualquier otra carga de trabajo, estas aplicaciones presentan su propio conjunto de complejidades y vulnerabilidades que deben abordarse. A medida que las empresas continúan adoptando la IA para impulsar la innovación, la eficiencia y la ventaja competitiva, garantizar la seguridad de estas aplicaciones se convierte en una necesidad esencial. Las buenas prácticas de ciberseguridad son la base para proteger cualquier carga de trabajo y, del mismo modo que se aplican al resto, es imprescindible mantenerlas también en el uso de la IA. Esto implica aplicar prácticas como la actualización regular de los sistemas, la autenticación multifactor, el control de acceso basado en funciones y la segmentación de la red. Estas medidas son fundamentales, pero la clave está en comprender cómo encajan estas capacidades en la arquitectura y el uso específicos de cada carga de trabajo.

En Dell, contamos con un profundo conocimiento de las cargas de trabajo de IA y de los desafíos de seguridad únicos que afrontan. Gracias a su capacidad para identificar las posibles amenazas contra estas cargas de trabajo, Dell puede ayudarle a crear una estrategia de seguridad sólida. Esta incluye abordar riesgos como la contaminación de los datos de entrenamiento, el robo o la alteración de modelos o la reconstrucción de conjuntos de datos, entre otros.

También nos centramos en gestionar los desafíos asociados a los datos de entrada de su modelo de IA, como prevenir la divulgación de información confidencial, mitigar temas no seguros o sesgos, y garantizar el cumplimiento normativo. En cuanto a los resultados, ayudamos a abordar cuestiones como la dependencia excesiva del modelo y los riesgos relacionados con el cumplimiento.

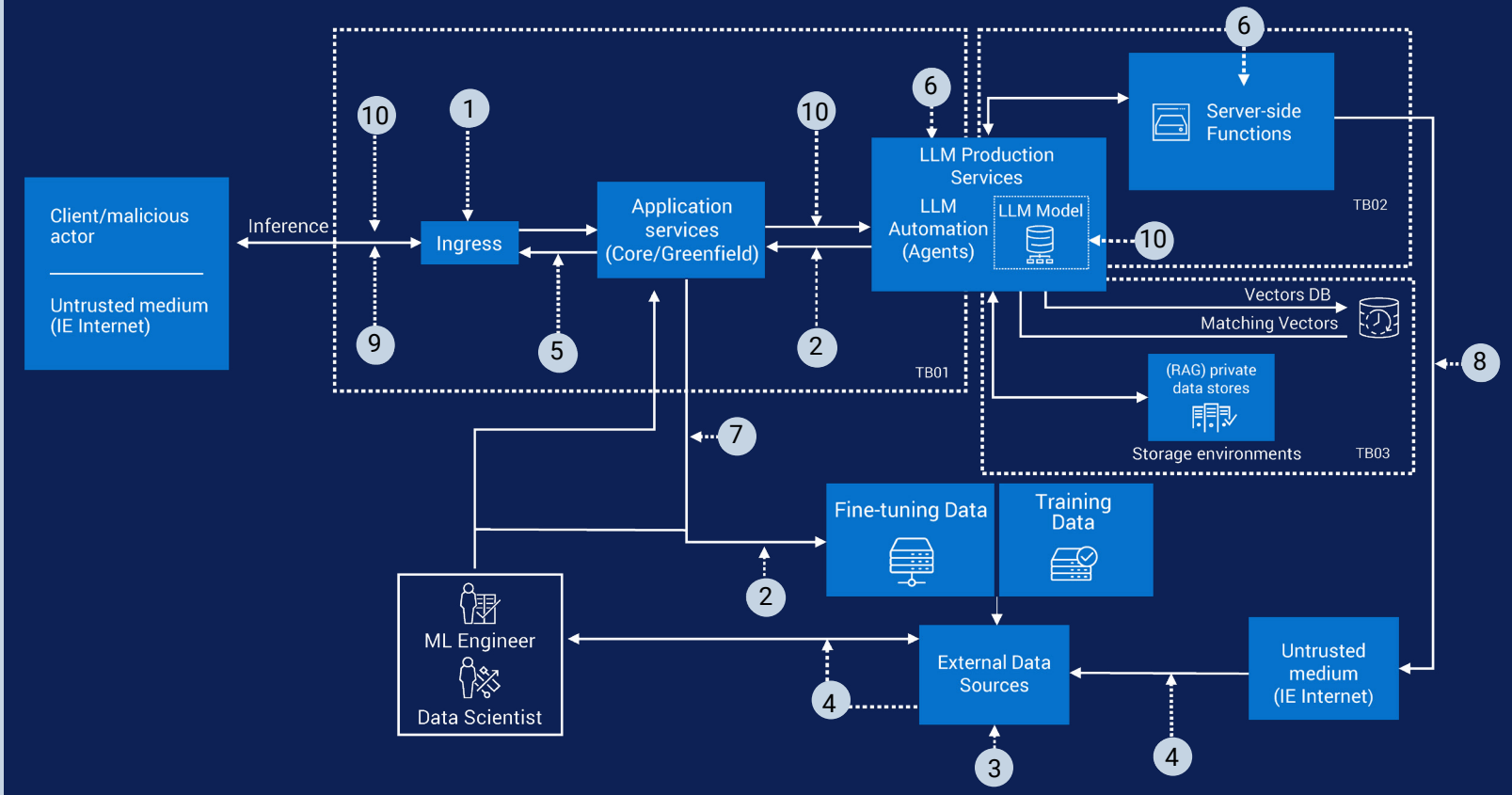
En Dell, ayudamos a las empresas a mitigar estos riesgos aprovechando sus soluciones de ciberseguridad existentes o explorando nuevas herramientas y prácticas para proteger sus sistemas. Nuestro objetivo es garantizar que la seguridad no limite la innovación. Al comprender cómo funcionan las cargas de trabajo de IA y las amenazas a las que se enfrentan, podemos ayudarle a fortalecer su estado de seguridad para que su entorno sea más resiliente y pueda innovar con confianza. Gracias a nuestra experiencia, le ayudamos a aprovechar el potencial de la IA con total confianza y a mantener una seguridad sólida en todo momento.

Los 10 principales riesgos de ciberseguridad para la IA generativa y los LLM

Estos son los principales riesgos que deben tenerse en cuenta para proteger los modelos de IA generativa y LLM, según OWASP.

Haga clic en cada riesgo para obtener más información:

1. Inyección de prompts
2. Divulgación de información confidencial
3. Cadena de suministro
4. Contaminación de datos del modelo
5. Gestión incorrecta de las salidas
6. Capacidad excesiva de actuación
7. Filtración de prompts
8. Debilidades en vectores e integraciones
9. Información errónea
10. Consumo descontrolado

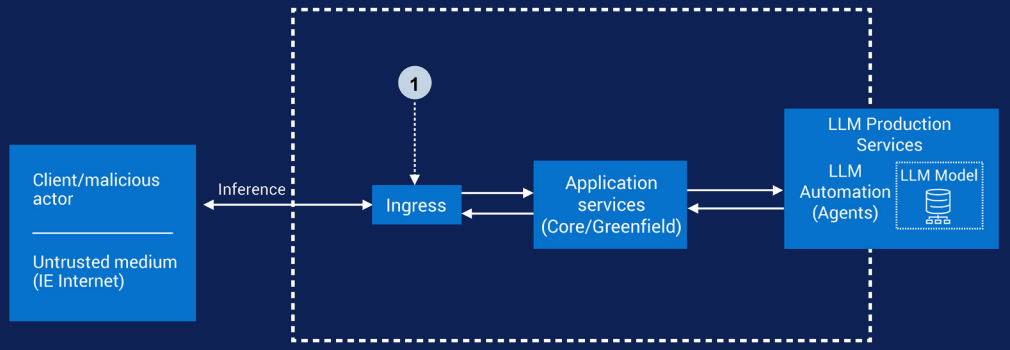


Riesgo n.º 1: Inyección de prompts

Estrategias para mitigar la inyección de prompts:

- **Saneamiento de datos y validación de entradas:** analice minuciosamente las entradas de los usuarios para eliminar contenido malicioso. Utilice técnicas de normalización y codificación para evitar usos indebidos.
- **Enfoques basados en procesamiento del lenguaje natural (PLN) y aprendizaje automático:** emplee PLN y aprendizaje automático para detectar y bloquear prompts manipulados o maliciosos.
- **Formato de salida claro y control de las respuestas:** establezca límites estrictos en las respuestas para garantizar que los resultados sigan los formatos previstos y evitar acciones no autorizadas. Utilice filtrado de prompts y validación de respuestas para mantener la integridad.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Ingeniería de prompts segura:** incorpore el diseño y el análisis seguros de prompts como parte de la estrategia general de seguridad del software para proteger el procesamiento de entradas.
- **Validación del modelo:** valide periódicamente los modelos de aprendizaje automático para asegurarse de que no se hayan alterado antes de su implementación y así preservar su precisión e integridad.
- **Filtrado, clasificación y validación de prompts:** analice y clasifique los prompts para garantizar que solo se procesen entradas seguras. Valide las respuestas para evitar usos indebidos.
- **Comprobaciones de robustez:** realice evaluaciones periódicas para identificar y corregir vulnerabilidades, y mantener la IA segura y fiable.

Volver al diagrama de arquitectura



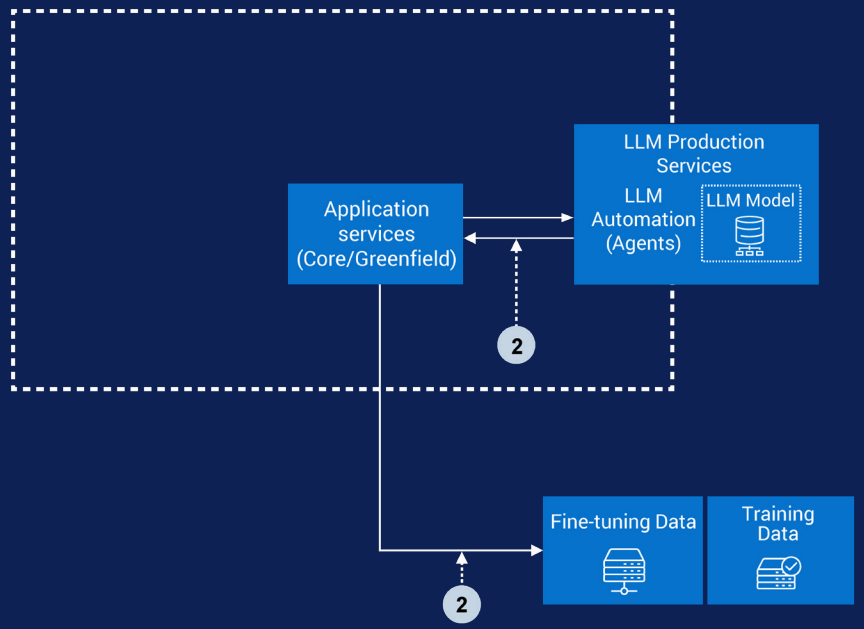
La inyección de prompts es un desafío emergente en el ámbito de la IA generativa, en el que se diseñan entradas maliciosas para manipular el comportamiento del modelo o comprometer su integridad. Estos ataques explotan vulnerabilidades en la forma en que los sistemas de IA procesan y responden a las entradas de los usuarios, lo que puede dar lugar a acciones no autorizadas, desinformación o exposición de datos confidenciales. A medida que la IA generativa se integra en flujos de trabajo empresariales críticos, abordar estos riesgos resulta esencial para mantener la confianza y la seguridad.

Riesgo n.º 2: Divulgación de información confidencial

Estrategias para mitigar la divulgación de información confidencial:

- **Saneamiento de datos y validación de entradas:** analice minuciosamente las entradas de los usuarios para eliminar contenido malicioso. Utilice técnicas de normalización y codificación para evitar usos indebidos.
- **Utilice cifrado homomórfico** para procesar datos confidenciales de forma segura sin exponer su contenido. De esta forma se garantiza que, incluso mientras se utilizan los datos, estos permanezcan cifrados y protegidos frente a brechas de seguridad.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Utilice API e interfaces de sistema seguras** para las interacciones de datos de IA y revise periódicamente las configuraciones para reducir la exposición y la superficie de ataque.
- **Proteja la recopilación, el almacenamiento y las políticas de datos,** y aplique políticas integrales de protección y gobernanza que garanticen el cumplimiento normativo y minimicen el riesgo asociado a los datos.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Desarrollo, configuración y auditorías seguros:** aplique prácticas de programación segura, utilice herramientas automatizadas de gestión de configuraciones y realice revisiones, auditorías y actualizaciones periódicas para mantener las configuraciones del sistema de IA seguras y actualizadas.
- **Formación y concienciación en seguridad:** ofrezca a usuarios y administradores formación continua y específica sobre seguridad para la IA con el fin de reducir el uso inadecuado y la divulgación accidental de datos.

Volver al diagrama de arquitectura



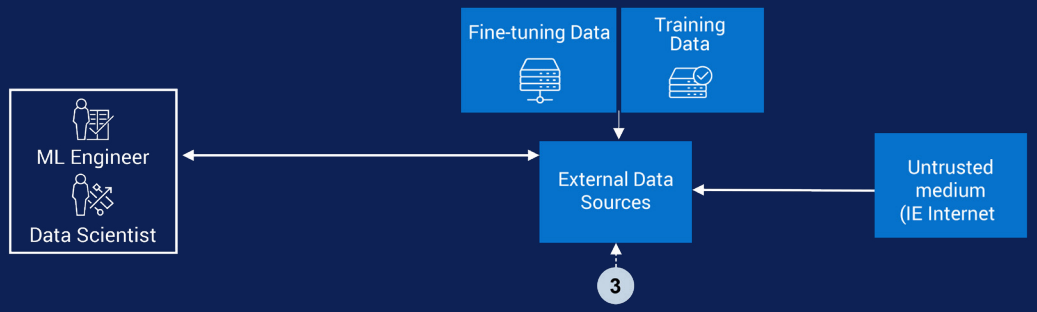
La IA generativa ha supuesto un avance extraordinario, pero también conlleva riesgos importantes, en concreto la exposición no intencionada de información confidencial. Ya se trate de datos personales identificables (PII) o de información empresarial propietaria, el uso indebido o la gestión inadecuada de las herramientas de IA generativa puede provocar filtraciones de datos, incumplimientos normativos o daños reputacionales. Por ello, resulta fundamental que las organizaciones comprendan estos riesgos y los aborden de forma proactiva para garantizar una implementación y un uso seguros de los sistemas de IA.

Riesgo n.º 3: Vulnerabilidades de la cadena de suministro

Estrategias para mitigar las vulnerabilidades en la cadena de suministro:

- **Evaluación de los proveedores para garantizar que la cadena de suministros cumpla la normativa:** evalúe a los proveedores y establezca acuerdos que prioricen la seguridad en la cadena de suministro.
- **Implemente listas de materiales de software:** supervise y verifique el origen de los componentes de software para garantizar la transparencia y reducir el riesgo de código comprometido.
- **Validación del modelo:** valide periódicamente los modelos de aprendizaje automático para asegurarse de que no se hayan alterado antes de su implementación y así preservar su precisión e integridad.
- **Ejecución de contenedores y pods con los privilegios mínimos:** reduzca el impacto potencial en caso de una brecha y limite el acceso no autorizado.
- **Implementación de firewalls:** bloquee las conexiones de red innecesarias para reducir la exposición a posibles amenazas y limitar las vías de ataque.
- **Protección de datos y anotaciones:** proteja sus datos y las anotaciones asociadas para evitar manipulaciones, accesos no autorizados o daños en información crítica.
- **Seguridad del hardware:** utilice hardware validado en materia de seguridad para prevenir vulnerabilidades derivadas de ataques basados en hardware y garantizar una base sólida para su infraestructura.
- **Seguridad de los componentes de software de ML:** emplee componentes de software de aprendizaje automático de confianza y verificados para reducir vulnerabilidades y reforzar la seguridad general de sus flujos de trabajo.
- **Desarrollo, configuración y auditorías seguros:** aplique prácticas de programación segura, utilice herramientas automatizadas de gestión de configuraciones y realice revisiones, auditorías y actualizaciones periódicas para mantener las configuraciones del sistema de IA seguras y actualizadas.

Volver al diagrama de arquitectura



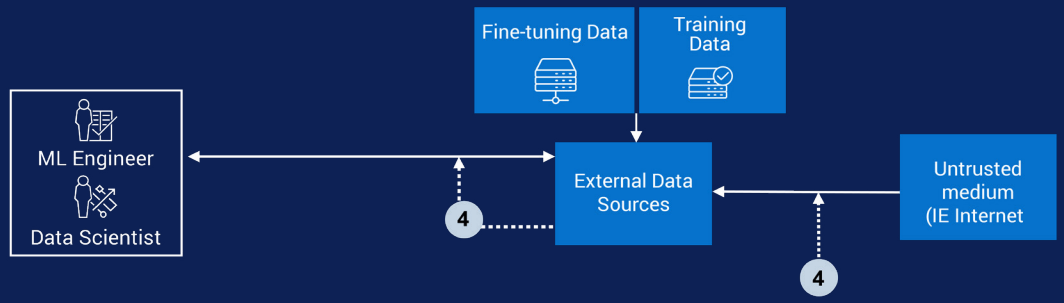
Explore las vulnerabilidades de la cadena de suministro de LLM, que pueden afectar componentes críticos como la integridad de los modelos preentrenados y los adaptadores de terceros. Los sistemas de IA dependen de hardware y software que pueden haberse visto comprometidos mucho antes de su implementación. Los adversarios pueden explotar debilidades en distintas fases de la cadena de suministro del aprendizaje automático, dirigiéndose al hardware de GPU, a los datos y sus anotaciones, a elementos de la pila de software de ML o incluso al propio modelo. Al comprometer estas partes específicas, los atacantes pueden obtener acceso inicial a los sistemas, con riesgos significativos para la seguridad y la integridad. Comprender y mitigar estas vulnerabilidades resulta esencial para construir soluciones de IA robustas y seguras.

Riesgo n.º 4: Contaminación de datos del modelo

Estrategias para mitigar la contaminación de los datos del modelo:

- **Detección de anomalías y validación de datos durante el entrenamiento:** identifique y corrija incoherencias en los datos para garantizar que solo se utilicen datos limpios y de alta calidad para entrenar el modelo.
- **Aísle los entornos durante las fases de ajuste preciso** para evitar accesos no autorizados o la contaminación del modelo en etapas críticas del desarrollo.
- **Validación del modelo:** valide periódicamente los modelos de aprendizaje automático para asegurarse de que no se hayan alterado antes de su implementación y así preservar su precisión e integridad.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Saneamiento de datos y validación de entradas:** analice minuciosamente las entradas de los usuarios para eliminar contenido malicioso. Utilice técnicas de normalización y codificación para evitar usos indebidos.
- **Desarrollo, configuración y auditorías seguros:** aplique prácticas de programación segura, utilice herramientas automatizadas de gestión de configuraciones y realice revisiones, auditorías y actualizaciones periódicas para mantener las configuraciones del sistema de IA seguras y actualizadas.
- **Comprobaciones de robustez:** realice evaluaciones periódicas para identificar y corregir vulnerabilidades, y mantener la IA segura y fiable.
- **Implemente la segmentación de red** para limitar el acceso a interfaces inseguras y a componentes críticos del sistema.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.

Volver al diagrama de arquitectura



La contaminación de datos del modelo es una amenaza de seguridad en el ciclo de vida de la IA en la que los atacantes contaminan deliberadamente los datos de entrenamiento con entradas corruptas, engañosas o maliciosas. Este riesgo puede afectar componentes críticos, desde la recopilación y anotación de datos sin procesar hasta la selección e integración de conjuntos de datos utilizados en el aprendizaje automático o en los modelos de lenguaje grande. La fiabilidad de los sistemas de IA depende de la integridad de sus fuentes de datos, que pueden verse expuestas a manipulaciones antes del entrenamiento, durante el preprocesamiento o a través de pipelines de datos externas.

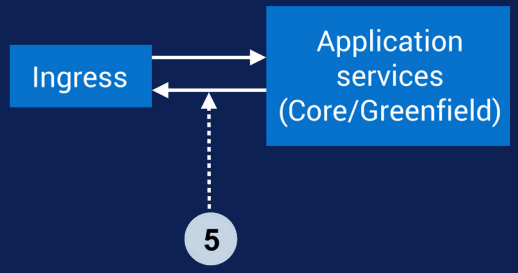
Los atacantes emplean la contaminación de datos para degradar la precisión del modelo, introducir vulnerabilidades o provocar resultados perjudiciales. Al explotar debilidades en el origen de los datos, la calidad de las anotaciones o los procesos de incorporación de conjuntos de datos, los adversarios pueden socavar la seguridad, la fiabilidad y la resiliencia del sistema. Reconocer y mitigar estas amenazas centradas en los datos es esencial para construir soluciones de IA sólidas y de confianza.

Riesgo n.º 5: Gestión incorrecta de las salidas

Estrategias para mitigar la gestión incorrecta de las salidas:

- **Codificación contextual de la salida:** aplique siempre técnicas de codificación y escape adaptadas al contexto específico en el que se utilizará la salida, como entornos HTML, SQL o API, para evitar vulnerabilidades como los ataques por inyección.
- **Saneamiento de la salida:** siga prácticas estrictas de validación y saneamiento de los resultados del modelo, de acuerdo con las directrices del Application Security Verification Standard (ASVS) del Open Web Application Security Project (OWASP), a fin de garantizar un uso seguro posterior y reducir los riesgos de seguridad.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Pruebas automatizadas de seguridad de las salidas:** realice pruebas de seguridad periódicas con herramientas automatizadas para identificar riesgos en los resultados, como vulnerabilidades de scripts entre sitios (XSS) o de inyección, y abórdelos de forma proactiva.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Supervisión humana en el proceso:** en aplicaciones de alto riesgo, como las de los sectores financiero o sanitario, establezca mecanismos de supervisión y revisión humana de los resultados del modelo para garantizar su precisión, seguridad y fiabilidad.
- **Privacidad y cumplimiento normativo:** integre técnicas de preservación de la privacidad en el proceso de generación de resultados y garantice el cumplimiento de las normativas y estándares pertinentes para el uso seguro de la información confidencial.

Volver al diagrama de arquitectura



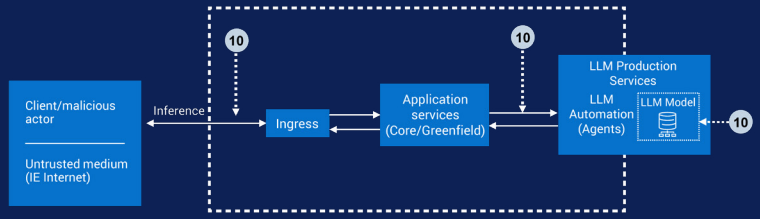
La validación o el saneamiento insuficientes de los resultados del modelo de IA pueden generar graves riesgos de seguridad, como la escalada de privilegios o las vulneraciones de datos. Cuando los modelos de IA producen resultados que no se comprueban ni filtran adecuadamente, los actores maliciosos pueden aprovechar esas vulnerabilidades para obtener acceso no autorizado o ampliar sus privilegios dentro de un sistema. Esta falta de control puede provocar el compromiso de datos, acciones no autorizadas y graves vulneraciones de la seguridad, lo que pone de relieve la importancia de establecer procesos sólidos de validación y saneamiento para cualquier salida generada por la IA.

Riesgo n.º 6: Capacidad excesiva de actuación

Estrategias para mitigar la capacidad excesiva de actuación:

- **Aplique el principio de privilegios mínimos:** conceda a los LLM y a los subsistemas de agentes únicamente los permisos imprescindibles para realizar las operaciones previstas y revise periódicamente los controles de acceso.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Definición de límites operativos:** establezca con claridad qué recursos o acciones pueden ejecutar los LLM o los agentes**.**
- **Supervisión humana en el proceso:** en aplicaciones de alto riesgo, como las de los sectores financiero o sanitario, establezca mecanismos de supervisión y revisión humana de los resultados del modelo para garantizar su precisión, seguridad y fiabilidad.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Limitación de la autonomía:** restringa las capacidades de los LLM para evitar accesos o controles sin restricciones.
- **Desarrollo, configuración y auditorías seguros:** aplique prácticas de programación segura, utilice herramientas automatizadas de gestión de configuraciones y realice revisiones, auditorías y actualizaciones periódicas para mantener las configuraciones del sistema de IA seguras y actualizadas.
- **Implementación de firewalls:** bloquee las conexiones de red innecesarias para reducir la exposición a posibles amenazas y limitar las vías de ataque.
- **Comprobaciones de robustez:** realice evaluaciones periódicas para identificar y corregir vulnerabilidades, y mantener la IA segura y fiable.

Volver al diagrama de arquitectura



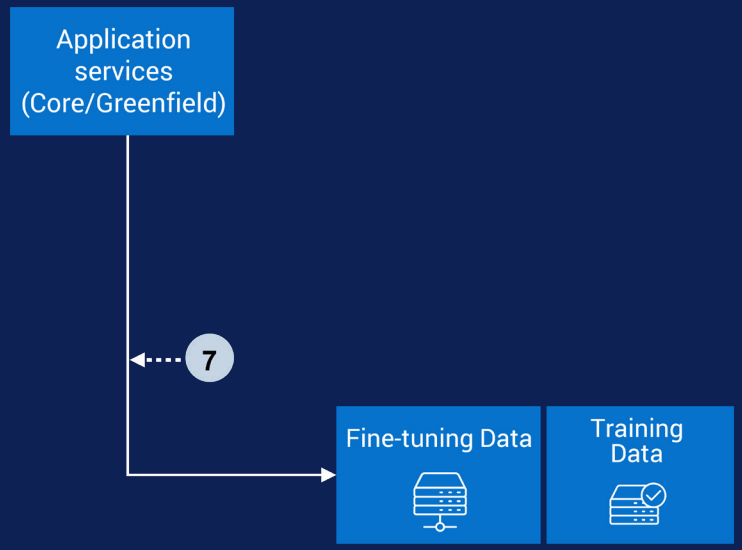
Conceder a los agentes o complementos de IA una autonomía excesiva o funcionalidades innecesarias dentro de los flujos de trabajo puede suponer riesgos importantes. Cuando un sistema de IA recibe privilegios o capacidades más amplias de lo necesario, aumenta la probabilidad de consecuencias no deseadas. Esto puede producirse cuando los sistemas basados en modelos de lenguaje grandes (LLM) se diseñan con permisos excesivos que les permiten realizar acciones o acceder a información que no deberían. Este exceso de control puede llevar a errores, el uso indebido de datos o incluso a vulnerabilidades de seguridad, lo que pone de relieve la importancia de limitar y supervisar cuidadosamente las capacidades de la IA para garantizar un uso seguro y responsable.

Riesgo n.º 7: Filtración de prompts

Estrategias para mitigar la filtración de prompts:

- **Evite incluir información confidencial en los prompts:** no introduzca credenciales, claves de API ni lógica propietaria en los prompts; gestione estos elementos de forma segura fuera del sistema.
- **Separe los controles de seguridad de los prompts:** gestione la autenticación, la autorización y las sesiones dentro de la lógica de la aplicación, no en los prompts.
- **Valide entradas y salidas:** aplique un saneamiento de prompts y respuestas mediante validaciones sólidas que bloqueen patrones o manipulaciones sospechosas.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Cifre y proteja los prompts:** almacene los prompts y las configuraciones en ubicaciones seguras y cifradas para evitar accesos no autorizados.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Revise los prompts periódicamente:** examine y sanee los prompts de forma regular para eliminar datos confidenciales y garantizar el cumplimiento de las normas de seguridad.
- **Pruebas de seguridad y equipos red team:** realice pruebas de tipo adversario para identificar y corregir vulnerabilidades en la gestión de prompts o en las salidas del modelo.
- **Aislamiento de prompts respecto a las entradas de los usuarios:** diseñe los sistemas de forma que las consultas de los usuarios no puedan manipular ni exponer los prompts.
- **Aplicación de límites de uso:** controle el consumo de API, limite la actividad sospechosa y bloquee ataques automatizados basados en prompts.

Volver al diagrama de arquitectura



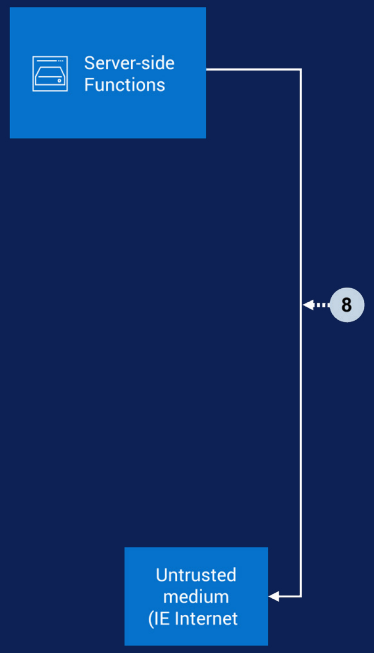
Un ataque de filtración del prompt del sistema en un modelo de lenguaje grande (LLM) o en un sistema de IA se produce cuando un atacante logra extraer o deducir las instrucciones ocultas, los "prompts del sistema", que guían el comportamiento del modelo y definen sus límites operativos. Estas instrucciones no se muestran a los usuarios finales, ya que contienen las reglas principales, las limitaciones y, en ocasiones, la lógica operativa confidencial. Mediante entradas especialmente diseñadas o la explotación de vulnerabilidades, un atacante puede inducir al LLM a revelar su prompt del sistema, total o parcialmente. Si esta información se filtra, puede utilizarse para invertir las restricciones, eludir los filtros de seguridad o desarrollar ataques dirigidos, lo que aumenta el riesgo de inyección de prompts, escalada de privilegios o uso indebido del modelo y de los sistemas dependientes de su integridad.

Riesgo n.º 8: Debilidades en vectores e integraciones

Estrategias para mitigar las debilidades en vectores e integraciones:

- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Cifrado:** proteja los datos vectoriales en tránsito y en reposo mediante estándares de cifrado robustos, como AES.
- **Supervisión y configuración seguras:** refuerce la seguridad de los sistemas, configúrelos de forma segura y supervise continuamente la detección de configuraciones incorrectas, accesos no autorizados o anomalías.
- **Gestión de vulnerabilidades:** actualice y aplique parches con regularidad a todo el software, las dependencias y los motores de almacenamiento vectorial para mitigar los riesgos de seguridad.
- **Saneamiento de datos y validación de entradas:** analice minuciosamente las entradas de los usuarios para eliminar contenido malicioso. Utilice técnicas de normalización y codificación para evitar usos indebidos.
- **Utilice API e interfaces de sistema seguras** para las interacciones de datos de IA y revise periódicamente las configuraciones para reducir la exposición y la superficie de ataque.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Seguridad del hardware:** utilice hardware validado en materia de seguridad para prevenir vulnerabilidades derivadas de ataques basados en hardware y garantizar una base sólida para su infraestructura.
- **Desarrollo, configuración y auditorías seguros:** aplique prácticas de programación segura, utilice herramientas automatizadas de gestión de configuraciones y realice revisiones, auditorías y actualizaciones periódicas para mantener las configuraciones del sistema de IA seguras y actualizadas.

Volver al diagrama de arquitectura



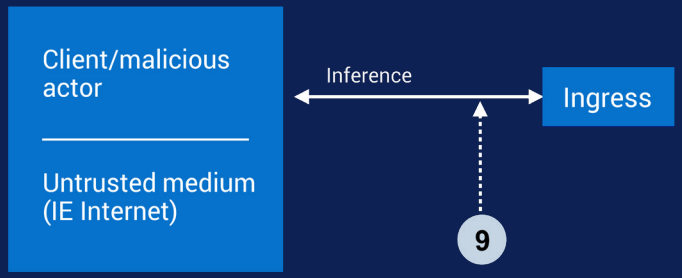
Un ataque que explota debilidades en vectores e integraciones en un modelo de lenguaje grande (LLM) o sistema de IA, especialmente en aquellos que utilizan la generación aumentada de recuperación (RAG), se dirige a las vulnerabilidades existentes en la forma en que la información se codifica, almacena y recupera como vectores numéricos e integraciones. Estas debilidades pueden aprovecharse mediante acciones maliciosas como la inversión de integraciones (reconstrucción de datos confidenciales a partir de integraciones), la contaminación de datos (la inyección de contenido sesgado o dañino para manipular el comportamiento del modelo), el acceso no autorizado a bases de datos vectoriales (con la consiguiente filtración de datos) o la alteración de los resultados de recuperación. Este tipo de ataques pone en riesgo la privacidad, la integridad y la fiabilidad, al permitir que los atacantes revelen información sensible, modifiquen los resultados o socaven la confianza de los usuarios en las aplicaciones basadas en IA. El control de accesos, la validación de datos, el cifrado y la supervisión continua son medidas fundamentales para defenderse de estas amenazas en constante evolución.

Riesgo n.º 9: Información errónea

Estrategias para mitigar la información errónea:

- **Generación aumentada mediante recuperación (RAG) con fuentes autorizadas:** utilice RAG para recuperar e integrar información procedente de bases de datos y repositorios de conocimiento verificados y de confianza, para reducir las alucinaciones del modelo.
- **Ajuste del modelo y calibración de las salidas:** realice ajustes precisos con conjuntos de datos diversos y aplique técnicas que minimicen los sesgos y la desinformación.
- **Verificación automática de hechos:** contraste los resultados con fuentes fiables y marque automáticamente la información falsa.
- **Supervisión de la incertidumbre:** señale las respuestas con bajo nivel de confianza para revisión humana en los casos críticos.
- **Supervisión humana en el proceso:** en aplicaciones de alto riesgo, como las de los sectores financiero o sanitario, establezca mecanismos de supervisión y revisión humana de los resultados del modelo para garantizar su precisión, seguridad y fiabilidad.
- **Comentarios de los usuarios:** permita que los usuarios informen de errores para mejorar continuamente el modelo y corregir con rapidez los flujos de desinformación.
- **Restricciones de acceso y supervisión humana:** aplique control de acceso basado en funciones (RBAC), autenticación multifactor (MFA) y gestión de identidades para limitar el acceso. Incluya la revisión humana en las decisiones críticas.
- **Desarrollo, configuración y auditorías seguros:** aplique prácticas de programación segura, utilice herramientas automatizadas de gestión de configuraciones y realice revisiones, auditorías y actualizaciones periódicas para mantener las configuraciones del sistema de IA seguras y actualizadas.
- **Comunicación de riesgos:** forme a los usuarios sobre las limitaciones de la IA y fomente la verificación independiente de los resultados.
- **Diseño intencional de la interfaz y las API:** destaque el contenido generado por IA y oriente a los usuarios hacia un uso responsable.

Volver al diagrama de arquitectura



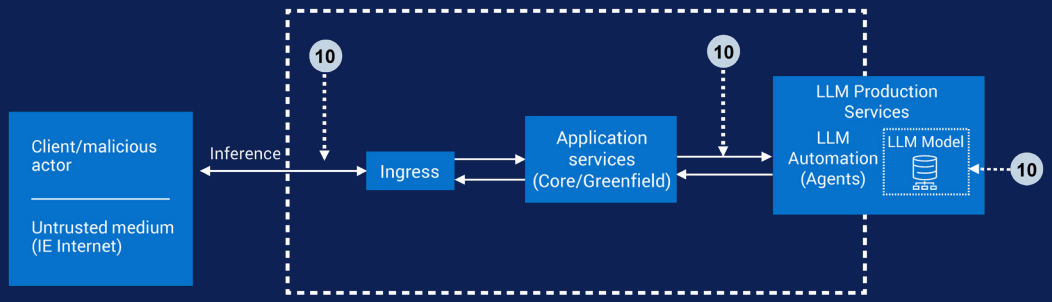
Un ataque de desinformación dirigido a un LLM o a un sistema de IA consiste en inducir al modelo a generar o difundir información falsa, engañosa o aparentemente verosímil, pero incorrecta, a través de sus resultados. Esta vulnerabilidad surge de varios factores: la tendencia del modelo a "alucinar" (generar contenido inventado, pero que parece auténtico), los sesgos o carencias presentes en los datos de entrenamiento y la influencia de prompts maliciosos. Las alucinaciones se producen porque los LLM generan texto de forma estadística, ajustándose a patrones lingüísticos sin una comprensión real de los hechos, lo que da lugar a respuestas que suenan autoritativas, pero carecen de fundamento. Los riesgos de este tipo de ataques incluyen vulneraciones de la seguridad, daños reputacionales e incluso responsabilidad legal, especialmente en entornos donde los usuarios confían excesivamente en las respuestas del LLM sin verificar su exactitud o validez, lo que puede dar lugar a la incorporación de errores o desinformación en decisiones y procesos críticos.

Riesgo n.º 10: Consumo descontrolado

Estrategias para un consumo descontrolado:

- **Aplicación de límites de uso y cuotas por usuario:** establezca límites estrictos de solicitudes, tokens o volumen de datos por usuario, clave de API o aplicación para evitar abusos.
- **Autenticación y segmentación de usuarios:** implemente autenticación sólida (por ejemplo, claves de API u OAuth) y asigne roles o niveles de acceso que permitan procesar solo las solicitudes autorizadas.
- **Validación de entradas y restricciones de tamaño:** compruebe el tamaño y la estructura de los prompts, bloqueando o recortando las consultas excesivamente grandes o con formato incorrecto.
- **Límites de tiempo de procesamiento y control de recursos:** defina tiempos máximos de ejecución y límites de recursos por solicitud para evitar operaciones prolongadas y sobrecarga del sistema.
- **Uso de almacenamiento en caché inteligente y eliminación de duplicados:** guarde en caché las respuestas a consultas duplicadas o similares para reducir el procesamiento innecesario.
- **Supervisión, registro y detección de anomalías:** supervise y registre de forma continua las actividades del sistema de IA mediante soluciones como MDR, XDR o SIEM, con el fin de detectar, investigar y responder rápidamente ante accesos no autorizados, anomalías o filtraciones de datos.
- **Supervisión de presupuesto y control de gastos:** utilice paneles de control y alertas para supervisar los costes y bloquear el uso al alcanzar los umbrales presupuestarios.
- **Aislamiento y ejecución en entornos controlados:** ejecute las cargas de trabajo en entornos aislados y con permisos limitados para reducir los riesgos.
- **Limitación de la profundidad de llamadas y del número de turnos de conversación:** imponga límites a las llamadas recursivas o a los pasos de interacción para evitar la explotación del sistema.
- **Asignación escalonada de modelos o recursos:** dirija las solicitudes de alta prioridad a modelos de nivel superior y el tráfico de menor prioridad a opciones más económicas.

Volver al diagrama de arquitectura



Una amenaza de consumo descontrolado en un LLM o sistema de IA hace referencia a una vulnerabilidad de seguridad que se produce cuando la aplicación permite a los usuarios, maliciosos o no, enviar solicitudes o prompts de inferencia en exceso y sin control, sin aplicar límites de uso, autenticación ni restricciones eficaces. Dado que la inferencia en LLM requiere una gran capacidad computacional, la falta de control puede explotarse de varias formas: los atacantes pueden provocar una denegación de servicio (DoS) saturando los recursos del sistema, generar pérdidas económicas en entornos de pago por uso o en despliegues en la cloud, o incluso consultar el modelo de forma sistemática para clonar su comportamiento y robar propiedad intelectual. Las consecuencias incluyen la interrupción del servicio, la degradación del rendimiento para otros usuarios, el impacto financiero y un mayor riesgo de filtración de información sensible del modelo. En definitiva, el consumo descontrolado se produce cuando el uso de los recursos no está debidamente gestionado, dejando a las aplicaciones basadas en LLM expuestas tanto a la explotación accidental como a la intencionada.

Por qué elegir Dell para proteger la IA

Dell ayuda a las organizaciones a proteger sus modelos de IA y LLM mediante un enfoque integral que abarca hardware, software y servicios gestionados. La seguridad se incorpora desde la cadena de suministro hasta el dispositivo, la infraestructura, los datos y las aplicaciones, todo ello alineado con los principios de confianza cero. En toda la cartera, las soluciones de Dell están diseñadas para reforzar las buenas prácticas de ciberseguridad, con funciones como MFA, RBAC, privilegios mínimos y verificación continua. Este enfoque integral, basado en el principio de "seguridad por diseño", permite a las organizaciones innovar con confianza en el uso de la IA y los LLM, minimizando los riesgos de robo de modelos, filtraciones de datos, ataques adversarios y otras ciberamenazas avanzadas.

Cadena de suministros

La cadena de suministro segura de Dell proporciona una protección fundamental para los modelos de IA y los LLM al integrar la seguridad en todas las fases del desarrollo, la fabricación y la entrega de productos. Mediante actualizaciones del BIOS y firmware firmadas criptográficamente, verificación de componentes seguros, listas de materiales de software centradas en IA (SBOM), trazabilidad de los conjuntos de datos, software y configuraciones de seguridad integrados, así como exhaustivas evaluaciones de riesgo de proveedores alineadas con los estándares internacionales, Dell minimiza los riesgos de manipulación, accesos no autorizados y ataques a la cadena de suministro. De este modo, las organizaciones pueden desplegar cargas de trabajo de IA fiables y resilientes, con plena transparencia, integridad y cumplimiento normativo.

PC con IA

Dell ofrece una seguridad fundamental para las cargas de trabajo de IA ejecutadas en el dispositivo. Los Dell Trusted Devices, que son los PC comerciales con IA más seguros del mundo*, están diseñados con la seguridad como prioridad. La seguridad en la cadena de suministro reduce el riesgo de vulnerabilidades y manipulaciones en los productos. Las defensas exclusivas integradas directamente en el hardware y el firmware protegen tanto el PC como al usuario final durante su uso. Dell SafeBIOS ofrece una visibilidad profunda a nivel de BIOS y detección de manipulaciones, mientras que Dell SafeID refuerza la seguridad de las credenciales y permite la autenticación sin contraseña. El software de nuestros partners proporciona una protección avanzada en los entornos de punto final, red y cloud.

Ciberresiliencia

Las soluciones de ciberresiliencia Dell PowerProtect protegen los datos de IA mediante copias de seguridad cifradas e inmutables, restauraciones rápidas y vaults de ciberrecuperación aislados. Estas capacidades impiden la destrucción de datos, reducen el impacto de actualizaciones maliciosas y facilitan el cumplimiento normativo y la recuperación tras un ataque.

Servidores

Los servidores PowerEdge incorporan informática confidencial para aislar y proteger los prompts y las integraciones de IA/LLM, junto con soluciones de generación aumentada de recuperación (RAG) basadas en fuentes autorizadas. Además, integran MFA, RBAC, raíz de confianza de silicio, firmware firmado y supervisión continua para proteger las cargas de trabajo de IA más críticas.

Para almacenamiento

La cartera de soluciones de almacenamiento de Dell garantiza un almacenamiento seguro y cifrado para los datos sensibles de IA, con un cifrado robusto AES-256 tanto en reposo como en tránsito. Algunas soluciones incorporan cifrado avanzado diseñado para resistir futuras amenazas cuánticas. La cartera incluye rendimiento NVMe de alta velocidad, módulos de cifrado conformes con FIPS para proteger los datos, incluidos los utilizados en cargas de trabajo de IA, instantáneas inmutables y vaults de ciberrecuperación con aislamiento físico para contrarrestar los ataques de ransomware. La arquitectura de confianza cero, la seguridad en la cadena de suministro y las capacidades de auditoría inalterable refuerzan la gobernanza. Los modelos de detección de anomalías y de AIOps ML integrados protegen las cargas de trabajo sin utilizar los datos del cliente para su entrenamiento, lo que reduce el riesgo de ataques basados en las entradas.

AIOps

Dell AIOps ofrece supervisión automatizada y continua para detectar configuraciones incorrectas y vulnerabilidades, incluidas las CVE, además de proporcionar visibilidad sobre los riesgos de la cadena de suministro que afectan a las cargas de trabajo de IA y LLM. El análisis en tiempo real de CVE, las alertas inteligentes y los paneles de control impulsados por IA facilitan una intervención rápida al señalar anomalías y realizar un seguimiento de los flujos de resolución. Las funciones integradas de cumplimiento normativo, el control de acceso basado en funciones y la generación automática de informes ayudan a mantener operaciones seguras en todas las cargas de trabajo. Asimismo, la integración fluida con EDR/XDR y las perspectivas operativas basadas en IA, incluidas las capacidades generativas en las soluciones compatibles, mejoran la eficiencia del entorno de TI.

Redes

Las soluciones de Dell Networking protegen los entornos de IA y LLM mediante una segmentación de red robusta que reduce al mínimo el movimiento lateral. Los canales de red cifrados y los controles de firewall integrados bloquean el acceso no autorizado a los datos de IA.

Servicios de resiliencia y seguridad de IA

Los servicios de seguridad y resiliencia de IA de Dell están diseñados para abordar los nuevos riesgos asociados a la integración de la IA en su organización. Nuestros servicios, diseñados para empezar a trabajar junto a sus equipos a medida que incorpora la IA lo antes posible, proporcionan conocimientos especializados para guiar en la planificación estratégica, la implementación de soluciones y los servicios gestionados de seguridad para aliviar las cargas operativas y poder innovar de forma segura con la IA. Cada uno de ellos está diseñado para ayudar a las organizaciones a abordar los riesgos de la IA en constante evolución y optimizar las implementaciones de IA seguras.

Dell AI Factory

Una cartera integrada de soluciones de seguridad específicas que incluye la cadena de suministro segura de Dell, capacidades de confianza cero para aplicar el principio de privilegios mínimos y soluciones AI MDR, diseñadas para mantener sus modelos protegidos y seguros.

* Según análisis internos de Dell en octubre de 2024 (Intel) y marzo de 2025 (AMD). Aplicable a PC con procesadores Intel y AMD. No todas las funciones están disponibles en todos los PC. Algunas funciones requieren compras adicionales. PC con tecnología Intel validados por Principled Technologies, julio de 2025.

Conclusión

Para construir marcos de IA resilientes, es esencial adoptar un enfoque colaborativo entre las organizaciones y los expertos en seguridad. A medida que la IA y los LLM continúan transformando los distintos sectores, resulta fundamental abordar los riesgos asociados, como la seguridad de los datos, la integridad de los modelos y los desafíos de cumplimiento normativo. Las organizaciones deben dar prioridad a estrategias proactivas que integren la seguridad en todas las fases de su recorrido hacia la IA.

Dell Technologies se consolida como un socio de confianza en esta misión, al ofrecer personalización integral de la IA generativa, asesoría sobre seguridad y soluciones integradas adaptadas a las necesidades específicas de cada cliente. Al aprovechar las sólidas soluciones de ciberseguridad de Dell, las empresas pueden mitigar eficazmente los riesgos asociados a la IA y los LLM, al tiempo que maximizan el valor de sus inversiones de seguridad existentes. Dell permite a las organizaciones proteger su infraestructura de IA mediante la integración fluida de tecnologías de seguridad avanzadas en sus marcos actuales, garantizando así un entorno seguro y preparado para el futuro.

Descubra cómo las soluciones integrales de IA de Dell pueden proteger sus entornos de IA generativa y LLM: Dell.com/CyberSecurityMonth