DELL Technologies

De 10 belangrijkste zorgen over cyberbeveiliging voor GenAI en LLM's

Aan de slag

Introductie

Kunstmatige intelligentie (AI) zorgt voor een revolutie in de manier waarop organisaties werken. Hierbij worden generatieve AI (GenAI) en grote taalmodellen (LLM's) kritieke workloads in moderne bedrijfsomgevingen.

Net als elke andere workload hebben deze applicaties hun eigen reeks complexiteiten en kwetsbaarheden om aan te pakken. Aangezien bedrijven AI steeds verder implementeren om innovatie, efficiëntie en concurrentievoordeel te stimuleren, wordt het waarborgen van de beveiliging van deze applicaties een absolute noodzaak. Goede cyberhygiëne vormt de basis voor het beveiligen van elke workload. Net zoals u prioriteit geeft aan beveiliging voor al uw workloads, is het essentieel om ook goede cyberhygiëne te hanteren voor AI. Dit omvat implementatieprocedures zoals de juiste systeempatches, meervoudige verificatie, op rollen gebaseerde toegang en netwerksegmentatie. Deze maatregelen zijn fundamenteel, maar de echte sleutel is inzicht in hoe deze mogelijkheden passen bij de specifieke architectuur en het gebruik van uw workload.

Bij Dell hebben we een diepgaand inzicht in AI-workloads en de unieke beveiligingsuitdagingen die hierbij optreden. Door vast te stellen op welke manieren bedreigingsactoren deze workloads kunnen targeten, helpt Dell u een robuuste beveiligingsstrategie te creëren. Dit omvat het aanpakken van risico's zoals vergiftiging van trainingsdata, diefstal of manipulatie van modellen, reconstructie van datasets en meer.

We richten ons ook op het beheer van uitdagingen in verband met de input voor uw AI-model, zoals voorkomen dat gevoelige informatie bekend wordt, onveilige onderwerpen of vooroordelen beperken en wettelijke naleving waarborgen. Aan de uitvoerzijde helpen we problemen aan te pakken zoals overafhankelijkheid van het model en nalevingsgerelateerde risico's.

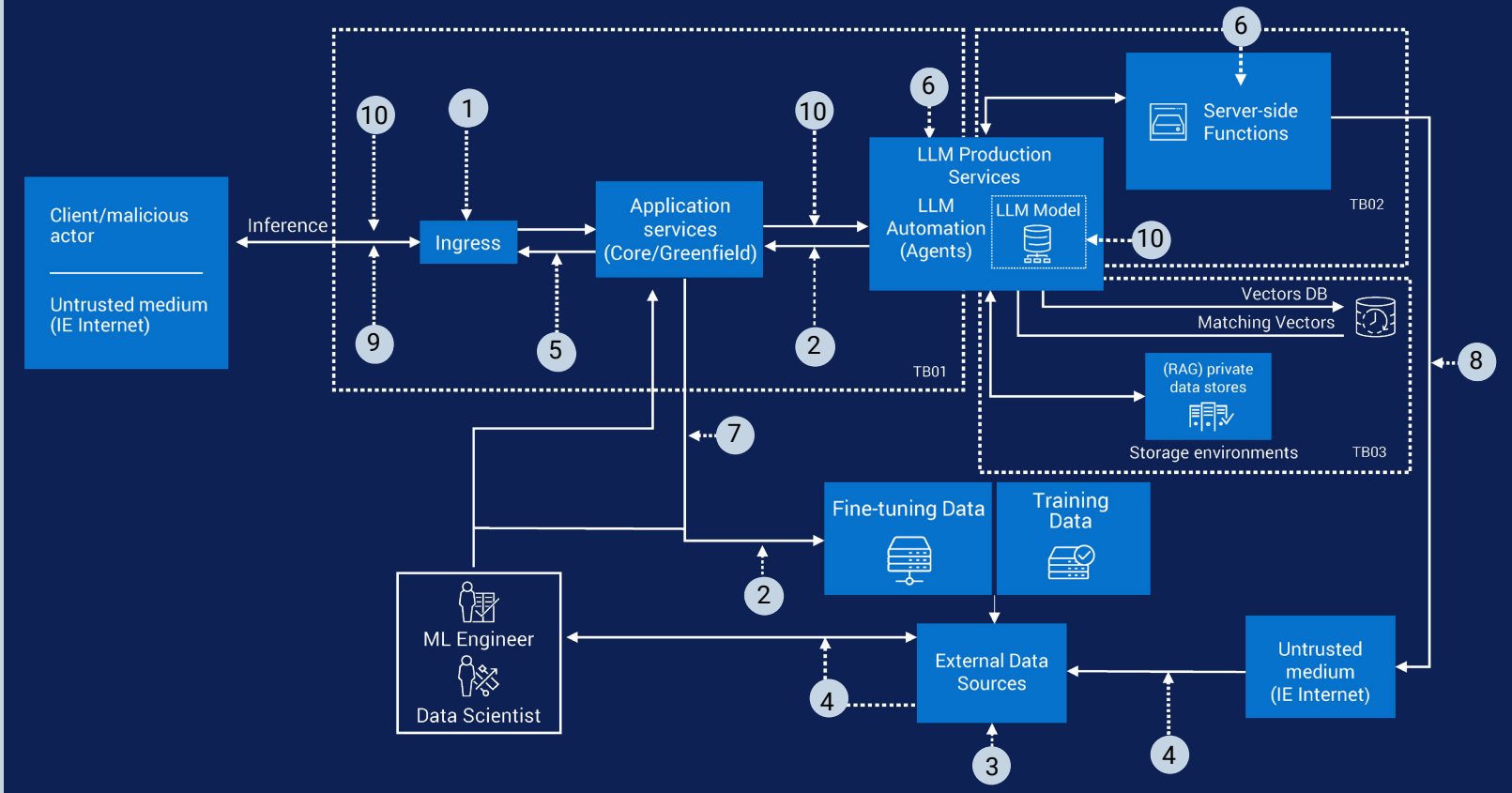
Bij Dell helpen we ondernemingen deze risico's te beperken door hun bestaande cyberbeveiligingsoplossingen te benutten of nieuwe tools en procedures te verkennen om hun systemen te beschermen. Ons doel is ervoor te zorgen dat beveiliging geen obstakel is voor innovatie. Door te begrijpen hoe AI-workloads werken en met welke beveiligingsdreigingen ze worden geconfronteerd, kunnen we u helpen een sterkere beveiligingsmentaliteit op te bouwen, zodat uw omgeving veerkrachtiger wordt en u vol vertrouwen kunt innoveren. Met onze expertise helpen we u het potentieel van AI met zekerheid te benutten en tegelijkertijd een robuuste beveiliging te handhaven.

De 10 belangrijkste zorgen over cyberbeveiliging voor GenAI en LLM

Dit zijn de belangrijkste zorgen voor de bescherming van GenAI en LLM-modellen volgens OWASP.

Klik op een zorg voor meer informatie:

1. Snelle implementatie
2. Openbaarmaking van gevoelige informatie
3. Leveringsketen
4. Vergiftiging van modeldata
5. Onjuiste verwerking van uitvoer
6. Buitensporige keuzevrijheid
7. Weglekken van systeemprompts
8. Zwakke punten in vectoren en inbeddingen
9. Verkeerde informatie
10. Onbeperkte consumptie

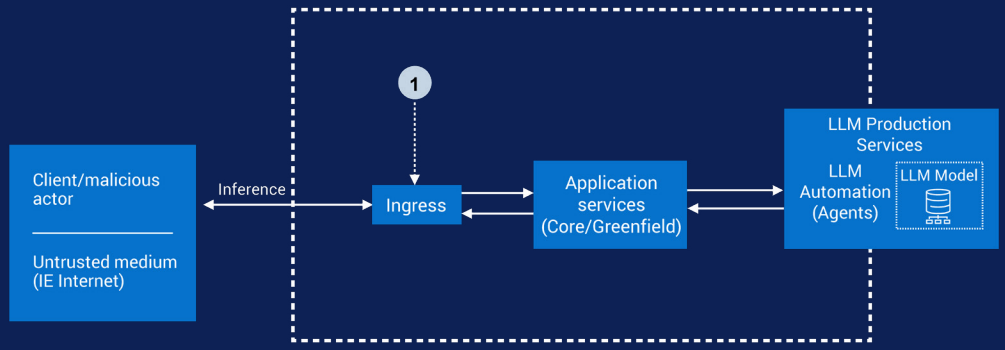


Eerste zorg: promptinjectie

Strategieën om promptinjectie te beperken:

- **Dataopschoning en invoervalidatie:** Controleer gebruikersinvoer grondig om schadelijke content te verwijderen. Gebruik normalisatie en codering om misbruik te voorkomen.
- **Op natuurlijke taalverwerking (NLP) en machine learning gebaseerde benaderingen:** Gebruik NLP en machine learning om gemanipuleerde of schadelijke prompts te detecteren en te blokkeren.
- **Duidelijke indeling voor uitvoer en controles op reacties:** Stel strikte reactiegrenzen in om ervoor te zorgen dat uitvoer de beoogde indelingen volgt en om ongeoorloofde acties te voorkomen. Gebruik promptfiltering en responsvalidatie om de integriteit te behouden.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Bewaking, logboekregistratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Veilige promptengineering:** Gebruik veilig promptontwerp en -analyse als onderdeel van de algehele softwarebeveiliging om de invoerverwerking te beschermen.
- **Modelvalidatie** Valideer ML-modellen regelmatig om ervoor te zorgen dat er niet mee is geknoeid vóór de implementatie, waardoor de nauwkeurigheid en integriteit ervan worden gewaarborgd.
- **Promptfiltering, rangschikking en responsvalidatie:** Analyseer en rangschik prompts om ervoor te zorgen dat alleen veilige invoer wordt verwerkt. Valideer antwoorden om misbruik te voorkomen.
- **Robuustheidscontroles:** Voer regelmatig evaluaties uit om kwetsbaarheden op te sporen en weg te nemen, zodat de AI veilig en betrouwbaar blijft.

Terug naar architectuurdiagram



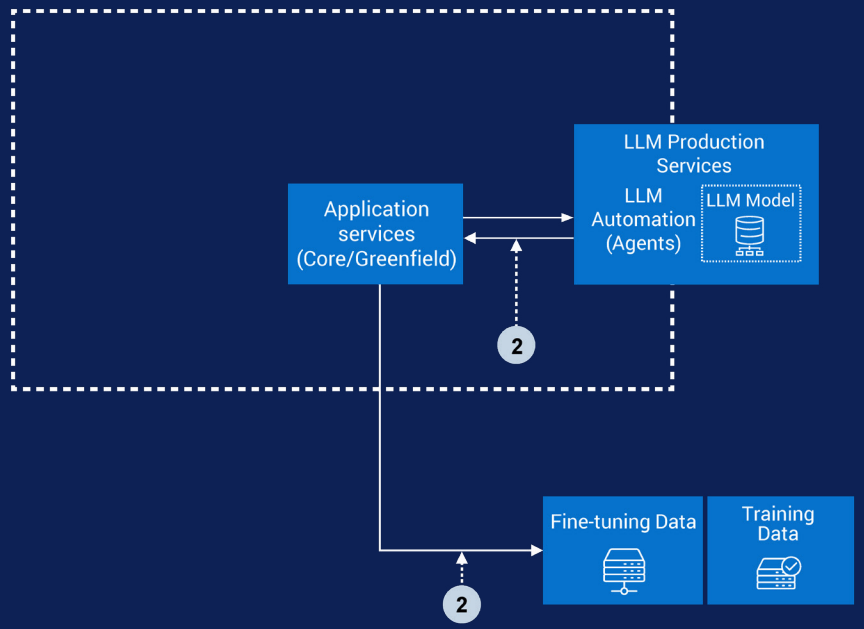
Promptinjectie is een opkomende uitdaging in de wereld van generatieve AI (GenAI), waarbij schadelijke invoer wordt gebruikt om het gedrag van het model te manipuleren of de integriteit ervan in gevaar te brengen. Deze aanvallen maken misbruik van kwetsbaarheden in de manier waarop AI-systemen gebruikersinvoer verwerken en daarop reageren, wat kan leiden tot ongeautoriseerde acties, onjuiste informatie of het uitlekken van gevoelige data. Naarmate GenAI steeds vaker wordt geïntegreerd in kritieke bedrijfsworkflows, is het aanpakken van deze risico's essentieel voor het behoud van vertrouwen en beveiliging

Tweede zorg: openbaarmaking van gevoelige informatie

Strategieën om openbaarmaking van gevoelige informatie te beperken:

- **Dataopschoning en invoervalidatie:** Controleer gebruikersinvoer grondig om schadelijke content te verwijderen. Gebruik normalisatie en codering om misbruik te voorkomen.
- **Gebruik homomorfische versleuteling** om gevoelige data veilig te verwerken zonder de inhoud ervan bloot te leggen. Zo zorgt u dat data versleuteld en beschermd blijven tegen inbreuken, ook wanneer ze in gebruik zijn.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Maak gebruik van beveiligde API's en systeeminterfaces** voor AI-data-interacties, waarbij u routinematig configuraties controleert om de blootstelling en het aanvalsoppervlak te minimaliseren.
- **Beveilig de dataverzameling, de storage en het databeleid** en implementeer uitgebreid beleid voor databescherming en -governance dat naleving van regelgeving garandeert en datarisico's minimaliseert.
- **Bewaking, logboekregistratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Veilige ontwikkeling, configuratie en audits:** Pas veilige coderingspraktijken toe, gebruik geautomatiseerde configuratiebeheertools en voer regelmatig evaluaties, audits en updates uit om AI-systeemconfiguraties veilig en actueel te houden.
- **Gebruikerstraining en beveiligingsbewustzijn:** Bied gebruikers en beheerders doorlopende, AI-specifieke training rond beveiligingsbewustzijn om onveilig gebruik en onbedoelde openbaarmaking van data te verminderen.

Terug naar architectuurdiagram



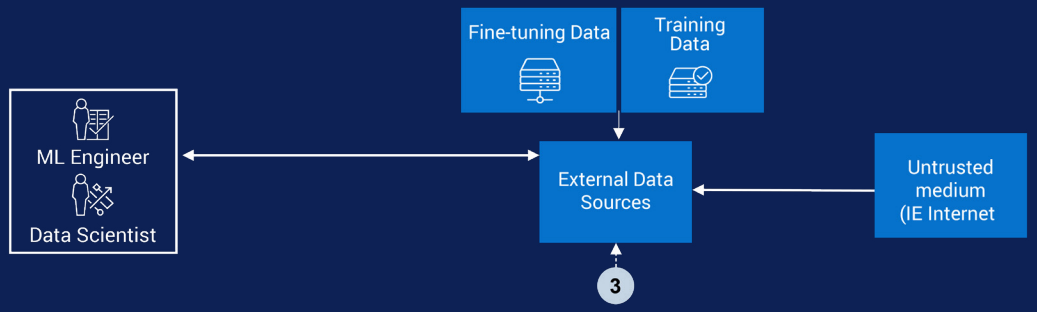
GenAI heeft ongelooflijke vooruitgang mogelijk gemaakt, maar brengt ook aanzienlijke risico's met zich mee, met name onbedoelde blootstelling van gevoelige informatie. Of het nu gaat om persoonlijk identificeerbare informatie (PII) of bedrijfseigen data, misbruik of verkeerd gebruik van GenAI-tools kan leiden tot datalekken, niet-naleving van wetgeving of reputatieschade. Hierdoor is het van cruciaal belang dat organisaties deze risico's begrijpen en proactief aanpakken om een veilige implementatie en veilig gebruik van AI-systemen te garanderen.

Derde zorg: kwetsbaarheden in de leveringsketen

Strategieën om kwetsbaarheden in de leveringsketen te beperken:

- **Controleer leveranciers en zorg voor naleving van veilige praktijken in de leveringsketen** Evalueer leveranciers en sluit overeenkomsten die prioriteit geven aan de beveiliging van de leveringsketen.
- **Implementeer een stuklijst voor software (SBOM)** Volg en verifieer de oorsprong van softwarecomponenten om transparantie te waarborgen en het risico op gecompromitteerde code te beperken.
- **Modelvalidatie** Valideer ML-modellen regelmatig om ervoor te zorgen dat er niet mee is geknoeid vóór de implementatie, wat de nauwkeurigheid en integriteit ervan waarborgt.
- **Voer containers en pods uit met de minste bevoegdheden** Zo vermindert u de mogelijke gevolgen in geval van een inbreuk en beperkt u onbevoegde toegang.
- **Implementeer firewalls** Blokkeer onnodige netwerkconnectiviteit om de blootstelling aan potentiële bedreigingen te verminderen en de mogelijkheden voor aanvallers te beperken.
- **Bescherm data en annotaties** Beveilig uw data en bijbehorende annotaties om sabotage, onbevoegde toegang en beschadiging van essentiële informatie te voorkomen.
- **Veilige hardware** Gebruik hardware waarvan de beveiliging is gevalideerd. Zo voorkomt u kwetsbaarheden die kunnen ontstaan door hardwareaanvallen en garandeert u een sterke basis voor uw infrastructuur.
- **Beveilig ML-softwarecomponenten** Gebruik vertrouwde en gescreende ML-softwarecomponenten om kwetsbaarheden te verminderen en de algehele beveiliging van uw machine learning-workflows te verbeteren.
- **Veilige ontwikkeling, configuratie en audits:** Pas veilige coderingspraktijken toe, gebruik geautomatiseerde configuratiebeheertools en voer regelmatig evaluaties, audits en updates uit om AI-systeemconfiguraties veilig en actueel te houden.

Terug naar architectuurdiagram



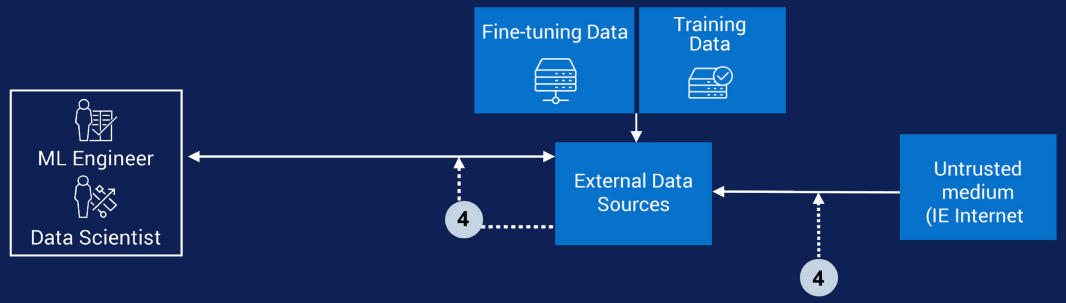
Ontdek kwetsbaarheden in de LLM-leveringsketen die van invloed kunnen zijn op kritieke componenten, zoals vooraf getrainde modelintegriteit en adapters van derden. AI-systemen zijn afhankelijk van hardware en software die al lang voor de implementatie gecompromitteerd kunnen zijn. Aanvallers kunnen zwakke punten in verschillende stadia van de leveringsketen voor machine learning uitbuiten, waarbij ze zich richten op GPU-hardware, data en annotaties, elementen van de ML-softwarestack of zelfs op het model zelf. Door deze unieke onderdelen te saboteren kunnen aanvallers initiële toegang krijgen tot systemen, wat aanzienlijke risico's voor de beveiliging en integriteit oplevert. Voor de ontwikkeling van robuuste, veilige AI-oplossingen is het cruciaal om deze kwetsbaarheden te begrijpen en op te lossen.

Vierde zorg: vergiftiging van modeldata

Strategieën om vergiftiging van modeldata te beperken:

- **Pas tijdens de training anomaliedetectie en datavalidatie toe** om inconsistenties in data op te sporen en aan te pakken, zodat alleen schone, hoogwaardige data worden gebruikt om het model te trainen.
- **Isoleer omgevingen tijdens fijnafstemmingsfasen** om onbevoegde toegang of vervuiling van het model tijdens kritieke ontwikkelingsfasen te voorkomen.
- **Modelvalidatie** Valideer ML-modellen regelmatig om ervoor te zorgen dat er niet mee is geknoeid vóór de implementatie, wat de nauwkeurigheid en integriteit ervan waarborgt.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Dataopschoning en invoervalidatie:** Controleer gebruikersinvoer grondig om schadelijke content te verwijderen. Gebruik normalisatie en codering om misbruik te voorkomen.
- **Veilige ontwikkeling, configuratie en audits:** Pas veilige coderingspraktijken toe, gebruik geautomatiseerde configuratiebeheertools en voer regelmatig evaluaties, audits en updates uit om AI-systeemconfiguraties veilig en actueel te houden.
- **Robuustheidscontroles:** Voer regelmatig evaluaties uit om kwetsbaarheden op te sporen en weg te nemen, zodat de AI veilig en betrouwbaar blijft.
- **Implementeer netwerksegmentatie** om de toegang tot onveilige interfaces en kritieke systeemcomponenten te beperken.
- **Bewaking, registratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.

Terug naar architectuurdiagram



Datavergiftiging van het Explore Model is een beveiligingsbedreiging in de AI-levenscyclus, waarbij aanvallers trainingsdata opzettelijk vervuilen met corrupte, misleidende of schadelijke invoer. Dit risico kan van invloed zijn op kritieke componenten van het verzamelen en annoteren van onbewerkte data, tot het samenstellen en integreren van datasets die worden gebruikt voor machine learning of grote taalmodellen. De betrouwbaarheid van AI-systemen hangt af van de integriteit van de bijbehorende databronnen. Deze kunnen voorafgaand aan de training, tijdens de voorverwerking of via externe datapijplijnen aan manipulatie worden blootgesteld.

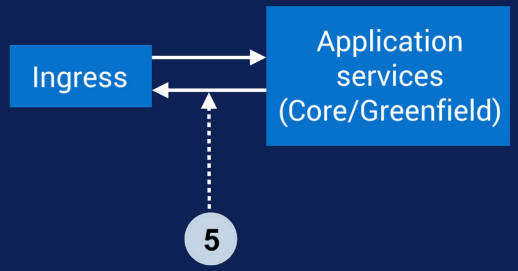
Aanvallers maken gebruik van datavergiftiging om de nauwkeurigheid van het model te verlagen, kwetsbaarheden te introduceren of schadelijke uitvoer te veroorzaken. Door zwakke punten in de processen voor dataherkomst, annotatiekwaliteit of opname van datasets aan te pakken, kunnen aanvallers de beveiliging, betrouwbaarheid en veerkracht ondermijnen. Het herkennen en beperken van deze datagerichte bedreigingen is essentieel om robuuste, betrouwbare AI-oplossingen te kunnen ontwikkelen.

Vijfde zorg: onjuiste verwerking van uitvoer

Strategieën om een onjuiste verwerking van uitvoer te beperken:

- **Contextbewuste uitvoercodering:** Pas altijd coderings- en ontsnappingstechnieken toe die zijn afgestemd op de specifieke context waarin de uitvoer wordt gebruikt, zoals HTML-, SQL- of API-omgevingen, om kwetsbaarheden zoals injectieaanvallen te voorkomen.
- **Uitvoeropschoning:** Volg strikte validatie- en opschoningspraktijken voor modeluitvoer in overeenstemming met de richtlijnen van de Application Security Verification Standard (ASVS) van het Open Web Application Security Project (OWASP) om veilig downstreamgebruik te garanderen en beveiligingsrisico's te beperken.
- **Bewaking, logboekregistratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Geautomatiseerde beveiligingstests voor de uitvoer:** Voer regelmatig beveiligingstests uit met behulp van geautomatiseerde tools om risico's in de uitvoer te identificeren, zoals cross-site scripting (XSS) of kwetsbaarheden in injecties. Pak deze proactief aan.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Human-in-the-loop-beoordeling:** Toepassingen met een verhoogd risico, zoals financiën of gezondheidszorg, vereisen menselijke bewaking en beoordeling van de modeluitvoer om nauwkeurigheid, beveiliging en veiligheid te garanderen.
- **Privacy en naleving:** Integreer technieken voor privacybescherming in het uitvoerproces en voldoe aan relevante regelgeving en normen voor het veilig gebruik van gevoelige informatie.

Terug naar architectuurdiagram



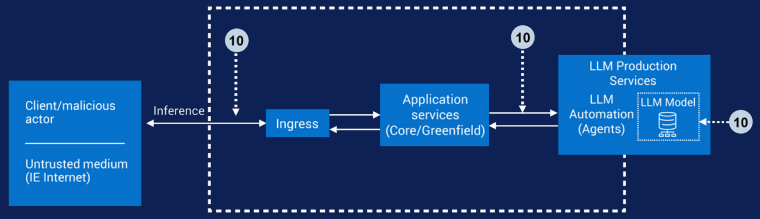
Onvoldoende validatie of opschoning van AI-modeluitvoer kan leiden tot ernstige beveiligingsrisico's, waaronder escalatie van bevoegdheden en datalekken. Wanneer AI-modellen uitvoer produceren die niet goed wordt gecontroleerd of gefilterd, kunnen aanvallers deze kwetsbaarheden uitbuiten om onbevoegde toegang te krijgen of hun bevoegdheden binnen een systeem uit te breiden. Dit gebrek aan toezicht kan leiden tot gecompromitteerde data, ongeautoriseerde acties en aanzienlijke beveiligingslekken. Dit benadrukt hoe belangrijk het is om robuuste validatie- en opschoningsprocessen te implementeren voor door AI gegenereerde uitvoer.

Zesde zorg: buitensporige keuzevrijheid

Strategieën om buitensporige keuzevrijheid te beperken

- **Dwing minimale bevoegdheden af:** Verleen LLM's en agentische subsystemen alleen de minimale machtigingen die nodig zijn om bedoelde bewerkingen uit te voeren en controleer regelmatig toegangscontroles.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Stel operationele grenzen in:** Definieer duidelijk welke toegang LLM's of agents mogen hebben en welke acties ze mogen uitvoeren.
- **Human-in-the-loop-beoordeling:** Toepassingen met een verhoogd risico, zoals financiën of gezondheidszorg, vereisen menselijke bewaking en beoordeling van de modeluitvoer om de nauwkeurigheid, beveiliging en veiligheid te garanderen.
- **Bewaking, registratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Beperk de autonomie:** Beperk de mogelijkheden van LLM's om onbeperkte toegang of controle te voorkomen.
- **Veilige ontwikkeling, configuratie en audits:** Pas veilige coderingspraktijken toe, gebruik geautomatiseerde configuratiebeheertools en voer regelmatig evaluaties, audits en updates uit om AI-systeemconfiguraties veilig en actueel te houden.
- **Implementeer firewalls** Blokkeer onnodige netwerkconnectiviteit om de blootstelling aan potentiële bedreigingen te verminderen en de mogelijkheden voor aanvallers te beperken.
- **Robuustheidscontroles:** Voer regelmatig evaluaties uit om kwetsbaarheden op te sporen en weg te nemen, zodat de AI veilig en betrouwbaar blijft.

Terug naar architectuurdiagram



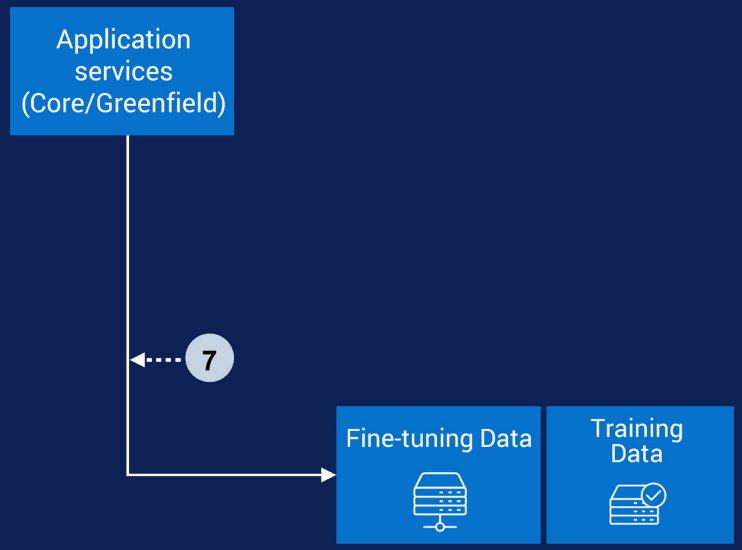
Als AI-agents of plug-ins buitensporige autonomie of onnodige functionaliteit krijgen in workflows, kan dat aanzienlijke risico's met zich meebrengen. Als een AI-systeem bevoegdheden of mogelijkheden krijgt die verder gaan dan wat nodig is, verhoogt dit de kans op onbedoelde gevolgen. Dit kan gebeuren wanneer systemen op basis van een Large Language Model (LLM) zijn ontworpen met overmatige machtigingen, waardoor ze acties kunnen ondernemen of toegang kunnen krijgen tot informatie die niet voor hen is bedoeld. Dit kan leiden tot fouten, misbruik van data of zelfs beveiligingslekken, wat het belang benadrukt van het zorgvuldig beperken en bewaken van AI-mogelijkheden om veilig en verantwoord gebruik te garanderen.

Zevende zorg: promptlekkage

Strategieën om promptlekkage te voorkomen

- **Neem geen gevoelige informatie op in prompts** Vermijd referenties, API-sleutels of bedrijfseigen logica in prompts. Beheer deze veilig buiten het systeem.
- **Scheid beveiligingscontroles van prompts** Voer authenticatie, autorisatie en sessiebeheeruit in applicatielogica, niet in prompts.
- **Valideer invoer en uitvoer** Schoon prompts en reacties op met robuuste validatie om verdachte patronen of manipulaties te blokkeren.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Versleutel en beveilig prompts** Sla prompts en configuraties op in versleutelde, veilige storage om onbevoegde toegang te voorkomen.
- **Bewaking, registratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Beoordeel prompts regelmatig** Controleer prompts regelmatig en schoon ze op om gevoelige data te verwijderen en naleving van beveiligingsregels te garanderen.
- **Test op zwakheden en schakel een red team in** Voer vijandige tests uit om kwetsbaarheden in promptbeheer of uitvoer te identificeren en op te lossen.
- **Isoleer prompts van gebruikersinvoer** Ontwerp systemen om te voorkomen dat gebruikersquery's prompts manipuleren of blootleggen.
- **Dwing snelheidslimieten af** Beperk API-gebruik, verminder verdachte activiteiten en blokkeer geautomatiseerde promptaanvallen.

Terug naar architectuurdiagram



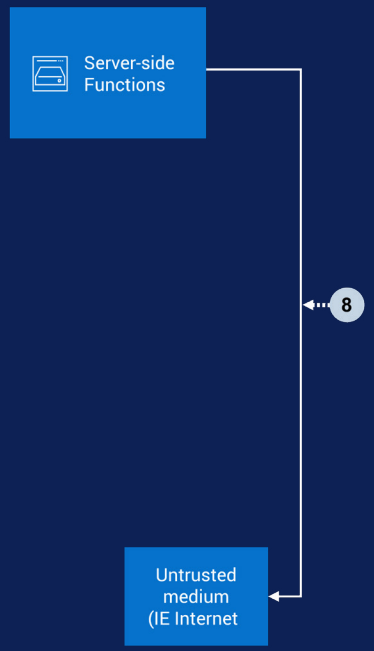
Een promptlekaanval op een groot taalmodel (LLM) of AI-systeem vindt plaats wanneer een aanvaller in staat is de verborgen instructies (de 'systeemprompts') te extraheren of af te leiden die het gedrag van het model sturen en operationele grenzen instellen. Het is meestal niet de bedoeling dat eindgebruikers prompts te zien krijgen, omdat deze basisregels, beperkingen en soms ook gevoelige operationele logica bevatten. Via speciaal samengestelde invoer of het misbruiken van beveiligingslekken kan een aanvaller de LLM misleiden om de systeemprompt geheel of gedeeltelijk weer te geven. Als deze informatie wordt gelekt, kan die worden gebruikt om restricties te reverse-engineeren, veiligheidsfilters te omzeilen of nieuwe gerichte aanvallen te ontwikkelen, waardoor uiteindelijk het risico toeneemt op directe injectie, escalatie van bevoegdheden of misbruik van het model en downstreamsystemen die afhankelijk zijn van de integriteit ervan.

Achtste zorg: zwakke punten in vectoren en inbeddingen

Strategieën om zwakke punten in vectoren en inbeddingen te beperken

- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Versleuteling:** Beveilig vectordata tijdens overdracht en in rust met behulp van robuuste versleutelingsstandaarden zoals AES.
- **Veilige configuratie en bewaking:** Versterk systemen, gebruik veilige configuraties en controleer continu op onjuiste configuraties, onbevoegde toegang of afwijkingen.
- **Kwetsbaarheidsbeheer:** Werk regelmatig alle software, afhankelijkheden en vectoropslagengines bij en patch deze om beveiligingsrisico's aan te pakken.
- **Dataopschoning en invoervalidatie:** Controleer gebruikersinvoer grondig om schadelijke content te verwijderen. Gebruik normalisatie en codering om misbruik te voorkomen.
- **Maak gebruik van beveiligde API's en systeeminterfaces** voor AI-data-interacties, waarbij u routinematig configuraties controleert om de blootstelling en het aanvalsoppervlak te minimaliseren.
- **Bewaking, logboekregistratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Veilige hardware** Gebruik hardware waarvan de beveiliging is gevalideerd. Zo voorkomt u kwetsbaarheden die kunnen ontstaan door hardwareaanvallen en garandeert u een sterke basis voor uw infrastructuur.
- **Veilige ontwikkeling, configuratie en audits:** Pas veilige coderingspraktijken toe, gebruik geautomatiseerde configuratiebeheertools en voer regelmatig evaluaties, audits en updates uit om AI-systeemconfiguraties veilig en actueel te houden.

Terug naar architectuurdiagram



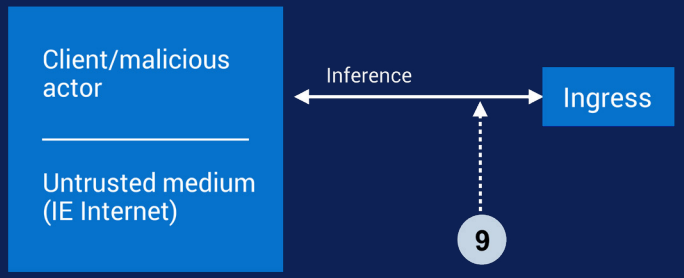
Een aanval op zwakke punten in vectoren en inbeddingen bij een groot taalmodel (LLM) of AI-systeem, met name modellen of systemen die Retrieval Augmented Generation (RAG) gebruiken, richt zich op kwetsbaarheden in de manier waarop informatie wordt gecodeerd, opgeslagen en opgehaald als numerieke vectoren en inbeddingen. Zwakke punten in deze mechanismen kunnen worden misbruikt via schadelijke acties, zoals inversie-inbedding (het reconstrueren van gevoelige data uit inbeddingen), datavergiftiging (het injecteren van schadelijke of bevooroordeelde content om modelgedrag te manipuleren), ongeautoriseerde toegang tot vectordatabases (wat leidt tot datalekken) of manipulatie van ophaaluitvoer. Deze aanvallen zijn een gevaar voor de privacy, integriteit en betrouwbaarheid, omdat aanvallers gevoelige informatie openbaar kunnen maken, uitvoer kunnen wijzigen of het vertrouwen van gebruikers in AI-gestuurde applicaties kunnen ondermijnen. Degelijke toegangscontroles, datavalidatie, versleuteling en doorlopende bewaking zijn van cruciaal belang om deze veranderende bedreigingen te bestrijden.

Negende zorg: onjuiste informatie

Strategieën om onjuiste informatie te beperken

- **Retrieval-Augmented Generation (RAG) met betrouwbare bronnen:** Gebruik RAG om informatie op te halen en te integreren uit geverifieerde, vertrouwde databases en kennisopslagplaatsen, waardoor hallucinaties worden verminderd.
- **Modelafstemming en uitvoerkalibratie:** Verfijn modellen met diverse datasets en pas technieken toe om vooroordelen en onjuiste informatie te minimaliseren.
- **Geautomatiseerde feitencontrole:** Vergelijk uitvoer met betrouwbare bronnen en markeer valse informatie automatisch.
- **Onzekerheidsbewaking:** Markeer reacties met weinig vertrouwen voor menselijke beoordeling in kritieke gevallen.
- **Human-in-the-loop-beoordeling:** Toepassingen met een verhoogd risico, zoals financiën of gezondheidszorg, vereisen menselijke bewaking en beoordeling van de modeluitvoer om nauwkeurigheid, beveiliging en veiligheid te garanderen.
- **Feedback van gebruikers**: Stel gebruikers in staat om fouten te melden, zodat u modellen continu kunt verbeteren en paden die tot onjuiste informatie leiden snel kunt corrigeren.
- **Toegangsbeperkingen en menselijk toezicht:** Pas op rollen gebaseerde toegangscontrole (RBAC), meervoudige verificatie (MFA) en identiteitsbeheer toe om de toegang te beperken. Gebruik menselijke beoordeling voor kritieke beslissingen.
- **Veilige ontwikkeling, configuratie en audits:** Pas veilige coderingspraktijken toe, gebruik geautomatiseerde configuratiebeheertools en voer regelmatig evaluaties, audits en updates uit om AI-systeemconfiguraties veilig en actueel te houden.
- **Communicatie over de risico's:** Informeer gebruikers over de beperkingen van AI en moedig onafhankelijke controle aan.
- **Doelbewust UI- en API-ontwerp:** Markeer door AI gegenereerde content en help gebruikers deze verantwoord te gebruiken.

Terug naar architectuurdiagram



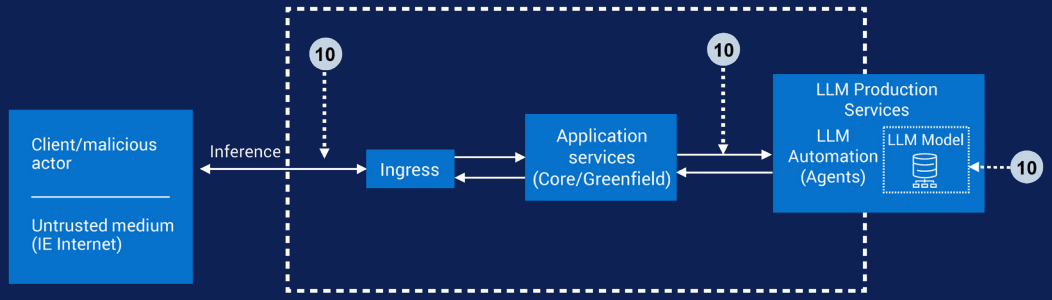
Een desinformatieaanval op een LLM- of AI-systeem is een opzettelijke poging om het model valse, misleidende of schijnbaar geloofwaardige, maar onjuiste, informatie te laten genereren of verspreiden via de uitvoer. Deze kwetsbaarheid is het gevolg van verschillende factoren: de neiging van het model om te 'hallucineren' (het genereren van gefabriceerde maar plausibel klinkende content), vooroordelen of hiaten in de trainingsdata en de invloed van vijandige prompts. Hallucinaties doen zich voor omdat LLM's statistisch tekst genereren die past bij een patroon, in plaats van feiten echt te begrijpen. Dit leidt tot antwoorden die betrouwbaar lijken, maar eigenlijk ongegrond zijn. De risico's van zulke aanvallen omvatten onder meer beveiligingslekken, reputatieschade en zelfs juridische aansprakelijkheid. Dit geldt in het bijzonder wanneer gebruikers te veel vertrouwen op LLM-reacties zonder de nauwkeurigheid of geldigheid daarvan te verifiëren, waardoor fouten of onjuiste informatie mogelijk deel worden van kritieke beslissingen en processen.

Tiende zorg: onbeperkte consumptie

Strategieën tegen onbeperkte consumptie

- **Dwing snelheidsbeperkingen en gebruikersquota af**
  Stel strikte limieten in voor aanvragen, tokens of data per gebruiker, API-sleutel of app om misbruik te voorkomen.
- **Vereis verificatie en gebruikerssegmentatie** Gebruik sterke verificatie (bijv. API-sleutels, OAuth) en wijs rollen of niveaus toe om alleen geautoriseerde aanvragen te verwerken.
- **Invoervalidatie en groottebeperkingen** Valideer de grootte en structuur van prompts, waardoor grote of verkeerd ingedeelde query's worden geblokkeerd of ingekort.
- **Pas verwerkingstime-outs en resourcebeperking toe** Stel voor elke aanvraag time-outs en resourcelimieten in om langdurige activiteiten en resourceverbruik te voorkomen.
- **Implementeer slimme caching en deduplicatie** Sla reacties voor dubbele of vergelijkbare query's op om onnodige verwerking te verminderen.
- **Bewaking, registratie en detectie van afwijkingen:** Bewaak en registreer AI-systeemactiviteiten voortdurend met behulp van oplossingen zoals MDR/XDR/SIEM, om onbevoegde toegang, afwijkingen of datalekken snel te detecteren, te onderzoeken en erop te reageren.
- **Budgetbewaking en uitgavencontroles** Gebruik dashboards en waarschuwingen om de kosten te bewaken en gebruik te blokkeren als budgetdrempels worden bereikt.
- **Sandboxing en isolatietechnieken** Voer workloads uit in geïsoleerde omgevingen met beperkte machtigingen om risico's te verminderen.
- **Beperk de gespreksdiepte en gespreksbeurten** Beperk recursieve gesprekken of gesprksstappen om misbruik te voorkomen.
- **Pas gelaagde toewijzing van modellen of resources toe**
  Leid aanvragen met hoge prioriteit door naar premium modellen en verkeer met lage prioriteit naar kosteneffectieve modellen.

Terug naar architectuurdiagram



Een onbeperkte consumptiedreiging op een LLM- of AI-systeem verwijst naar een beveiligingsprobleem waarbij de applicatie gebruikers in staat stelt om (kwaadaardig of anderszins) buitensporige, ongecontroleerde inferentieaanvragen of prompts in te dienen zonder effectieve snelheidsbeperking, verificatie of gebruiksbeperkingen. Omdat de LLM-inferentie veel rekenkracht kost, kan dit gebrek aan controle op verschillende manieren worden misbruikt: aanvallers kunnen denial of service (DOS) veroorzaken door systeembronnen te overweldigen, onvoorziene economische verliezen veroorzaken bij pay-per-use- of cloud-gehoste implementaties, of systematisch het model vragen om zijn gedrag te klonen en intellectueel eigendom te stelen. Dit kan onder andere leiden tot verstoring van de service, verminderde prestaties voor andere gebruikers, financiële druk en een verhoogd risico op het weglekken van gevoelige modellen. In wezen vindt onbeperkte consumptie plaats wanneer het resourceverbruik niet goed wordt geregeld, waardoor op LLM gebaseerde applicaties zowel onbedoeld als opzettelijk worden misbruikt.

Waarom zou u Dell kiezen voor AI-beveiliging?

Dell helpt organisaties AI-modellen en LLM's te beveiligen via een uitgebreide aanpak die hardware, software en beheerde services omvat. Beveiliging is geïntegreerd van de leveringsketen tot apparaat, infrastructuur, data en applicaties, allemaal afgestemd op Zero Trust-principes. Alle oplossingen in het portfolio van Dell zijn gebouwd om cyberhygiëne te verbeteren met functies zoals MFA, RBAC, minimale bevoegdheden en continue verificatie. Deze uitgebreide, 'secure by design'-aanpak zorgt ervoor dat organisaties vol vertrouwen kunnen innoveren met AI en LLM's, waarbij het risico op modeldiefstal, datalekken, vijandige aanvallen en andere geavanceerde cyberdreigingen wordt geminimaliseerd.

Leveringsketen

De beveiligde leveringsketen van Dell biedt fundamentele bescherming voor AI-modellen en LLM's door beveiliging te integreren in elke fase van productontwikkeling, productie en levering. Door middel van cryptografisch ondertekende BIOS- en firmware-updates, Secured Component Verification, AI-gerichte stuklijsten voor software (SBOM), het volgen van datasetlijnen, geïntegreerde beveiligingssoftware en -configuratie en strenge risicobeoordelingen van leveranciers in overeenstemming met wereldwijde normen, minimaliseert Dell risico's op sabotage, onbevoegde toegang en aanvallen op de leveringsketen. Zo kunnen organisaties betrouwbare, veerkrachtige AI-workloads implementeren met volledige transparantie, integriteit en naleving van regelgeving.

AI PC's

Dell biedt fundamentele beveiliging voor AI-workloads op apparaten. Dell Trusted Devices, de veiligste zakelijke AI PC's ter wereld*, zijn ontworpen met het oog op beveiliging. Beveiliging van de leveringsketen minimaliseert het risico op kwetsbaarheden en sabotage van producten. De unieke verdediging die rechtstreeks in hardware en firmware is ingebouwd, houdt de pc en eindgebruiker beschermd tijdens het gebruik. Dell SafeBIOS biedt diepgaande zichtbaarheid op BIOS-niveau en detectie van sabotage, terwijl Dell SafeID de beveiliging van referenties verbetert en verificatie zonder wachtwoord mogelijk maakt. Partnersoftware biedt geavanceerde bescherming voor eindpunten-, netwerk- en cloudomgevingen.

Cybertolerantie

De PowerProtect cybertolerantieoplossingen van Dell beveiligen AI-data met versleutelde, onveranderlijke back-ups, snel herstel en geïsoleerde kluizen voor cyberherstel. Deze mogelijkheden voorkomen vernietiging, beperken de impact van schadelijke updates en ondersteunen naleving en herstel na een aanval.

Servers

PowerEdge servers zijn voorzien van vertrouwelijke computing om AI-/LLM-prompts en -inbeddingen te isoleren en te beveiligen. Daarnaast omvatten ze vertrouwde RAG-oplossingen (Retrieval-Augmented Generation) die zijn gebaseerd op betrouwbare bronnen, plus MFA, RBAC, Silicon Root of Trust, ondertekende firmware en continue bewaking om kritieke AI-workloads te beschermen.

Storage

Het storageportfolio van Dell zorgt voor veilige, versleutelde storage voor gevoelige AI-data met robuuste AES-256-versleuteling voor data in rust en tijdens overdracht. Bij selecteerde aanbiedingen is geavanceerde versleuteling beschikbaar die is ontworpen om toekomstige kwantumbedreigingen te kunnen weerstaan. Het portfolio bevat snelle NVMe-prestaties, FIPS-compatibele versleutelingsmodules om data te beveiligen (waaronder data die wordt gebruikt in AI-workloads), onveranderlijke snapshots en air-gapped cyberherstelkluizen om ransomware-aanvallen tegen te gaan. Zero Trust-architectuur, beveiliging van de leveringsketen en sabotagebestendige auditmogelijkheden verbeteren de governance. Geïntegreerde anomaliedetectie en AIOps ML-modellen beschermen workloads zonder klantdata te gebruiken voor training, waardoor het risico op aanvallen op basis van invoer wordt geminimaliseerd.

AIOps

Dell AIOps biedt geautomatiseerde, continue bewaking om onjuiste configuraties en kwetsbaarheden (inclusief CVE's) te detecteren, en ondersteunt risicobewustzijn in de leveringsketen die van invloed is op AI/LLM-workloads. Realtime CVE-scans, slimme waarschuwingen en AI-gestuurde dashboards maken snel ingrijpen mogelijk door afwijkingen te markeren en resolutieworkflows te volgen. Geïntegreerde nalevingsfuncties, op rollen gebaseerde toegangscontroles en geautomatiseerde rapportage helpen ervoor te zorgen dat activiteiten in verschillende workloads veilig blijven, terwijl naadloze EDR/XDR-integratie en AI-gestuurde operationele inzichten, waaronder generatieve mogelijkheden in ondersteunde oplossingen, de IT-efficiëntie verder verbeteren.

Networking

Dell Networking oplossingen beschermen AI/LLM-omgevingen door middel van robuuste netwerksegmentatie, waardoor laterale verplaatsing wordt geminimaliseerd. Versleutelde netwerkpaden en geïntegreerde firewallcontroles blokkeren onbevoegde toegang tot AI-data.

Services voor AI-beveiliging en -veerkracht

De AI-beveiligings- en veerkrachtservices van Dell zijn ontworpen om nieuwe risico's aan te pakken die gepaard gaan met de integratie van AI in uw organisatie. Onze services zijn ontworpen om samen te werken met uw teams terwijl u AI zo snel mogelijk integreert. Ze bieden expertise om u te begeleiden bij strategische planning, implementatie van oplossingen en beheerde beveiligingsservices om de operationele lasten te verlichten, zodat u veilig kunt innoveren met AI. Elk service wordt op maat gemaakt om organisaties te helpen opkomende AI-risico's aan te pakken en veilige AI-implementaties te optimaliseren.

Dell AI Factory

Een geïntegreerd portfolio van speciaal gebouwde beveiliging, zoals de beveiligde leveringsketen van Dell, Zero Trust-mogelijkheden om minimale bevoegdheden af te dwingen en AI MDR-oplossingen die zijn ontworpen om uw model veilig te houden.

*Gebaseerd op interne analyse van Dell, oktober 2024 (Intel) en maart 2025 (AMD). Van toepassing op pc's met Intel en AMD-processors. Niet alle functies zijn beschikbaar voor alle pc's. Extra aankoop vereist voor sommige functies. Pc's op basis van Intel zijn gevalideerd door Principled Technologies, juli 2025 .

Conclusie

Om veerkrachtige AI-frameworks op te bouwen, is een gezamenlijke aanpak tussen organisaties en beveiligingsexperts essentieel. Aangezien AI en LLM's industrieën steeds verder hervormen, is het van cruciaal belang om de risico's aan te pakken die deze technologieën met zich meebrengen, waaronder uitdagingen op het gebied van databeveiliging, modelintegriteit en naleving. Organisaties moeten prioriteit geven aan proactieve strategieën die beveiliging integreren in elke fase van hun AI-traject.

Dell Technologies steunt deze missie als een vertrouwde partner en biedt end-to-end GenAI-aanpassing, beveiligingsadvies en geïntegreerde oplossingen die zijn afgestemd op uw unieke behoeften. Door gebruik te maken van de robuuste cyberbeveiligingsoplossingen van Dell, kunnen ondernemingen AI- en LLM-risico's effectief beperken en tegelijkertijd het potentieel van hun bestaande beveiligingsinvesteringen maximaliseren. Dell stelt organisaties in staat hun AI-infrastructuur te beschermen door geavanceerde beveiliging naadloos te integreren in hun huidige frameworks, waardoor een toekomstbestendige, veilige omgeving wordt gegarandeerd.

Ontdek hoe de uitgebreide AI-oplossingen van Dell uw GenAI- en LLM-omgevingen kunnen beveiligen: Dell.com/CyberSecurityMonth