DELL Technologies

I 10 principali problemi di sicurezza informatica per la GenAI e gli LLM

Guida introduttiva

Introduzione

L'intelligenza artificiale (AI) sta rivoluzionando il modo di operare delle organizzazioni, con AI generativa (GenAI) e modelli linguistici di grandi dimensioni (LLM) che diventano carichi di lavoro critici negli ambienti aziendali moderni.

Come qualsiasi altro carico di lavoro, queste applicazioni implicano una serie di complessità e vulnerabilità da affrontare. Man mano che le aziende continuano ad adottare l'AI per promuovere l'innovazione, l'efficienza e il vantaggio competitivo, garantire la sicurezza di queste applicazioni diventa una necessità fondamentale. Una buona igiene informatica è alla base della sicurezza di qualsiasi carico di lavoro e, proprio così come si dà priorità alla sicurezza in tutti i carichi di lavoro, è essenziale adottare una buona igiene informatica anche per l'AI. Ciò significa implementare prassi come l'applicazione di patch di sistema appropriate, l'autenticazione a più fattori, l'accesso basato sui ruoli e la segmentazione della rete. Queste misure sono fondamentali, ma la chiave sta nel comprendere in che modo queste funzionalità si integrano nell'architettura e nell'utilizzo specifici del tuo carico di lavoro.

In Dell, abbiamo una profonda conoscenza dei carichi di lavoro di AI e delle particolari sfide che comportano nel campo della sicurezza. Attraverso l'identificazione dei modi in cui gli attori delle minacce potrebbero prendere di mira questi carichi di lavoro, Dell ti aiuta a creare una solida strategia di sicurezza. Ciò include la gestione di rischi come l'inquinamento dei dati di addestramento, il furto o la manipolazione dei modelli, la ricostruzione dei dataset e altro ancora.

Ci concentriamo inoltre sulla gestione delle sfide associate all'input del tuo modello di AI, come la prevenzione della divulgazione di informazioni sensibili, la mitigazione di argomenti non sicuri o pregiudizi e la garanzia di conformità alle normative. Per quanto riguarda l'output, aiutiamo ad affrontare problemi come l'eccessiva dipendenza dal modello e i rischi correlati alla conformità.

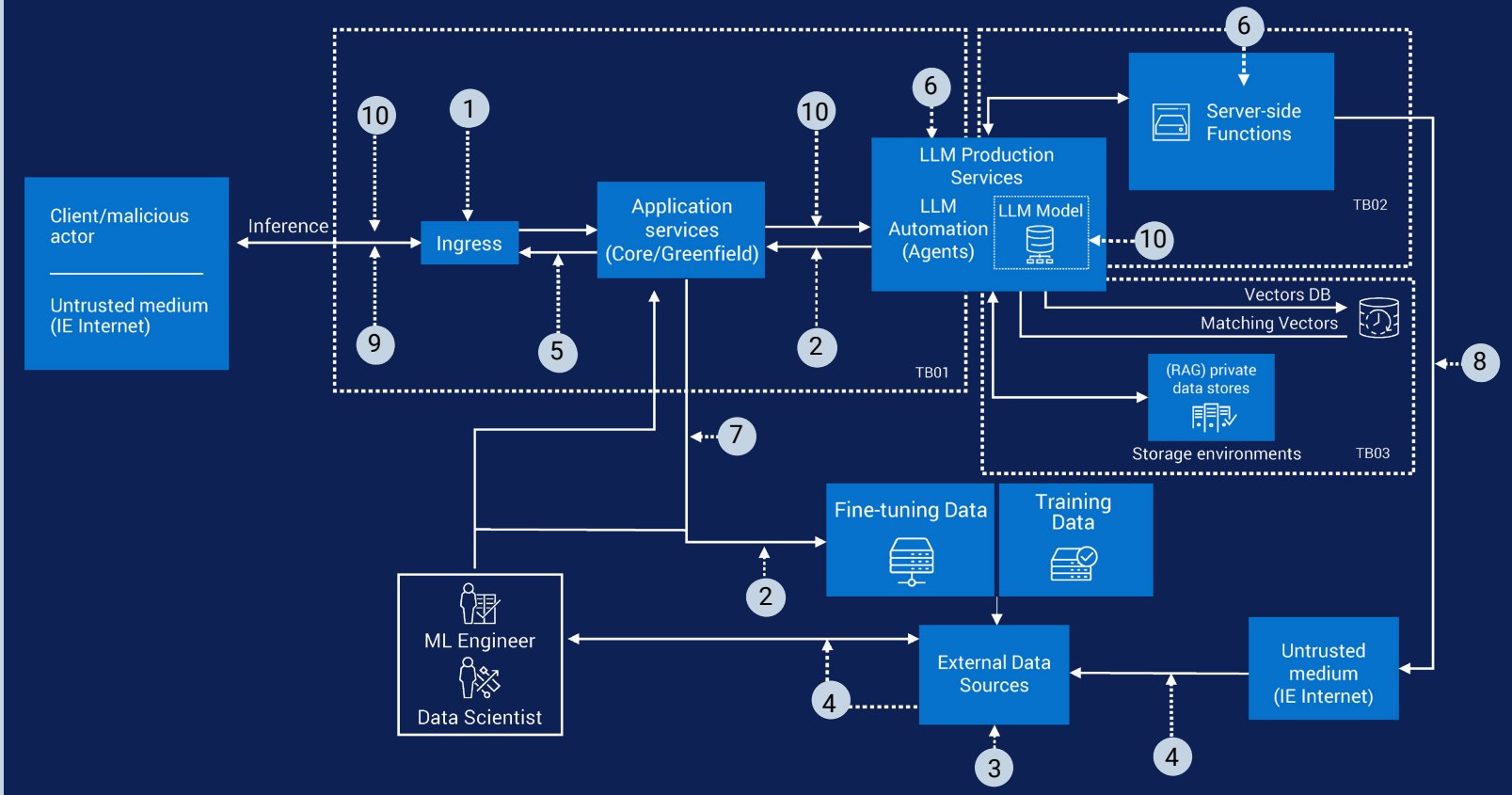
Con Dell, le aziende hanno la possibilità di mitigare questi rischi sfruttando le soluzioni di sicurezza informatica esistenti o esplorando nuovi strumenti e pratiche per proteggere i propri sistemi. Il nostro obiettivo è garantire che la sicurezza non ostacoli l'innovazione. Conoscendo il funzionamento dei carichi di lavoro di AI e le minacce alla sicurezza a cui sono esposti, ti forniamo assistenza per creare un profilo di sicurezza più solido, in modo da rendere più resiliente il tuo ambiente e al contempo consentirti di innovare in tutta sicurezza. Grazie alla nostra competenza, puoi sfruttare con fiducia il potenziale dell'AI, mantenendo parallelamente una sicurezza solida.

I 10 principali problemi di sicurezza informatica per la GenAI e gli LLM

Di seguito sono riportati i principali problemi per la protezione dei modelli di GenAI e LLM, come indicato da OWASP.

Clicca su ognuno di essi per saperne di più:

1. Prompt injection
2. Divulgazione di informazioni sensibili
3. Supply chain
4. Inquinamento dei dati del modello
5. Gestione impropria dell'output
6. Excessive agency
7. System prompt leakage
8. Punti deboli nei vettori e negli embedding
9. Informazioni errate
10. Consumo non vincolato

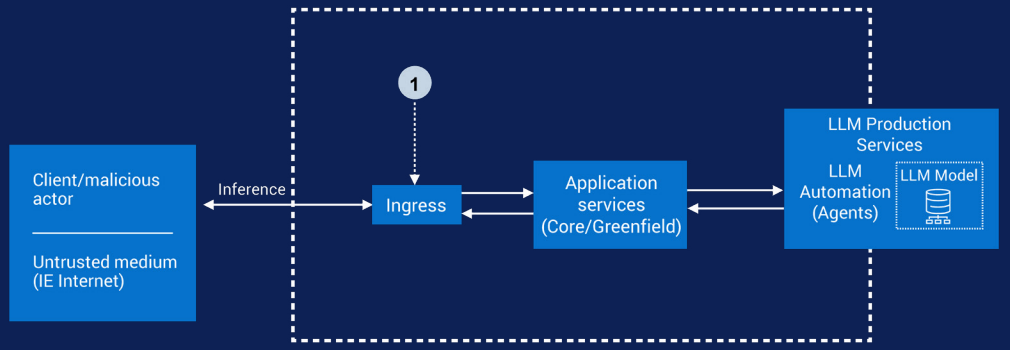


Problema n. 1: prompt injection

Strategie per mitigare il rischio di prompt injection:

- **Sanificazione dei dati e convalida degli input:** sottoponi a screening gli input degli utenti per rimuovere i contenuti dannosi. Utilizza la normalizzazione e la codifica per evitare usi impropri.
- **Approcci basati sull'elaborazione del linguaggio naturale (NLP) e sull'apprendimento automatico:** utilizza l'NLP e l'apprendimento automatico per rilevare e bloccare i prompt manipolati o malevoli.
- **Controlli precisi sulla formattazione e sulla risposta degli output:** imposta limiti rigorosi di risposta per garantire che gli output seguano i formati previsti e prevenire azioni non autorizzate. Utilizza il filtraggio dei prompt e la convalida delle risposte per preservarne l'integrità.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Prompt engineering sicuro:** utilizza la progettazione e l'analisi sicure dei prompt nell'ambito della sicurezza complessiva del software per proteggere l'elaborazione degli input.
- **Convalida dei modelli:** convalida regolarmente i modelli di ML per garantire che non siano stati manomessi prima dell'implementazione, salvaguardandone l'accuratezza e l'integrità.
- **Filtraggio e classificazione dei prompt e convalida delle risposte:** analizza e classifica i prompt per garantire che vengano elaborati solo input sicuri. Convalida le risposte per evitare usi impropri.
- **Controlli della solidità:** esegui valutazioni periodiche per identificare e correggere le vulnerabilità, mantenendo sicura e affidabile l'AI.

Torna al diagramma dell'architettura



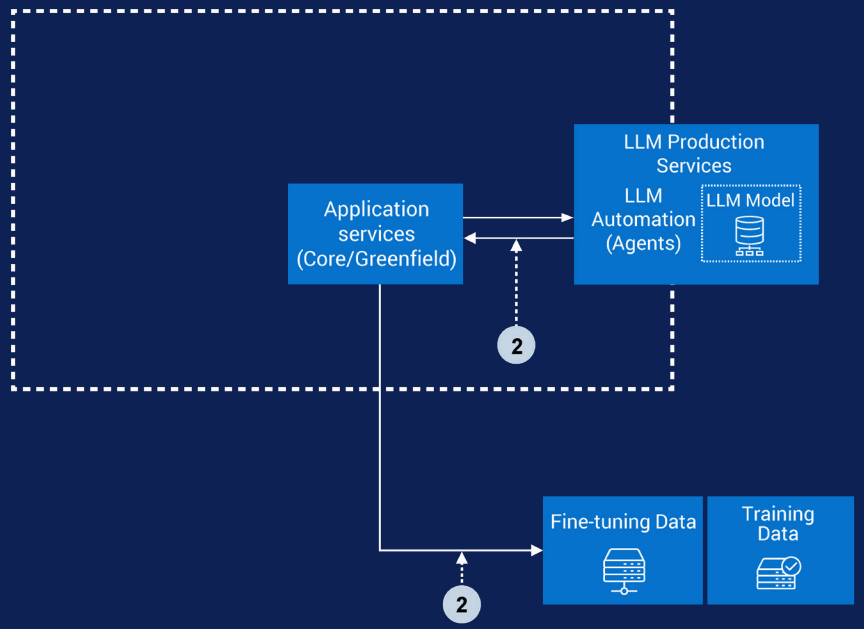
Quella della prompt injection è una sfida emergente nel mondo dell'AI generativa (GenAI), in cui gli aggressori creano input malevoli per manipolare il comportamento del modello o comprometterne l'integrità. Questi attacchi sfruttano le vulnerabilità nel modo in cui i sistemi di AI elaborano e rispondono agli input degli utenti, con la potenziale conseguenza di azioni non autorizzate, informazioni errate o esposizione di dati sensibili. Poiché la GenAI viene sempre più integrata nei flussi di lavoro aziendali critici, affrontare questi rischi è essenziale per preservare la fiducia e la sicurezza.

Problema n. 2: divulgazione di informazioni sensibili

Strategie per mitigare il rischio di divulgazione di informazioni sensibili:

- **Sanificazione dei dati e convalida degli input:** sottoponi a screening gli input degli utenti per rimuovere i contenuti dannosi. Utilizza la normalizzazione e la codifica per evitare usi impropri.
- **Utilizzo della crittografia omomorfica** per elaborare i dati sensibili in modo sicuro senza esporne il contenuto. Ciò garantisce che i dati, anche quando sono in uso, rimangano crittografati e protetti dalle violazioni.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Utilizzo di API e interfacce di sistema sicure** per le interazioni dei dati di AI, con revisione regolare delle configurazioni per ridurre al minimo l'esposizione e la superficie di attacco.
- **Protezione della data collection, dello storage e delle policy** e applicazione di policy complete di governance e protezione dei dati che garantiscono la conformità alle normative e riducono al minimo il rischio per i dati.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Sviluppo e configurazione sicuri e audit:** applica pratiche di codifica sicure, utilizza strumenti automatizzati di gestione della configurazione e conduci revisioni, audit e aggiornamenti regolari per mantenere sicure e aggiornate le configurazioni dei sistemi di AI.
- **Formazione degli utenti e sensibilizzazione alla sicurezza:** fornisci agli utenti e agli amministratori formazione continua di sensibilizzazione alla sicurezza per ridurre l'utilizzo non sicuro e la divulgazione accidentale dei dati.

Torna al diagramma dell'architettura



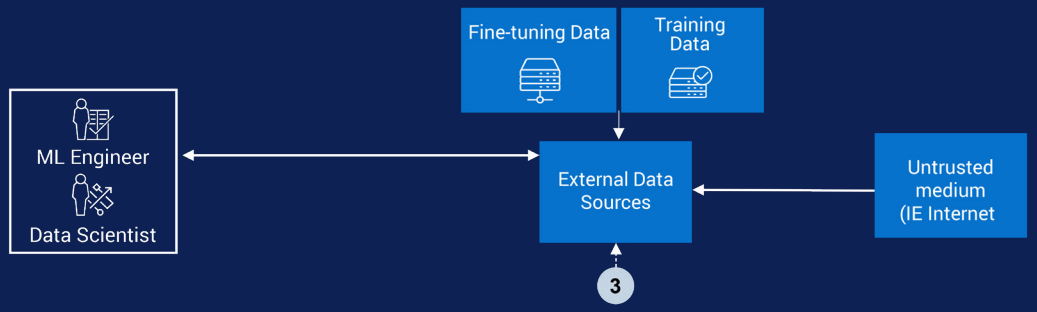
La GenAI ha prodotto progressi incredibili, ma comporta anche rischi notevoli, in particolare l'esposizione involontaria di informazioni sensibili. Che si tratti di informazioni di identificazione personale (PII) o di dati aziendali proprietari, l'uso improprio o la gestione errata degli strumenti di GenAI può causare fughe di dati, mancata conformità alle normative o danni alla reputazione. È quindi cruciale per le organizzazioni comprendere questi rischi e affrontarli in modo proattivo per garantire l'implementazione e l'utilizzo sicuri dei sistemi di AI.

Problema n. 3: vulnerabilità nella supply chain

Strategie per mitigare il rischio di vulnerabilità nella supply chain:

- **Verifica dei fornitori e garanzia di conformità con procedure sicure per la supply chain:** valuta i fornitori e stabilisci contratti che diano priorità alla sicurezza della supply chain.
- **Implementazione di distinte base del software:** monitora e verifica le origini dei componenti software per garantire trasparenza e ridurre il rischio di codice compromesso.
- **Convalida dei modelli:** convalida regolarmente i modelli di ML per garantire che non siano stati manomessi prima dell'implementazione, salvaguardandone l'accuratezza e l'integrità.
- **Esecuzione di container e pod con privilegi minimi:** in questo modo si riduce il potenziale impatto in caso di compromissione e si limita l'accesso non autorizzato.
- **Implementazione di firewall:** blocca la connettività di rete non necessaria, riducendo l'esposizione alle potenziali minacce e limitando le vie di accesso per gli autori di attacchi informatici.
- **Protezione dei dati e delle annotazioni:** proteggi i dati e le annotazioni associate per evitare manomissioni, accessi non autorizzati e danneggiamento delle informazioni critiche.
- **Hardware sicuro:** utilizza hardware convalidato per la sicurezza per prevenire le vulnerabilità che potrebbero derivare da attacchi basati su hardware, assicurando una solida base per l'infrastruttura.
- **Protezione dei componenti software per l'apprendimento automatico:** utilizza componenti software affidabili e verificati per l'apprendimento automatico per ridurre le vulnerabilità e migliorare la sicurezza complessiva dei flussi di lavoro di apprendimento automatico.
- **Sviluppo e configurazione sicuri e audit:** applica pratiche di codifica sicure, utilizza strumenti automatizzati di gestione della configurazione e conduci revisioni, audit e aggiornamenti regolari per mantenere sicure e aggiornate le configurazioni dei sistemi di AI.

Torna al diagramma dell'architettura



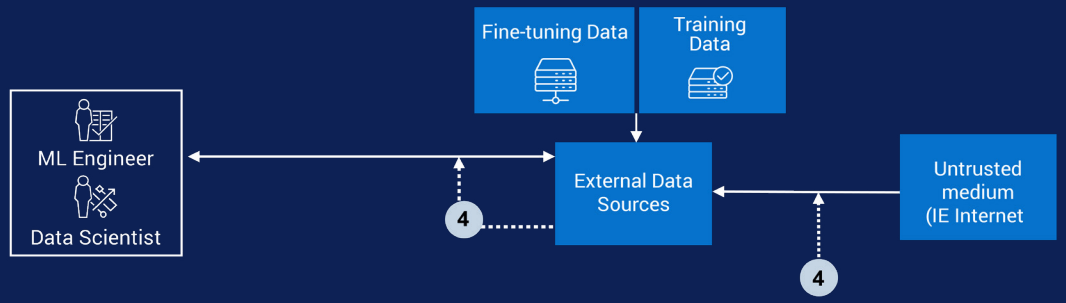
Analizza le vulnerabilità nella supply chain degli LLM, poiché possono influire sui componenti critici come l'integrità dei modelli pre-addestrati e gli adattatori di terze parti. I sistemi di AI si basano su hardware e software che potrebbero essere compromessi molto prima dell'implementazione. Gli avversari sfruttano i punti deboli in varie fasi della supply chain per l'apprendimento automatico, puntando all'hardware delle GPU, ai dati e alle relative annotazioni, agli elementi dello stack software per l'apprendimento automatico o persino al modello stesso. Compromettendo queste parti specifiche, gli autori degli attacchi possono ottenere l'accesso iniziale ai sistemi e di conseguenza creare rischi notevoli per la sicurezza e l'integrità. Comprendere e mitigare queste vulnerabilità è fondamentale per creare soluzioni di AI solide e sicure.

Problema n. 4: inquinamento dei dati del modello

Strategie per mitigare il rischio di inquinamento dei dati del modello:

- **Utilizzo del rilevamento delle anomalie e convalida dei dati durante l'addestramento** per identificare e risolvere le incoerenze nei dati e garantire che vengano utilizzati solo dati puliti e di alta qualità per addestrare il modello.
- **Isolamento degli ambienti durante le fasi di ottimizzazione** per impedire l'accesso non autorizzato o la contaminazione del modello durante le fasi critiche di sviluppo.
- **Convalida dei modelli:** convalida regolarmente i modelli di ML per garantire che non siano stati manomessi prima dell'implementazione, salvaguardandone l'accuratezza e l'integrità.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Sanificazione dei dati e convalida degli input:** sottoponi a screening gli input degli utenti per rimuovere i contenuti dannosi. Utilizza la normalizzazione e la codifica per evitare usi impropri.
- **Sviluppo e configurazione sicuri e audit:** applica pratiche di codifica sicure, utilizza strumenti automatizzati di gestione della configurazione e conduci revisioni, audit e aggiornamenti regolari per mantenere sicure e aggiornate le configurazioni dei sistemi di AI.
- **Controlli della solidità:** esegui valutazioni periodiche per identificare e correggere le vulnerabilità, mantenendo sicura e affidabile l'AI.
- **Implementazione della segmentazione della rete** per limitare l'accesso a interfacce non sicure e componenti critici del sistema.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.

Torna al diagramma dell'architettura



L'inquinamento dei dati del modello è una minaccia alla sicurezza durante il ciclo di vita dell'AI in cui gli avversari contaminano intenzionalmente i dati di addestramento con input danneggiati, fuorvianti o malevoli. Questo rischio può influire su componenti critici, dalla raccolta e annotazione dei dati non elaborati alla curation e all'integrazione dei dataset utilizzati per l'apprendimento automatico o per i modelli linguistici di grandi dimensioni. L'affidabilità dei sistemi di AI dipende dall'integrità delle origini dati, che potrebbero essere esposte a manipolazioni prima dell'addestramento, durante la pre-elaborazione o tramite pipeline di dati esterne.

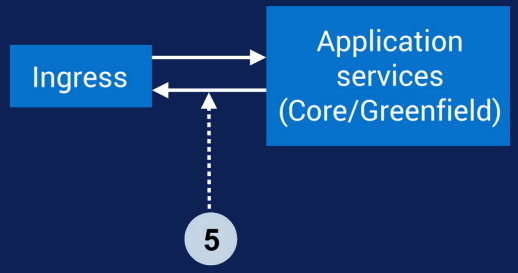
Gli autori degli attacchi sfruttano l'inquinamento dei dati per ridurre l'accuratezza dei modelli, introdurre vulnerabilità o attivare output dannosi. Individuando i punti deboli nella provenienza dei dati, nella qualità delle annotazioni o nei processi di acquisizione dei dataset, gli avversari possono compromettere la sicurezza, l'affidabilità e la resilienza. Riconoscere e mitigare queste minacce incentrate sui dati è essenziale per creare soluzioni di AI solide e affidabili.

Problema n. 5: gestione impropria dell'output

Strategie per mitigare il rischio di gestione impropria dell'output:

- **Codifica dell'output con riconoscimento del contesto:** applica sempre tecniche di codifica ed escaping personalizzate in base al contesto specifico in cui verrà utilizzato l'output, ad esempio ambienti HTML, SQL o API, per prevenire vulnerabilità come gli attacchi di tipo injection.
- **Sanificazione dell'output:** segui rigorose pratiche di convalida e sanificazione per gli output dei modelli in linea con le linee guida Application Security Verification Standard (ASVS) dell'Open Web Application Security Project (OWASP) per garantire un utilizzo downstream sicuro e ridurre i rischi per la sicurezza.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Test automatizzati di sicurezza degli output:** conduci di test di sicurezza regolari con strumenti automatizzati per identificare i rischi negli output, ad esempio il cross-site scripting (XSS) o le vulnerabilità di tipo injection, e affrontali in modo proattivo.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Revisione Human-in-the-loop:** per le applicazioni ad alto rischio, ad esempio nel settore finanziario o sanitario, è necessaria la supervisione umana e la revisione degli output del modello per garantire accuratezza, protezione e sicurezza.
- **Privacy e conformità:** integra tecniche di tutela della privacy nel processo di output e garantisci la conformità alle normative e agli standard pertinenti per l'utilizzo sicuro delle informazioni sensibili.

Torna al diagramma dell'architettura



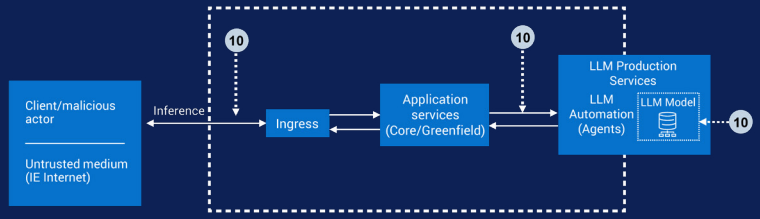
La convalida o la sanificazione insufficienti dell'output del modello di AI può comportare gravi rischi per la sicurezza, tra cui escalation dei privilegi e violazioni dei dati. Quando i modelli di AI producono output non correttamente controllati o filtrati, i malintenzionati hanno la possibilità di sfruttare queste vulnerabilità per ottenere accesso non autorizzato o eseguire l'escalation dei privilegi all'interno di un sistema. Questa mancanza di supervisione può causare compromissione dei dati, azioni non autorizzate e violazioni significative della sicurezza, il che evidenzia l'importanza di implementare solidi processi di convalida e sanificazione per qualsiasi output generato dall'AI.

Problema n. 6: excessive agency

Strategie per mitigare il rischio di excessive agency

- **Applicazione del privilegio minimo:** concedi a LLM e sottosistemi agentici solo le autorizzazioni minime necessarie per eseguire le operazioni previste e rivedi regolarmente i controlli degli accessi.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Impostazione di limiti operativi:** definisci chiaramente quali LLM/agenti possono accedere o eseguire operazioni.
- **Revisione Human-in-the-loop:** per le applicazioni ad alto rischio, ad esempio nel settore finanziario o sanitario, è necessaria la supervisione umana e la revisione degli output del modello per garantire accuratezza, protezione e sicurezza.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Limitazione dell'autonomia:** limita le capacità degli LLM per evitare l'accesso o il controllo senza restrizioni.
- **Sviluppo e configurazione sicuri e audit:** applica pratiche di codifica sicure, utilizza strumenti automatizzati di gestione della configurazione e conduci revisioni, audit e aggiornamenti regolari per mantenere sicure e aggiornate le configurazioni dei sistemi di AI.
- **Implementazione di firewall:** blocca la connettività di rete non necessaria, riducendo l'esposizione alle potenziali minacce e limitando le vie di accesso per gli autori di attacchi informatici.
- **Controlli della solidità:** esegui valutazioni periodiche per identificare e correggere le vulnerabilità, mantenendo sicura e affidabile l'AI.

Torna al diagramma dell'architettura



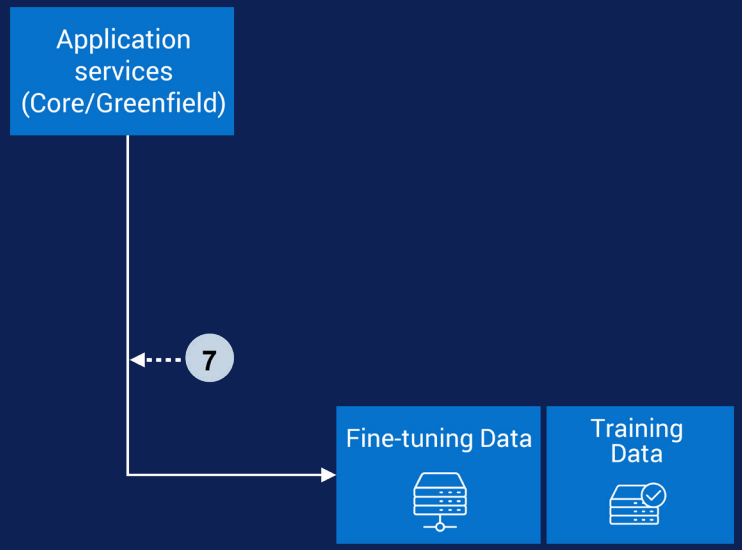
Concedere ai plug-in o agli agenti AI un'eccessiva autonomia o funzionalità non necessarie all'interno dei flussi di lavoro può creare rischi significativi. Quando a un sistema di AI vengono forniti privilegi o capacità al di là di quanto necessario, aumenta la probabilità di conseguenze indesiderate. Ciò può verificarsi quando i sistemi basati su modelli linguistici di grandi dimensioni (LLM) sono progettati con autorizzazioni eccessive, consentendo loro di intraprendere azioni o accedere a informazioni che non dovrebbero. Tale eccesso può causare errori, uso improprio dei dati o persino vulnerabilità della sicurezza, il che evidenzia l'importanza di limitare e monitorare attentamente le capacità di AI per garantire un utilizzo sicuro e responsabile.

Problema n. 7: prompt leakage

Strategie per mitigare il rischio di prompt leakage

- **Nessun incorporamento di informazioni sensibili nei prompt:** non includere mai credenziali, chiavi API o logica proprietaria nei prompt e gestisci in modo sicuro questi elementi all'esterno del sistema.
- **Separazione dei controlli di sicurezza dai prompt:** gestisci l'autenticazione, l'autorizzazione e la gestione delle sessioni nella logica delle applicazioni, non nei prompt.
- **Convalida degli input e degli output:** sanifica i prompt e le risposte con una convalida affidabile per bloccare manipolazioni o modelli sospetti.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Crittografia e protezione dei prompt:** archivia i prompt e le configurazioni in uno storage crittografato e sicuro per prevenire l'accesso non autorizzato.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Revisione periodica dei prompt** esamina e sanifica periodicamente i prompt per rimuovere i dati sensibili e garantire la conformità della sicurezza.
- **Test e attività di red teaming per individuare i punti deboli:** conduci test di simulazione di attacchi per identificare e correggere le vulnerabilità nella gestione dei prompt o negli output.
- **Isolamento dei prompt dagli input degli utenti:** progetta i sistemi in modo da evitare che le query degli utenti manipolino o espongano i prompt.
- **Applicazione di limiti di frequenza:** limita l'utilizzo delle API, circoscrivi le attività sospette e blocca gli attacchi automatizzati ai prompt.

Torna al diagramma dell'architettura



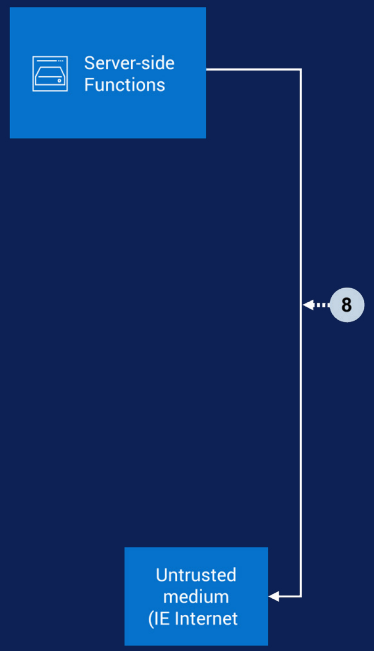
Un attacco di tipo system prompt leakage a un modello linguistico di grandi dimensioni (LLM) o un sistema di AI si verifica quando un malintenzionato è in grado di estrarre o derivare le istruzioni nascoste, ovvero i "prompt di sistema", che guidano il comportamento del modello e definiscono i limiti operativi. Questi prompt non sono in genere destinati a essere visibili agli utenti finali, in quanto contengono regole di base, limitazioni e logiche operative talvolta sensibili. Tramite input appositamente realizzati o lo sfruttamento di vulnerabilità, un malintenzionato può indurre l'LLM a rivelare in tutto o in parte il prompt di sistema. In caso di fuga, queste informazioni possono essere utilizzate per decompilare le restrizioni, bypassare i filtri di sicurezza o sviluppare nuovi attacchi mirati, aumentando alla fine il rischio di prompt injection, escalation dei privilegi o uso improprio del modello e dei sistemi downstream che dipendono dalla loro integrità.

Problema n. 8: punti deboli nei vettori e negli embedding

Strategie per mitigare il rischio di punti deboli nei vettori e negli embedding

- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Crittografia:** proteggi i dati vettoriali inattivi e in transito utilizzando standard di crittografia affidabili come AES.
- **Configurazione e monitoraggio sicuri:** rafforza i sistemi, esegui la configurazione in modo sicuro e monitora costantemente per rilevare la presenza di configurazioni errate, accessi non autorizzati o anomalie.
- **Gestione delle vulnerabilità:** aggiorna e applica regolarmente le patch a tutti i software, le dipendenze e gli engine degli archivi vettoriali per affrontare i rischi per la sicurezza.
- **Sanificazione dei dati e convalida degli input:** sottoponi a screening gli input degli utenti per rimuovere i contenuti dannosi. Utilizza la normalizzazione e la codifica per evitare usi impropri.
- **Utilizzo di API e interfacce di sistema sicure** per le interazioni dei dati di AI, con revisione regolare delle configurazioni per ridurre al minimo l'esposizione e la superficie di attacco.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Hardware sicuro:** utilizza hardware convalidato per la sicurezza per prevenire le vulnerabilità che potrebbero derivare da attacchi basati su hardware, assicurando una solida base per l'infrastruttura.
- **Sviluppo e configurazione sicuri e audit:** applica pratiche di codifica sicure, utilizza strumenti automatizzati di gestione della configurazione e conduci revisioni, audit e aggiornamenti regolari per mantenere sicure e aggiornate le configurazioni dei sistemi di AI.

Torna al diagramma dell'architettura



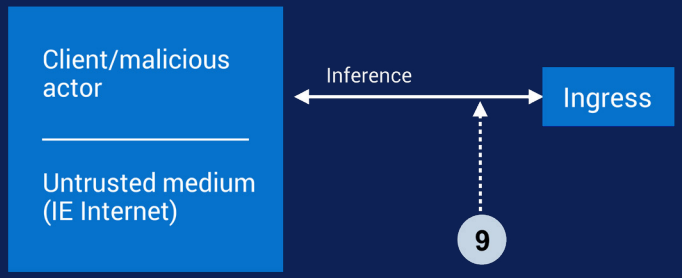
Un attacco ai punti deboli nei vettori e negli embedding di un modello linguistico di grandi dimensioni (LLM) o un sistema di AI, in particolare quelli che utilizzano tecniche di Retrieval-Augmented Generation (RAG), mira a vulnerabilità correlate al modo in cui le informazioni vengono codificate, archiviate e recuperate, ad esempio a vettori numerici ed embedding. I punti deboli di questi meccanismi possono essere sfruttati attraverso azioni malevole come l'inversione degli embedding (ricostruzione dei dati sensibili dagli embedding), l'inquinamento dei dati (inserimento di contenuti dannosi o distorti per manipolare il comportamento del modello), l'accesso non autorizzato ai database vettoriali (con conseguenti fughe di dati) o la manipolazione degli output di recupero. Questi attacchi sono una minaccia per la privacy, l'integrità e l'affidabilità in quanto consentono agli autori degli attacchi di divulgare informazioni sensibili, alterare gli output o minare la fiducia degli utenti nelle applicazioni basate sull'AI. Controlli degli accessi adeguati, convalida dei dati, crittografia e monitoraggio costante sono misure fondamentali per difendersi da queste minacce in continua evoluzione.

Problema n. 9: informazioni errate

Strategie per mitigare il rischio di informazioni errate

- **Retrieval-Augmented Generation (RAG) con origini autorevoli:** utilizza la RAG per recuperare e integrare informazioni da database e repository di conoscenze verificati e affidabili, riducendo le allucinazioni.
- **Ottimizzazione dei modelli e calibrazione dell'output:** ottimizza i modelli con dataset diversificati e applica tecniche per ridurre al minimo pregiudizi e informazioni errate.
- **Controllo automatizzato delle informazioni:** incrocia gli output con origini affidabili e segnala le informazioni false in modo automatico.
- **Monitoraggio delle incertezze:** segnala le risposte a baso livello di attendibilità per la revisione umana in casi critici.
- **Revisione Human-in-the-loop:** per le applicazioni ad alto rischio, ad esempio nel settore finanziario o sanitario, è necessaria la supervisione umana e la revisione degli output del modello per garantire accuratezza, protezione e sicurezza.
- **Feedback degli utenti:** fai in modo che gli utenti abbiano la possibilità di segnalare gli errori per il miglioramento continuo del modello e la correzione rapida dei percorsi delle informazioni errate.
- **Restrizioni di accesso e supervisione umana:** applica il controllo degli accessi basato sui ruoli (RBAC), l'autenticazione a più fattori (MFA) e la gestione delle identità per limitare l'accesso. Utilizza la revisione umana per l'adozione di decisioni critiche.
- **Sviluppo e configurazione sicuri e audit:** applica pratiche di codifica sicure, utilizza strumenti automatizzati di gestione della configurazione e conduci revisioni, audit e aggiornamenti regolari per mantenere sicure e aggiornate le configurazioni dei sistemi di AI.
- **Comunicazione dei rischi:** istruisci gli utenti sui limiti dell'AI e sull'importanza della verifica indipendente.
- **Progettazione intenzionale di interfacce utente e API:** evidenzia i contenuti generati dall'AI e guida gli utenti sull'utilizzo responsabile.

Torna al diagramma dell'architettura



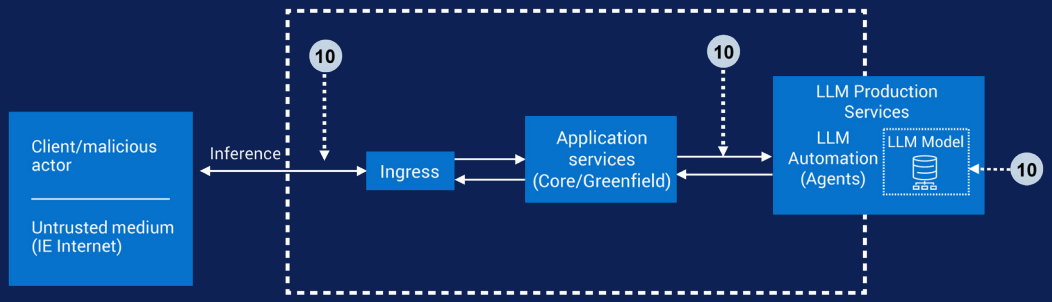
Un attacco di informazioni errate a un sistema LLM o AI è il tentativo intenzionale di far sì che il modello generi o diffonda informazioni false, fuorvianti o errate, ma apparentemente credibili, attraverso i suoi output. Questa vulnerabilità deriva da diversi fattori: tendenza del modello a creare "allucinazioni" (generazione di contenuti inventati ma che sembrano plausibili), pregiudizi o lacune presenti nei dati di addestramento e influenza di prompt antagonisti. Le allucinazioni si verificano quando gli LLM, anziché comprendere realmente i dati, generano statisticamente contenuti che seguono comunque uno schema, con conseguenti risposte che sembrano autorevoli, ma che sono in realtà infondate. I rischi di tali attacchi includono violazioni della sicurezza, danni alla reputazione e persino conseguenze legali, in particolare negli ambienti in cui gli utenti pongono un'eccessiva fiducia nelle risposte generate dagli LLM senza verificarne l'accuratezza o la validità, integrando potenzialmente errori o informazioni errate in decisioni e processi critici.

Problema n. 10: consumo non vincolato

Strategie per ridurre il rischio di consumo non vincolato

- **Applicazione di limiti di frequenza e quote utente:** imposta limiti rigorosi di richieste, token o dati per utente, chiavi API o app al fine di evitare usi impropri.
- **Autenticazione e segmentazione degli utenti:** utilizza tecniche di autenticazione avanzata (ad esempio chiavi API, OAuth) e assegna ruoli o tier per elaborare solo le richieste autorizzate.
- **Convalida dell'input e restrizioni sulle dimensioni:** convalida le dimensioni e la struttura dei prompt, bloccando o riducendo le query di grandi dimensioni o non valide.
- **Applicazione di timeout di elaborazione e limitazione delle risorse:** imposta timeout e limiti di risorse per ogni richiesta al fine di evitare operazioni a esecuzione prolungata e il consumo delle risorse.
- **Implementazione di caching intelligente e deduplica:** memorizza le risposte nella cache per eventuali query duplicate o simili in modo da ridurre l'elaborazione inutile.
- **Monitoraggio, registrazione e rilevamento delle anomalie:** monitora e registra costantemente le attività dei sistemi di AI con soluzioni come MDR/XDR/SIEM per rilevare, analizzare e rispondere rapidamente ad accessi non autorizzati, anomalie o fughe di dati.
- **Monitoraggio del budget e controlli di spesa:** utilizza dashboard e avvisi per monitorare i costi e blocca l'utilizzo al raggiungimento delle soglie di budget.
- **Tecniche di sandboxing e isolamento:** esegui i carichi di lavoro in ambienti isolati con autorizzazioni limitate per ridurre i rischi.
- **Limitazione della profondità delle chiamate e dei turni di conversazione:** imponi limiti alle chiamate ricorsive o ai passaggi di conversazione per evitare lo sfruttamento delle risorse.
- **Applicazione di un modello a più livelli o dell'allocazione delle risorse:** indirizza le richieste ad alta priorità ai modelli premium e il traffico a bassa priorità a quelli a costi più contenuti.

Torna al diagramma dell'architettura



La minaccia di consumo non vincolato a un sistema LLM o AI si riferisce a una vulnerabilità della sicurezza in cui l'applicazione consente agli utenti, malintenzionati e non, di inviare richieste di inferenza o prompt in numero eccessivo senza efficaci restrizioni di limite di frequenza, autenticazione o utilizzo. Poiché l'inferenza degli LLM è dispendiosa dal punto di vista computazionale, questa mancanza di controllo può essere sfruttata in diversi modi: i malintenzionati possono causare un attacco Denial of Service (DoS) sovraccaricando le risorse di sistema, generare perdite economiche impreviste nelle implementazioni pay-per-use o in hosting nel cloud oppure interrogare sistematicamente il modello per clonarne il comportamento e sottrarre la proprietà intellettuale. Le conseguenze includono interruzioni del servizio, riduzione delle prestazioni per altri utenti, conseguenze finanziarie e aumento del rischio di fuga di modelli sensibili. In sostanza, il consumo non vincolato si verifica quando l'utilizzo delle risorse non è governato correttamente, lasciando le applicazioni basate su LLM esposte a exploit sia accidentali che deliberati.

Perché scegliere Dell per la sicurezza dell'AI

Dell favorisce la protezione dei modelli di AI e degli LLM delle organizzazioni attraverso un approccio completo che include hardware, software e servizi gestiti. La sicurezza è integrata dalla supply chain fino a dispositivi, infrastrutture, dati e applicazioni, il tutto in linea con i principi Zero Trust. Le soluzioni dell'intero portafoglio Dell sono progettate per migliorare l'igiene informatica con funzionalità come MFA, RBAC, privilegio minimo e verifica continua. Questo approccio completo e "sicuro by design" garantisce alle organizzazioni la possibilità di innovare in tutta sicurezza con l'AI e gli LLM, riducendo al minimo il rischio di furto di modelli, fuga di dati, attacchi antagonisti e altre minacce informatiche avanzate.

Supply chain

La supply chain sicura Dell offre la protezione di base necessaria per i modelli di AI e gli LLM integrandola sicurezza in ogni fase dello sviluppo, della produzione e della consegna dei prodotti. Grazie a funzionalità quali aggiornamenti del BIOS e del firmware con firma crittografica, Secured Component Verification, distinta base del software (SBOM) incentrata sull'AI, monitoraggio della derivazione dei dataset, configurazione e software di sicurezza integrati e rigorosi assessment dei rischi dei vendor in linea con gli standard globali, Dell riduce al minimo i rischi derivanti da manomissioni, accessi non autorizzati e attacchi alla supply chain, garantendo alle organizzazioni la possibilità di implementare carichi di lavoro di AI affidabili e resilienti con la massima trasparenza, integrità e conformità alle normative.

AI PC

Dell offre sicurezza fondamentale per i carichi di lavoro di AI on-device. I Dell Trusted Device, gli AI PC commerciali più sicuri al mondo*, sono progettati pensando alla sicurezza. La sicurezza della supply chain riduce al minimo il rischio di vulnerabilità e manomissioni nei prodotti. Le difese esclusive integrate direttamente nell'hardware e nel firmware mantengono protetti il PC e l'utente finale durante l'uso. Dell SafeBIOS offre visibilità profonda a livello di BIOS e rilevamento delle manomissioni, mentre Dell SafeID migliora la sicurezza delle credenziali e abilita l'autenticazione senza password. Il software dei partner fornisce protezione avanzata in tutti gli ambienti endpoint, di rete e cloud.

Cyber-resilienza

Le soluzioni di cyber-resilienza Dell PowerProtect proteggono i dati di AI con backup crittografati e immutabili, ripristino rapido e vault di Cyber Recovery isolati. Queste funzionalità impediscono la distruzione dei dati, riducono l'impatto di aggiornamenti malevoli e supportano la conformità e il ripristino dopo un attacco.

Server

I server PowerEdge sono dotati di elaborazione riservata per isolare e proteggere i prompt e gli embedding di AI e LLM, soluzioni di Retrieval-Augmented Generation (RAG) affidabili basate su origini autorevoli, unitamente a funzionalità MFA, RBAC, Silicon Root of Trust, firmware con firma e monitoraggio continuo per proteggere i carichi di lavoro di AI critici.

Storage

Il portafoglio di storage Dell garantisce l'archiviazione sicura e crittografata dei dati di AI sensibili con la solida crittografia AES-256 per i dati inattivi e in transito. Su specifiche offerte è disponibile una crittografia avanzata progettata per essere resiliente alle future minacce quantistiche. Il portafoglio include prestazioni NVMe ad alta velocità, moduli di crittografia conformi a FIPS per proteggere i dati, inclusi quelli utilizzati nei carichi di lavoro di AI, snapshot non modificabili e vault di Cyber Recovery con air gap per contrastare gli attacchi ransomware. L'architettura Zero Trust, la sicurezza della supply chain e le funzionalità di audit a prova di manomissione migliorano la governance. Il rilevamento delle anomalie integrato e i modelli ML AIOps proteggono i carichi di lavoro senza utilizzare i dati dei clienti per l'addestramento, riducendo al minimo i rischi di attacchi basati sull'input.

AIOps

Dell AIOps offre monitoraggio continuo automatizzato per rilevare configurazioni errate e vulnerabilità (incluse le CVE) e supporta la consapevolezza dei rischi della supply chain che influisce sui carichi di lavoro di AI/LLM. La scansione delle CVE in tempo reale, gli avvisi intelligenti e i dashboard basati sull'AI consentono un intervento rapido grazie alla segnalazione delle anomalie e al monitoraggio dei flussi di lavoro di risoluzione. Le funzionalità di conformità integrate, i controlli degli accessi basati sui ruoli e il reporting automatizzato contribuiscono a mantenere sicure le operazioni nei carichi di lavoro, mentre l'integrazione perfetta di EDR/XDR e le informazioni operative basate sull'AI, incluse le capacità generative nelle soluzioni supportate, migliorano ulteriormente l'efficienza IT.

Networking

Le soluzioni Dell Networking proteggono gli ambienti AI/LLM attraverso un'affidabile segmentazione della rete, riducendo al minimo il movimento laterale. I percorsi di rete crittografati e i controlli firewall integrati bloccano l'accesso non autorizzato ai dati di AI.

Servizi di sicurezza e resilienza per l'AI

I servizi Dell di sicurezza e resilienza per l'AI sono attentamente realizzati per affrontare i nuovi rischi associati all'integrazione dell'AI nell'organizzazione. Progettati per la collaborazione con i team dell'organizzazione durante l'integrazione dell'AI nel più breve tempo possibile, i nostri servizi offrono le competenze necessarie per guidare la pianificazione strategica e l'implementazione delle soluzioni, nonché servizi di sicurezza gestiti per alleggerire i carichi operativi, in modo da poter innovare in modo sicuro con l'AI. Ognuno di essi è personalizzato per aiutare le organizzazioni ad affrontare i rischi dell'AI in continua evoluzione e ottimizzare le implementazioni sicure dell'AI.

Dell AI Factory

Portafoglio integrato di sicurezza specifica, tra cui supply chain sicura Dell, funzionalità Zero Trust per applicare il privilegio minimo e soluzioni MDR per l'AI progettate per mantenere sicuro e protetto il tuo modello.

* Dati basati su un'analisi interna Dell, ottobre 2024 (Intel) e marzo 2025 (AMD). Applicabile ai PC con processori Intel e AMD. Non tutte le funzionalità sono disponibili per tutti i PC. Sono necessari ulteriori acquisti per alcune funzionalità. PC basati su Intel convalidati da Principled Technologies, luglio 2025.

Conclusioni

Per creare framework di AI resilienti, è essenziale adottare un approccio collaborativo tra organizzazioni ed gli esperti di sicurezza. Poiché l'AI e gli LLM continuano a rimodellare i settori, è fondamentale affrontare i rischi che comportano, tra cui la sicurezza dei dati, l'integrità dei modelli e i problemi di conformità. Le organizzazioni devono dare priorità a strategie proattive che integrino la sicurezza in ogni fase del loro percorso verso l'AI.

Dell Technologies è un partner di fiducia in questa missione, offrendo personalizzazione della GenAI end-to-end, consulenza per la sicurezza e soluzioni integrate su misura per le tue esigenze specifiche. Sfruttando le solide soluzioni di sicurezza informatica di Dell, le aziende riducono efficacemente i rischi associati all'AI e agli LLM e al contempo massimizzano il potenziale degli investimenti in sicurezza esistenti. Dell consente alle organizzazioni di proteggere l'infrastruttura AI attraverso la perfetta integrazione di misure di sicurezza avanzata nei loro framework attuali, garantendo un ambiente sicuro e orientato al futuro.

Scopri in che modo le soluzioni AI complete di Dell proteggono gli ambienti di GenAI e LLM: Dell.com/CyberSecurityMonth