DELL Technologies

De tio främsta cybersäkerhetsproblemen för GenAI och stora språkmodeller (LLM:er)

Kom i gång

Introduktion

Artificiell intelligens (AI) håller på att revolutionera det sätt på vilket organisationer arbetar, och generativ AI (GenAI) och stora språkmodeller (Large Language Models, LLM) blir till kritiska arbetsbelastningar i moderna företagsmiljöer.

Precis som alla andra arbetsbelastningar har dessa program sina egna komplexiteter och sårbarheter som behöver hanteras. I takt med att allt fler verksamheter inför AI för att förbättra innovationerna, effektiviteten och konkurrensfördelarna blir det helt nödvändigt att upprätthålla säkerheten för dessa program. God cyberhygien är grunden för att säkra alla arbetsbelastningar, och precis som säkerheten för samtliga arbetsbelastningar är prioriterad är det viktigt att tillämpa god cyberhygien även för AI. Detta omfattar implementeringsmetoder som korrekt systemkorrigering, multifaktorautentisering, rollbaserad åtkomst och nätverkssegmentering. Dessa åtgärder är grundläggande, men nyckeln ligger i att förstå hur dessa funktioner passar in i den aktuella arkitekturen och användningen av din arbetsbelastning.

Vid Dell har vi en god förståelse för AI-arbetsbelastningen och de unika säkerhetsutmaningar den medför. Genom att identifiera hur hotaktörer kan rikta in sig på dessa arbetsbelastningar kan Dell hjälpa dig att skapa en robust säkerhetsstrategi. Detta inkluderar att hantera risker som förgiftning av träningsdata, modellstöld och manipulering, rekonstruktion av datauppsättningar och mycket mer.

Vi fokuserar också på att hantera utmaningar som är förknippade med indata till din AI-modell, till exempel att förhindra utlämnande av känslig information, minska osäkra ämnen och missvisande information och säkerställa efterlevnad av regelverk. På resultatsidan hjälper vi till att hantera problem som överdrivet stort beroende av modellen och efterlevnadsrelaterade risker.

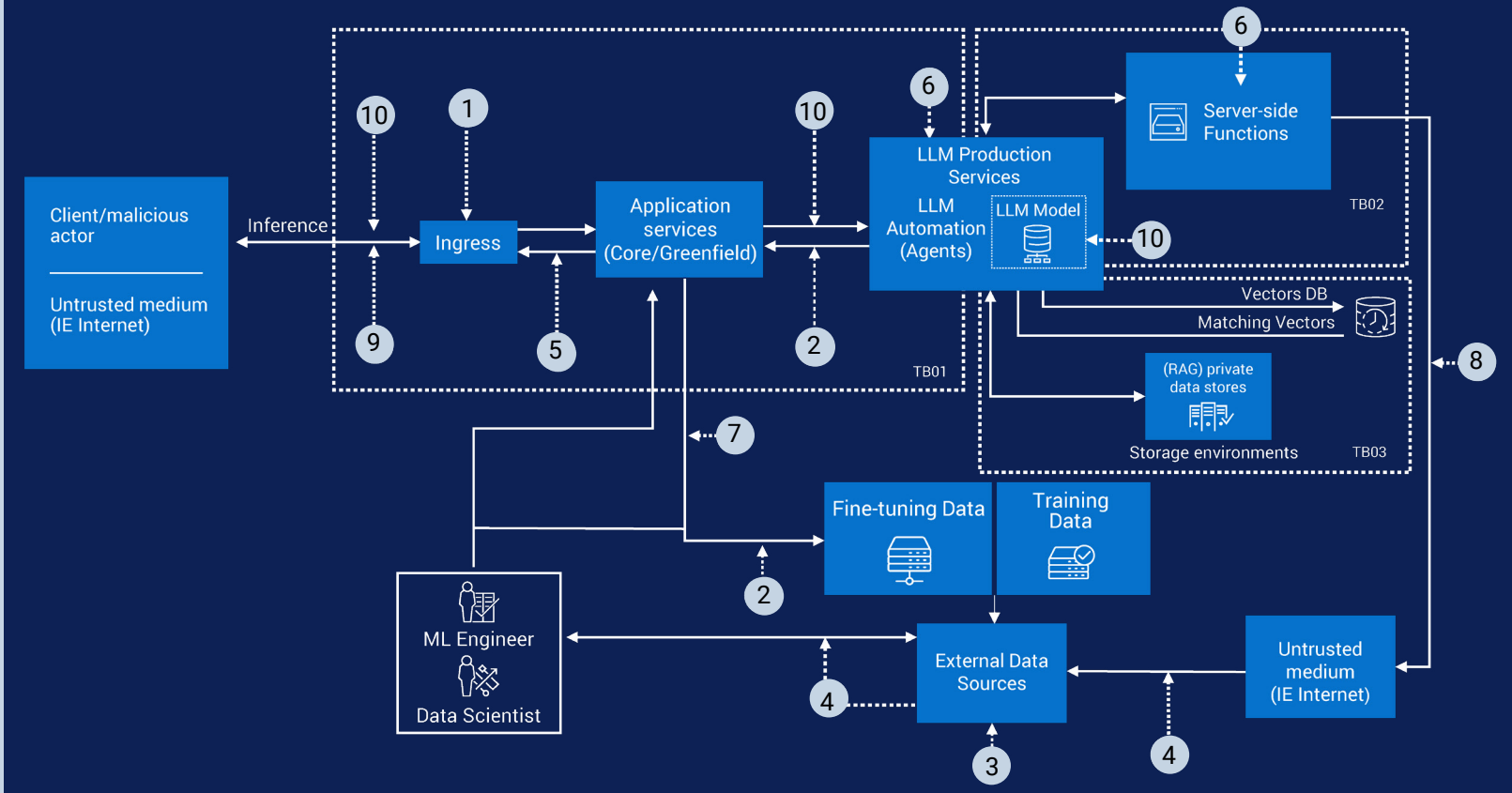
På Dell ger vi företag möjlighet att minska dessa risker genom att utnyttja befintliga cybersäkerhetslösningar eller utforska nya verktyg och metoder för att skydda systemen. Vårt mål är att säkerställa att säkerheten inte hindrar innovationer. Genom att förstå hur AI-arbetsbelastningar fungerar och de säkerhetshot de kan utsättas för kan vi hjälpa dig att bygga en starkare säkerhetsställning, vilket gör miljön mer motståndskraftig samtidigt som den inte förhindrar innovationer. Med våra expertkunskaper hjälper vi till med att tryggt utnyttja potentialen hos AI samtidigt som en robust säkerhet kan upprätthållas.

De tio största cybersäkerhetsproblemen för generativ AI och stora språkmodeller

Det här är de centrala problemen när det kommer till att skydda GenAI-/LLM-modeller, enligt OWASP.

Klicka på ett problem om du vill veta mer:

1. Promptinmatning
2. Avslöjande av känslig information
3. Leverantörskedja
4. Modelldataförgiftning
5. Felaktig hantering av utdata
6. Överdriven handlingsfrihet
7. Systempromptsläckage
8. Vektorsvagheter och inbäddade svagheter
9. Felaktig information
10. Obegränsad konsumtion

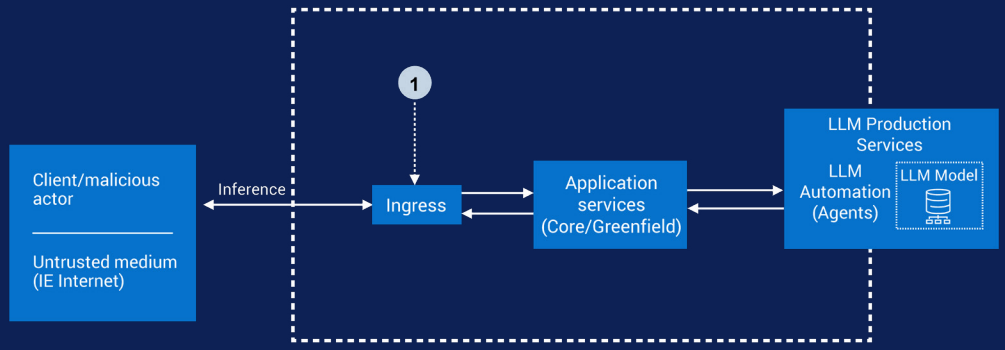


Problem nr 1: Promptinmatning

Strategier för att förhindra promptinmatning:

- **Datasanering och validering av indata:** Gör en grundlig genomsökning av användarindata för att ta bort skadligt innehåll. Använd normalisering och kodning för att förhindra missbruk.
- **Natural Language Processing (NLP) och maskininlärningsbaserade metoder:** Använd NLP och maskininlärning för att upptäcka och blockera manipulerade och skadliga prompter.
- **Rensa utdataformatering och svarskontroller:** Ställ in strikta svarsgränser för att säkerställa att utdata följer avsedda format och förhindra obehöriga åtgärder. Använd promptfiltrering och svarsvalidering för att upprätthålla integriteten.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Säker prompтteknik:** Använd säker promptutformning och analys som en del av den övergripande mjukvarusäkerheten för att skydda inmatningsbearbetningen.
- **Modellvalidering:** Validera ML-modeller regelbundet för att säkerställa att de inte har manipulerats före driftsättning, vilket skyddar deras noggrannhet och integritet.
- **Promptfiltrering, rangordning och svarsvalidering:** Analysera och rangordna prompter för att säkerställa att endast säkra inmatningar bearbetas. Validera svar för att förhindra felaktig användning.
- **Robusthetskontroller:** Genomför regelbundna utvärderingar för att identifiera och åtgärda sårbarheter och hålla AI:n säker och tillförlitlig.

Tillbaka till arkitekturdiagrammet



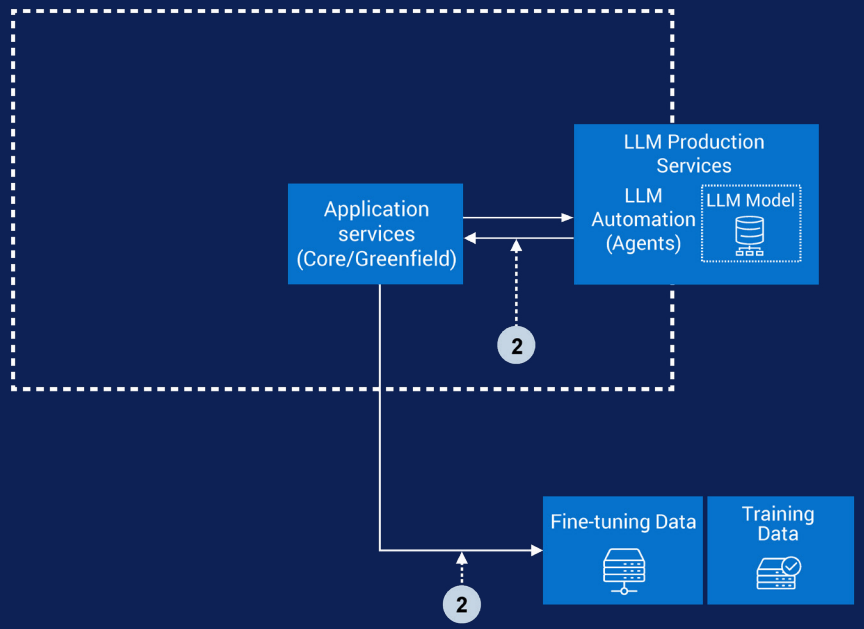
Promptinmatning är en växande utmaning inom generativ AI (GenAI), där skadliga inmatningar skapas för att manipulera modellens beteende eller äventyra dess integritet. Dessa attacker utnyttjar sårbarheter i hur AI-system bearbetar och reagerar på användarinmatningar, vilket kan leda till obehöriga åtgärder, felaktig information eller exponering av känsliga data. I takt med att GenAI blir alltmer integrerat i centrala verksamhetsarbetsflöden är det viktigt att hantera dessa risker för att upprätthålla förtroendet och säkerheten

Problem nr 2: Spridning av känslig information

Strategier för att minska spridningen av känslig information:

- **Datasanering och validering av indata:** Gör en grundlig genomsökning av användarindata för att ta bort skadligt innehåll. Använd normalisering och kodning för att förhindra missbruk.
- **Använd homomorf kryptering** för att bearbeta känsliga data på ett säkert sätt utan att innehållet exponeras. Detta säkerställer att data förblir krypterade och skyddade mot intrång även när de används.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Dra nytta av säkra API:er och systemgränssnitt** för AI-datainteraktioner och granska konfigurationer rutinmässigt i syfte att minimera exponering och attackyta.
- **Säkra datainsamling, lagring och policyer** och upprätthåll omfattande policyer för dataskydd och styrning som säkerställer regelefterlevnad och minimerar datarisker.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Säker utveckling, konfiguration och revisioner:** Tillämpa säker kodningspraxis, använd automatiserade konfigurationshanteringsverktyg och genomför regelbundna granskningar, revisioner och uppdateringar för att hålla AI-systemkonfigurationerna säkra och aktuella.
- **Användarutbildning och säkerhetsmedvetenhet:** Tillhandahåll löpande AI-specifik utbildning i säkerhetsmedvetenhet för användare och administratörer för att minska osäker användning och oavsiktlig dataspridning.

Tillbaka till arkitekturdiagrammet



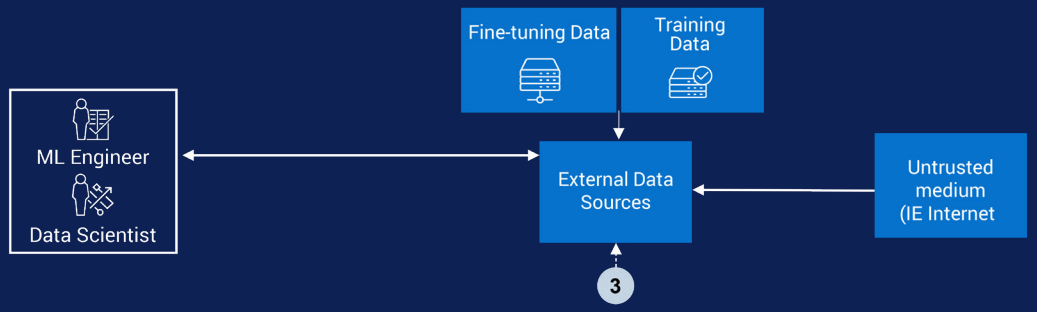
GenAI har fört med sig otroliga framsteg, men det medför också betydande risker, i synnerhet oavsiktlig exponering av känslig information. Oavsett om det handlar om identitetsuppgifter eller företagsdata kan felaktig användning eller felaktig hantering av GenAI-verktyg leda till dataläckage, bristande regelefterlevnad eller ryktesskada. Därför är det viktigt för en organisation att förstå dessa risker och hantera dem proaktivt för att säkerställa en säker implementering och användning av AI-system.

Problem nr 3: Sårbarheter i leverantörskedjan

Strategier för att minimera sårbarheter i leverantörskedjan:

- **Utvärdera leverantörer och se till att de följer praxis för en säker leverantörskedja** Utvärdera leverantörer och upprätta avtal som prioriterar säkerheten i leverantörskedjan.
- **Implementera en mjukvaruförteckning** Spåra och verifiera ursprunget för mjukvarukomponenterna, säkerställ transparensen och minska risken för kompromitterad kod.
- **Modellvalidering** Validera ML-modeller regelbundet för att säkerställa att de inte har manipulerats före driftsättning, vilket skyddar deras noggrannhet och integritet.
- **Kör behållare och POD-enheter med minsta behörighet** Detta minskar de potentiella följdverkningarna i händelse av intrång och begränsar obehörig åtkomst.
- **Installera brandväggar** Blockera onödiga nätverksanslutningar, minska exponeringen för potentiella hot och begränsa möjligheterna för angripare.
- **Skydda data och anteckningar** Skydda dina data och tillhörande anteckningar för att förhindra manipulering, obehörig åtkomst och korruption av kritisk information.
- **Säker hårdvara** Använd hårdvara som validerats för säkerhet för att förhindra sårbarheter som kan uppstå vid hårdvarubaserade attacker, vilket säkerställer en stark grund för infrastrukturen.
- **Säkra Ml-mjukvarukomponenter** Använd betrodda och granskade ML-mjukvarukomponenter för att minska sårbarheter och förbättra den övergripande säkerheten för dina arbetsflöden för maskininlärning.
- **Säker utveckling, konfiguration och revision:** Tillämpa säker kodningspraxis, använd automatiserade konfigurationshanteringsverktyg och genomför regelbundna granskningar, revisioner och uppdateringar för att hålla AI-systemkonfigurationerna säkraoch aktuella.

Tillbaka till arkitekturdiagrammet



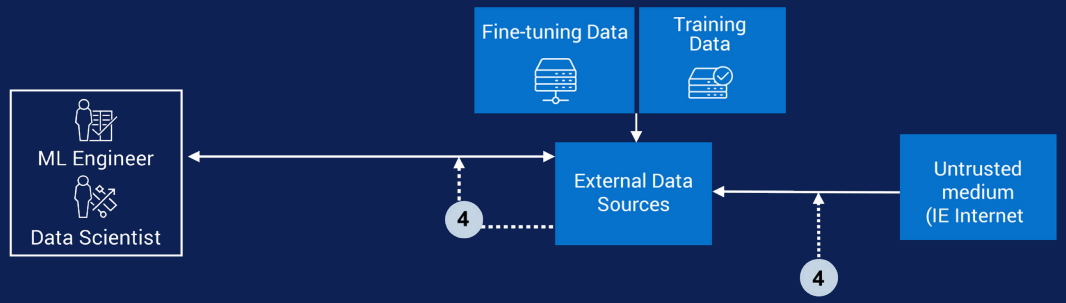
Utforska sårbarheter i LLM-leverantörskedjan som kan påverka kritiska komponenter som förhandstränad modellintegritet och adaptrar från tredje part. AI-system förlitar sig på både hård- och mjukvara som kan äventyras långt före driftsättning. Angripare kan utnyttja svagheter i olika skeden av leverantörskedjan för maskininlärning och rikta in sig på GPU-hårdvara, data och dess anteckningar, delar av ML-mjukvarustacken eller till och med själva modellen. Genom att kompromissa med dessa unika delar kan angripare få initial åtkomst till system, vilket utgör en betydande risk för säkerheten och integriteten. Att förstå och åtgärda dessa sårbarheter är avgörande för att bygga robusta, säkra AI-lösningar.

Problem nr 4: Modelldataförgiftning

Strategier för att åtgärda modelldataförgiftning:

- **Använd avvikelsedetektering och datavalidering under träningen** för att identifiera och åtgärda inkonsekvenser i data och säkerställa att endast rena data av hög kvalitet används för att träna modellen.
- **Isolera miljöer under finjusteringsfaserna** för finjustering i syfte att förhindra obehörig åtkomst och förorening av modellen under kritiska utvecklingsfaser.
- **Modellvalidering** Validera ML-modeller regelbundet för att säkerställa att de inte har manipulerats före driftsättning, vilket skyddar deras noggrannhet och integritet.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Datasanering och validering av indata:** Gör en grundlig genomsökning av användarindata för att ta bort skadligt innehåll. Använd normalisering och kodning för att förhindra missbruk.
- **Säker utveckling, konfiguration och revisioner:** Tillämpa säker kodningspraxis, använd automatiserade konfigurationshanteringsverktyg och genomför regelbundna granskningar, revisioner och uppdateringar för att hålla AI-systemkonfigurationerna säkra och aktuella.
- **Robusthetskontroller:** Genomför regelbundna utvärderingar för att identifiera och åtgärda sårbarheter och hålla AI:n säker och tillförlitlig.
- **Implementera nätverkssegmentering** för att begränsa åtkomsten till osäkra gränssnitt och kritiska systemkomponenter.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.

Tillbaka till arkitekturdiagrammet



Utforska Modelldataförgiftning är ett säkerhetshot i AI-livscykeln där motståndare avsiktligt förorenar träningsdata med skadade, vilseledande eller skadliga inmatningar. Denna risk kan påverka kritiska komponenter, från insamling av rådata och anteckningsskapande till framtagning och integrering av datauppsättningar som används för maskininlärning och stora språkmodeller. Tillförlitligheten hos AI-system beror på integriteten hos deras datakällor, som kan utsättas för manipulering före träning, under förbearbetning eller via externa datapipelines.

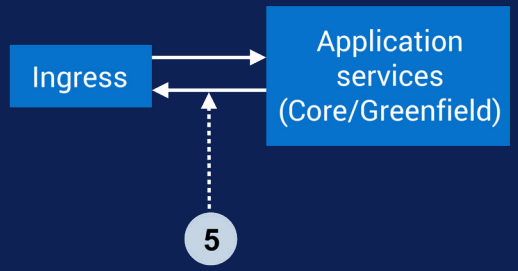
Angripare utnyttjar dataförgiftning för att försämra modellnoggrannheten, introducera sårbarheter och utlösa skadliga utdata. Genom att rikta in sig på svagheter i dataursprung, anteckningskvalitet och dataintagsprocesser kan angripare underminera säkerheten, pålitligheten och motståndskraften. Att identifiera och avhjälpa dessa datacenterhot är avgörande för att bygga upp robusta, pålitliga AI-lösningar.

Problem nr 5: Felaktig hantering av utdata

Strategier för att minska felaktig hantering av utdata:

- **Kontextmedveten utdatakodning:** Tillämpa alltid kodnings- och utlösningsmetoder som är anpassade till den specifika kontext där utdata används, till exempel HTML-, SQL- eller API-miljöer, för att förhindra sårbarheter som inmatningsattacker.
- **Sanering av utdata:** Följ strikta validerings- och saneringsmetoder för modellutdata i linje med riktlinjerna för OWASP (Open Web Application Security Project) Application Security Verification Standard) för att säkerställa säker nedströmsanvändning och minska säkerhetsriskerna.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Automatiserad utdatasäkerhetstestning:** Utför regelbundna säkerhetstester med hjälp av automatiserade verktyg för att identifiera risker i utdata, till exempel serveröverskridande skript (XSS) eller sårbarheter vid inmatning, och åtgärda dem proaktivt.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Granskning med mänsklig inblandning:** För högrisktillämpningar, såsom finans och sjukvård, krävs mänsklig övervakning och granskning av modellens utdata i syfte att säkerställa noggrannheten, säkerheten och tryggheten.
- **Sekretess och efterlevnad:** Integrera sekretessbevarande tekniker i utdataprocessen och säkerställ överensstämmelse med relevanta regelverk och standarder för säker användning av känslig information.

Tillbaka till arkitekturdiagrammet



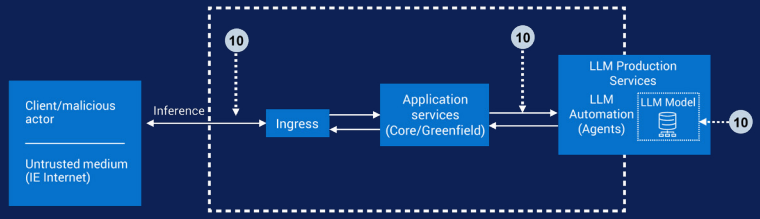
Otillräcklig validering eller sanering av AI-modellutdata kan leda till allvarliga säkerhetsrisker, däribland utökning av privilegier och dataintrång. När AI-modeller producerar utdata som inte kontrolleras eller filtreras korrekt kan skadliga aktörer utnyttja dessa sårbarheter för att få obehörig åtkomst eller eskalera sina behörigheter inom ett system. Denna brist på tillsyn kan leda till komprometterade data, obehöriga åtgärder och betydande säkerhetsintrång, vilket understryker vikten av att implementera robusta validerings- och saneringsprocesser för alla AI-genererade utdata.

Problem nr 6: För stor handlingsfrihet

Strategier för att begränsa alltför stor handlingsfriheten

- **Tillämpa minsta behörighet:** Bevilja LLM:er och agentiska delsystem endast de minimala behörigheter som krävs för att utföra avsedda åtgärder och regelbundet granska åtkomstkontroller.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Ställ upp driftsgränser:** Definiera tydligt vilka LLM:er/agenter som kan komma åt och utföra.
- **Granskning med mänsklig inblandning:** För högrisktillämpningar, såsom finans och sjukvård, krävs mänsklig övervakning och granskning av modellens utdata i syfte att säkerställa noggrannheten, säkerheten och tryggheten.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Begränsa autonomin:** Begränsa LLM-funktionerna för att undvika obegränsad åtkomst och kontroll.
- **Säker utveckling, konfiguration och revisioner:** Tillämpa säker kodningspraxis, använd automatiserade konfigurationshanteringsverktyg och genomför regelbundna granskningar, revisioner och uppdateringar för att hålla AI-systemkonfigurationerna säkra och aktuella.
- **Installera brandväggar** Blockera onödiga nätverksanslutningar, minska exponeringen för potentiella hot och begränsa möjligheterna för angripare.
- **Robusthetskontroller:** Genomför regelbundna utvärderingar för att identifiera och åtgärda sårbarheter och hålla AI säker och tillförlitlig.

Tillbaka till arkitekturdiagrammet



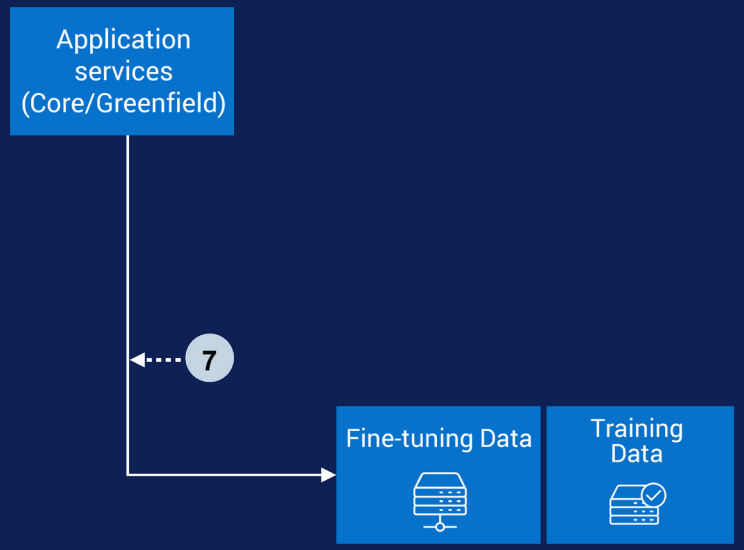
Att ge AI-agenter eller insticksprogram alltför stor autonomi eller onödiga funktioner i arbetsflöden kan medföra betydande risker. När ett AI-system får behörigheter och funktioner utöver vad som krävs ökar sannolikheten för oavsiktliga konsekvenser. Detta kan inträffa när system baserade på stora språkmodeller (LLM) är utformade med alltför omfattande behörigheter, vilket gör det möjligt för dem att vidta åtgärder eller komma åt information de inte borde. Detta kan leda till fel, missbruk av data eller till och med säkerhetsrisker, vilket understryker vikten av att noggrant begränsa och övervaka AI-funktionerna för att säkerställa en säker och ansvarsfull användning.

Problem nr 7: Promptläckage

Strategier för att minska promptläckage

- **Undvik att bädda in känslig information i prompter** Inkludera aldrig inloggningsuppgifter, API-nycklar eller tillverkarspecifik logik i prompter – hantera dem säkert utanför systemet.
- **Separera säkerhetskontrollerna från prompterna** Hantera autentisering, auktorisering och sessionshantering i programlogik, inte i prompter.
- **Validera in- och utdata** Sanera prompter och svar med robust validering för att blockera misstänkta mönster och manipulation.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Kryptera och säkra prompter** Lagra prompter och konfigurationer i krypterad, säker lagring för att förhindra obehörig åtkomst.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Granska prompterna regelbundet** Granska och sanera prompterna regelbundet för att ta bort känsliga data och säkerställa säkerhetskompatibiliteten.
- **Testa och simulera attacker för att hitta svagheter** Utför angripartester för att identifiera och åtgärda sårbarheter i prompthantering och utdata.
- **Isolera prompter från användarinmatningar** Utforma system för att förhindra att användarfrågor manipulerar eller exponerar prompter.
- **Tillämpa hastighetsgränser** Sätt gränser för API-användningen, begränsa misstänkt aktivitet och blockera automatiserade promptattacker.

Tillbaka till arkitekturdiagrammet



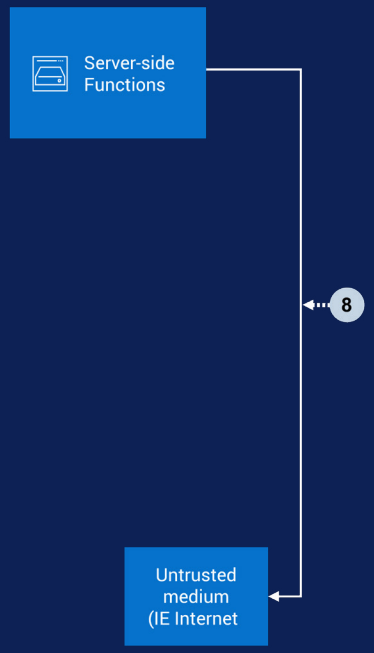
En systempromptsläckageattack på en stor språkmodell (LLM) eller ett AI-system inträffar när en angripare kan extrahera eller sluta sig till de dolda instruktionerna – ”systemprompterna” – som styr modellens beteende och anger driftsgränserma. Dessa prompter är vanligtvis inte avsedda att vara synliga för slutanvändaren, eftersom de innehåller grundläggande regler, begränsningar och ibland känslig driftlogik. Genom specialutformade indata eller genom att utnyttja sårbarheter kan en angripare lura LLM-modellen att avslöja sin systemprompt, antingen helt eller delvis. Om den här informationen läcker ut kan den användas för att bakåtkonstruera begränsningar, kringgå säkerhetsfilter och skapa nya riktade attacker, vilket i slutänden ökar risken för promptinmatning, utökning av privilegier och missbruk av modellen och nedströmssystem som är beroende av dess integritet.

Problem nr 8: Svagheter i vektor och inbäddning

Strategier för att minska vektorsvagheter/ inbyggda svagheter

- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Kryptering** säkrar vektordata under transport och vila med robusta krypteringsstandarder som AES.
- **Säker konfiguration och övervakning** stärker systemen, konfigurerar dem säkert och övervakar kontinuerligt för att upptäcka felkonfigurationer, obehörig åtkomst och avvikelser.
- **Sårbarhetshantering** uppdaterar och korrigerar regelbundet all mjukvara, alla beroenden och alla vektorlagringsmotorer för att hantera säkerhetsrisker.
- **Datasanering och validering av indata:** Gör en grundlig genomsökning av användarindata för att ta bort skadligt innehåll. Använd normalisering och kodning för att förhindra missbruk.
- **Dra nytta av säkra API:er och systemgränssnitt** för AI-datainteraktioner och granska konfigurationer rutinmässigt i syfte att minimera exponering och attackyta.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Säker hårdvara** Använd hårdvara som validerats för säkerhet för att förhindra sårbarheter som kan uppstå vid hårdvarubaserade attacker, vilket säkerställer en stark grund för infrastrukturen.
- **Säker utveckling, konfiguration och revisioner:** Tillämpa säker kodningspraxis, använd automatiserade konfigurationshanteringsverktyg och genomför regelbundna granskningar, revisioner och uppdateringar för att hålla AI-systemkonfigurationerna säkra och aktuella.

Tillbaka till arkitekturdiagrammet



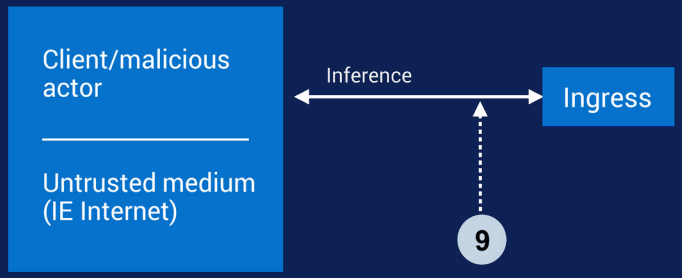
En vektor- och inbäddningssvaghetsattack på en stor språkmodell (LLM) eller ett AI-system – i synnerhet om RAG (Retrieval Augmented Generation) används – riktar in sig på sårbarheter i hur informationen kodas, lagras och hämtas som numeriska vektorer och inbäddningar. Svagheter i dessa mekanismer kan utnyttjas genom skadliga handlingar som inbäddningsinversion (återskapa känsliga data från inbäddningar), dataförgiftning (inmatning av skadligt eller missvisande innehåll för att manipulera modellbeteende), obehörig åtkomst till vektordatabaser (som leder till dataläckor), eller manipulation av hämtningsutdata. Dessa attacker hotar integriteten och tillförlitligheten genom att göra det möjligt för angripare att komma åt känslig information, ändra utdata och underminera användarnas förtroende för AI-drivna program. Korrekta åtkomstkontroller, datavalidering, kryptering och kontinuerlig övervakning är avgörande för att det ska vara möjligt att försvara sig mot dessa föränderliga hot.

Problem nr 9: Felaktig information

Strategier för att åtgärda felaktig information

- **RAG (Retrieval-Augmented Generation) med auktoritativa källor:** Använd RAG för att hämta och integrera information från verifierade, betrodda databaser och kunskapsdatabaser, vilket minskar hallucinationer.
- **Modelljustering och kalibrering av utdata:** Finjustera modellerna med olika datauppsättningar och tillämpa tekniker för att minimera missvisande och felaktig information.
- **Automatiserad faktakontroll:** Korsreferera utdata med tillförlitliga källor och flagga falsk information automatiskt.
- **Osäkerhetsövervakning:** Flagga svar med lågt förtroende för mänsklig granskning i kritiska fall.
- **Granskning med mänsklig inblandning:** För högrisktillämpningar, såsom finans och sjukvård, krävs mänsklig övervakning och granskning av modellens utdata i syfte att säkerställa noggrannheten, säkerheten och tryggheten.
- **Återkoppling från användare:** Gör det möjligt för användarna att rapportera fel för kontinuerlig modellförbättring och snabb korrigering av desinformationsvägar.
- **Åtkomstbegränsningar och mänsklig tillsyn:** Tillämpa rollbaserad åtkomstkontroll (RBAC), multifaktorautentisering (MFA) och identitetshantering för att begränsa åtkomsten. Använd mänsklig granskning inför kritiska beslut.
- **Säker utveckling, konfiguration och revision:** Tillämpa säker kodningspraxis, använd automatiserade konfigurationshanteringsverktyg och genomför regelbundna granskningar, revisioner och uppdateringar för att hålla AI-systemkonfigurationerna säkra och aktuella.
- **Riskkommunikation:** Utbilda användare om AI-begränsningar och uppmuntra oberoende verifiering.
- **Avsiktlig UI- och API-design:** Framhäv AI-genererat innehåll och ge användaren information om ansvarsfull användning.

Tillbaka till arkitekturdiagrammet



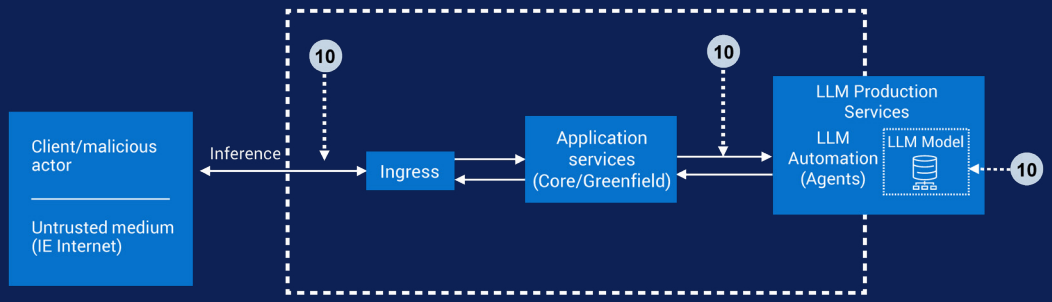
En desinformationsattack på ett LLM- eller AI-system är ett avsiktligt försök att få modellen att generera eller sprida falsk, vilseledande eller till synes trovärdig – men felaktig – information via sina utdata. Den här sårbarheten beror på flera faktorer: Modellens tendens att ”hallucinera” (generera påhittat innehåll som låter rimligt), träningsdata som är missvisande eller innehåller luckor samt påverkan från angriparprompter. Hallucinationer uppstår eftersom stora språkmodeller statistiskt genererar text som passar ett mönster, snarare än att verkligen förstå fakta, vilket leder till svar som verkar auktoritativa men faktiskt är ogrundade. Riskerna med sådana attacker inbegriper säkerhetsintrång, ryktesskada och till och med rättsligt ansvar, i synnerhet i miljöer där användarna förlitar sig alltför mycket på LLM-svar utan att verifiera att de är korrekta eller validerade, och potentiellt bäddar in fel eller felaktig information i kritiska beslut och processer.

Problem nr 10: Obegränsad konsumtion

Strategier för obegränsad konsumtion

- **Tillämpa hastighetsbegränsningar och användarkvoter** för att sätta strikta gränser för förfrågningar, token eller data per användare, API-nyckel eller app för att förhindra missbruk.
- **Kräv autentisering och användarsegmentering** använder stark autentisering (t.ex. API-nycklar, OAuth) och tilldela roller eller nivåer för att bearbeta endast autentiserade begäranden.
- **Indatavalidering och storleksbegränsningar** validerar promptstorlek och -struktur, vilket blockerar/rensar stora respektive felaktigt utformade frågor.
- **Använd tidsgränser för behandling och resursbegränsningar** för varje begäran för att undvika långsamma operationer och resursutarmning.
- **Distribuera smart cachelagring och deduplicering** Cachelagra svar för att hitta dubbletter och liknande frågor för att minska onödig bearbetning.
- **Övervakning, loggning och avvikelsedetektering:** Övervaka och logga kontinuerligt AI-systemaktiviteter – med lösningar som MDR/XDR/SIEM – för att snabbt upptäcka, undersöka och reagera på obehörig åtkomst, avvikelser och dataläckage.
- **Budgetspårning och utgiftskontroller** har instrumentpaneler och varningar för att övervaka kostnader och blockera användningen vid budgetgränser.
- **Sandboxnings- och isoleringstekniker** kör arbetsbelastningar i isolerade miljöer med begränsade behörigheter för att minska riskerna.
- **Begränsa samtalsdjup och konversationshantering** begränsar rekursiva samtal och konversationssteg för att förhindra utnyttjande.
- **Tillämpa nivåindelade modeller eller resursfördelning** dirigerar högprioriterade begäranden till premiummodeller och lågprioriterad trafik till kostnadseffektiva modeller.

Tillbaka till arkitekturdiagrammet



Ett obegränsat konsumtionshot mot ett LLM- eller AI-system avser en säkerhetssårbarhet där programmet tillåter användare – illasinnade eller andra – att skicka in för stora, okontrollerade inferensförfrågningar eller prompter utan effektiv hastighetsbegränsning, autentisering eller användningsbegränsningar. Eftersom LLM-inferens tar stora beräkningsresurser i anspråk kan denna brist på kontroll utnyttjas på flera sätt: Angripare kan orsaka överbelastningsattacker (DoS-attacker) genom att överbelasta systemresurserna, generera oförutsedda ekonomiska förluster i distributioner med användningsbaserad betalning eller molnvärdsbaserade distributioner, eller systematiskt fråga modellen för att klona dess beteende och stjäla immateriell egendom. Konsekvenserna inkluderar tjänstestörningar, försämrad prestanda för andra användare, ekonomisk belastning och ökad risk för känsligt modelläckage. Obegränsad konsumtion uppstår när resursanvändningen inte styrs korrekt, vilket gör att LLM-baserade program utsätts för både oavsiktlig och avsiktlig exploatering.

Varför du ska välja Dell för AI-säkerhet

Dell hjälper organisationer att skydda AI-modeller och LLM-modeller genom en omfattande metod som inbegriper hårdvara, mjukvara och hanterade tjänster. Säkerheten är inbäddad från leverantörskedjan till enheten, infrastrukturen, data och program, alltihop anpassat till nollförtroendeprinciper. I hela produktportföljen är Dells lösningar byggda för att främja cyberhygien med funktioner som MFA, RBAC, minsta behörighet och kontinuerlig verifiering. Den här heltäckande metoden är säkert designad och säkerställer att organisationer tryggt kan förnya sig med AI- och LLM-modeller, vilket minimerar risken för modellstöld, dataläckage, attacker från angripare och andra avancerade cyberhot.

Leverantörskedja

Dells säkra leverantörskedja ger grundläggande skydd för AI-modeller och stora språkmodeller genom att bädda in säkerhet i varje steg av produktutvecklingen, tillverkningen och leveransen. Genom kryptografiskt signerade uppdateringar av BIOS och fast mjukvara, säker komponentverifiering, AI-fokuserad mjukvaruförteckning (SBOM), spårning av datauppsättningar, integrerad säkerhetsprogramvara och -konfiguration samt rigorösa leverantörsriskbedömningar som är anpassade till globala standarder minimerar Dell riskerna för manipulering, obehörig åtkomst och attacker på leverantörskedjan – vilket säkerställer att organisationer kan distribuera betrodda, motståndskraftiga AI-arbetsbelastningar med fullständig transparens, integritet och regelefterlevnad.

AI-datorer

Dell erbjuder grundläggande säkerhet för AI-arbetsbelastningar på enheten. Dell Trusted Devices – världens säkraste kommersiella AI-datorer* – är utformade med säkerhet i åtanke. Säkerhet i leverantörskedjan minimerar risken för produktsårbarheter och manipulering. Unika försvar som är inbyggt direkt i hårdvara och fast mjukvara skyddar datorn och slutanvändaren under användning. Dell SafeBIOS ger djup synlighet på BIOS-nivå och manipuleringsdetektering, medan Dell SafeID förbättrar säkerheten för inloggningsuppgifter och möjliggör lösenordsfri autentisering. Partnerprogramvara ger avancerat skydd för slutpunkter, nätverk och molnmiljöer.

Cyberelasticitet

Dells PowerProtect-lösningar för cyberelasticitet skyddar AI-data med krypterade, oföränderliga säkerhetskopior, snabb återställning och isolerade cyberåterställningsvalv. Dessa funktioner förhindrar förstörelse, minskar påverkan från skadliga uppdateringar och stöder efterlevnad och återställning efter en attack.

Servrar

PowerEdge-servrar har konfidentiell databehandling för att isolera och säkra AI-/LLM-prompter och -inbäddningar, pålitliga RAG-lösningar (Retrieval-Augmented Generation) förankrade i auktoritativa källor, tillsammans med MFA, RBAC, kiselbaserad förtroenderot, signerad fast mjukvara och kontinuerlig övervakning för att skydda kritiska AI-arbetsbelastningar.

Lagring

Dells lagringsportfölj säkerställer säker, krypterad lagring för känsliga AI-data med robust AES-256-kryptering för data i vila och under överföring. Avancerad kryptering som är utformad för att vara tålig mot framtida kvanthot finns tillgänglig med utvalda erbjudanden. Portföljen innehåller NVMe-prestanda med hög hastighet, FIPS-kompatibla krypteringsmoduler för att skydda data – inklusive sådana som används i AI-arbetsbelastningar – oföränderliga ögonblicksbilder och valv med luftgap för cyberåterställning för att motverka attacker med utpressningsvirus. Nollförtroendearkitektur, säkerhet i leverantörskedjan och manipuleringssäkra revisionsfunktioner förbättrar styrningen. Inbyggd avvikelsedetektering och AIOps ML-modeller skyddar arbetsbelastningar utan att använda kunddata för träning, vilket minimerar inmatningsbaserade attackrisker.

AIOps

Dell AIOps tillhandahåller automatiserad, kontinuerlig övervakning för att upptäcka felkonfigurationer, sårbarheter (inklusive CVE:er) och stöder medvetenhet om leverantörskedjerisker som påverkar AI-/LLM-arbetsbelastningar. CVE-skanning i realtid, smarta varningar och AI-drivna instrumentpaneler möjliggör snabba åtgärder genom att flagga avvikelser och spåra arbetsflöden för upplösning. Inbyggda efterlevnadsfunktioner, rollbaserade åtkomstkontroller och automatiserad rapportering bidrar till att upprätthålla säker drift över arbetsbelastningar, medan smidig EDR-/XDR-integrering och AI-drivna driftsinsikter – inklusive generativa funktioner i lösningar som stöds – förbättrar IT-effektiviteten ytterligare.

Nätverk

Dell Networking-lösningar skyddar AI-/LLM-miljöer genom robust nätverkssegmentering, vilket minimerar laterala förflyttningar. Krypterade nätverkssökvägar och integrerade brandväggskontroller blockerar obehörig åtkomst till AI-data.

Tjänster för AI-säkerhet och motståndskraft

Dells tjänster för AI-säkerhet och motståndskraft är utformade för att hantera nya risker som är förknippade med att AI integreras i organisationen. Våra tjänster är byggda för att fungera tillsammans med dina team så snabbt som möjligt när du använder AI och tillhandahåller expertis för att vägleda i strategisk planering, implementering av lösningar och hanterade säkerhetstjänster för att underlätta driftsbördan så att du kan förnya på ett säkert sätt med AI. Var och en är skräddarsydd för att hjälpa organisationer att hantera föränderliga AI-risker och optimera säkra AI-distributioner.

Dell AI Factory

En integrerad portfölj med specialbyggd säkerhet som Dells säkra leverantörskedja, nollförtroendefunktioner för att upprätthålla minsta möjliga privilegier och AI MDR-lösningar som utformats för att skydda modellen.

* Enligt Dells interna analyser, oktober 2024 (Intel) och mars 2025 (AMD). Gäller datorer med Intel och AMD processorer. Alla funktioner är inte tillgängliga för alla datorer. För vissa funktioner krävs ytterligare köp. Intel-baserade datorer validerade av Principled Technologies, juli 2025.

Sammanfattning

För att kunna bygga tåliga AI-ramverk är en samarbetsinriktad strategi mellan organisationer och säkerhetsexperter avgörande. I takt med att AI och LLM:er fortsätter att omforma olika branscher är det viktigt att hantera de risker som de medför, inklusive utmaningar inom datasäkerhet, modellintegritet och efterlevnad. Organisationer måste prioritera proaktiva strategier som integrerar säkerheten i varje steg av AI-resan.

Dell Technologies är en betrodd partner i det här arbetet och erbjuder heltäckande GenAI-anpassning, säkerhetskonsultation och integrerade lösningar som är anpassade efter användarens unika behov. Genom att utnyttja Dells robusta cybersäkerhetslösningar kan företagen effektivt minska AI- och LLM-riskerna samtidigt som de maximerar potentialen i sina befintliga säkerhetsinvesteringar. Dell gör det möjligt för organisationer att skydda sin AI-infrastruktur genom att sömlöst integrera avancerad säkerhet i nuvarande ramverk, vilket säkerställer en framtidssäkrad och trygg miljö.

Läs mer om hur Dells omfattande AI-lösningar kan skydda GenAI- och LLM-miljöer: Dell.com/CyberSecurityMonth