DELL Technologies

As dez principais questões de segurança cibernética para IA generativa e LLMs

Começar

Introdução

A Inteligência Artificial (IA) está revolucionando a maneira como as organizações operam, com a Inteligência artificial generativa (GenAI) e os Grandes modelos de linguagem (LLMs) se tornando cargas de trabalho essenciais nos ambientes empresariais modernos.

Assim como qualquer outra carga de trabalho, essas aplicações trazem consigo um conjunto próprio de complexidades e vulnerabilidades que devem ser abordadas. À medida que as empresas continuam a adotar a IA para promover a inovação, a eficiência e a vantagem competitiva, garantir a segurança dessas aplicações se torna uma necessidade fundamental. Uma boa higiene cibernética é a base para proteger qualquer carga de trabalho. Assim como você prioriza a segurança em todas as cargas de trabalho, é essencial praticar uma boa higiene cibernética para a IA também. Isso inclui a implementação de práticas, como aplicação adequada de patches do sistema, autenticação baseada em vários fatores, acesso baseado em função e segmentação de rede. Essas medidas são fundamentais, mas a chave está em entender como esses recursos se encaixam na arquitetura específica e no uso de sua carga de trabalho.

Na Dell, temos uma compreensão profunda da carga de trabalho de IA e dos desafios exclusivos de segurança que ela enfrenta. Ao identificar as maneiras como os agentes de ameaça podem atingir essas cargas de trabalho, a Dell pode ajudar você a criar uma estratégia de segurança robusta. Isso inclui lidar com riscos como: envenenamento de dados de treinamento, roubo ou manipulação de modelos, reconstrução de conjuntos de dados e muito mais.

Também nos concentramos no gerenciamento de desafios associados à entrada em seu modelo de IA, como evitar a divulgação de informações confidenciais, mitigar vieses ou tópicos inseguros e garantir a conformidade com as normas. Do lado do resultado, ajudamos a lidar com problemas como a dependência excessiva do modelo e os riscos relacionados à conformidade.

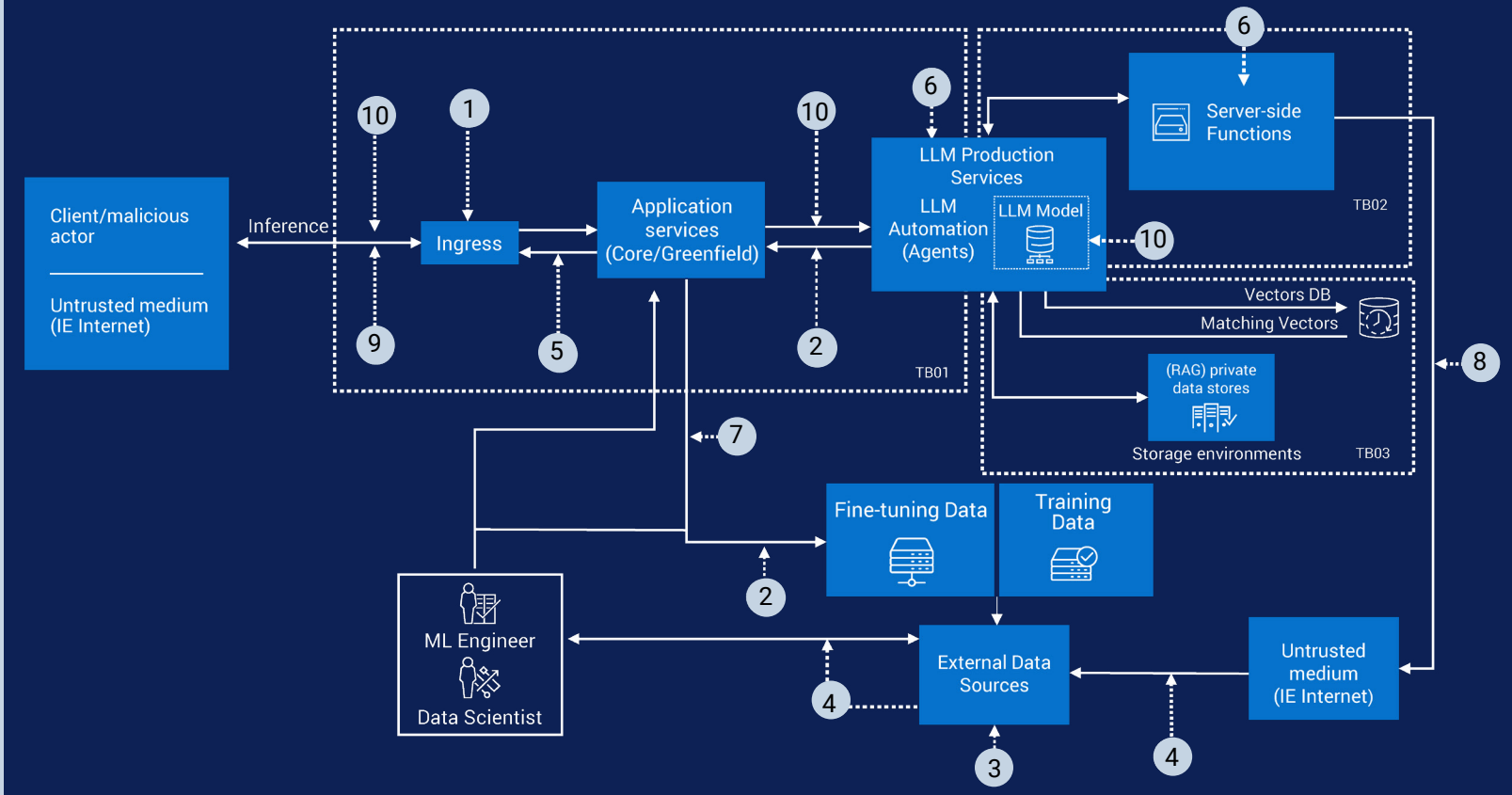
Na Dell, capacitamos as empresas a reduzir esses riscos aproveitando suas soluções de segurança cibernética existentes ou explorando novas ferramentas e práticas para proteger seus sistemas. Nosso objetivo é garantir que a segurança não atrapalhe sua inovação. Ao entender como as cargas de trabalho de IA funcionam e as ameaças de segurança que elas enfrentam, podemos ajudar você a criar uma postura de segurança mais forte, tornando seu ambiente mais resiliente e, ao mesmo tempo, permitindo que você inove com confiança. Com nossa experiência, ajudamos você a aproveitar com confiança o potencial da IA e, ao mesmo tempo, manter uma segurança robusta.

As dez principais questões de segurança cibernética para IA generativa e LLM

Essas são as principais preocupações para proteger modelos de IA generativa/LLM, conforme descrito pelo OWASP.

Clique em cada questão saber mais:

1. Injeção de prompts
2. Divulgação de informações confidenciais
3. Cadeia de suprimentos
4. Envenenamento de dados do modelo
5. Tratamento inapropriado de saídas
6. Excesso de agências
7. Vazamento de prompts do sistema
8. Vulnerabilidades de incorporações e vetores
9. Informações incorretas
10. Consumo ilimitado

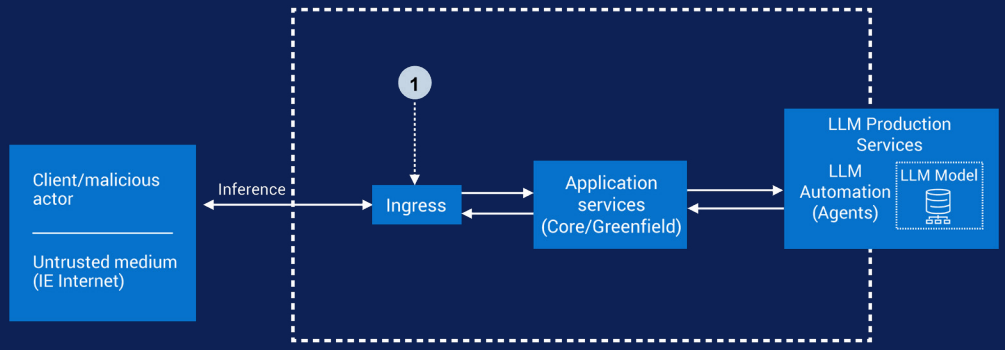


Problema nº 1: injeção de prompt

Estratégias para atenuar a injeção de prompt:

- **Limpeza de dados e validação de entrada:** analise cuidadosamente as entradas do usuário para remover conteúdo prejudicial. Use normalização e codificação para evitar o uso indevido.
- **Abordagens baseadas em processamento de linguagem natural (NLP) e aprendizado de máquina:** use o NLP e o aprendizado de máquina para detectar e bloquear prompts manipulados ou mal-intencionados.
- **Formatação de saída clara e controles de resposta:** defina limites rígidos de resposta para garantir que as saídas sigam os formatos pretendidos e impeçam ações não autorizadas. Use a filtragem de prompts e a validação de respostas para manter a integridade.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Engenharia de prompts seguros:** use o design e a análise de prompts seguros como parte da segurança geral do software para proteger o processamento de entradas.
- **Validação de modelos:** valide regularmente os modelos de ML para garantir que eles não tenham sido adulterados antes da implementação, protegendo sua precisão e integridade.
- **Filtragem, classificação e validação de respostas de prompts:** analise e classifique prompts para garantir que apenas entradas seguras sejam processadas. Valide as respostas para evitar o uso indevido.
- **Verificações de robustez:** realize avaliações regulares para identificar e corrigir vulnerabilidades, mantendo a IA segura e confiável.

Voltar ao Diagrama de arquitetura



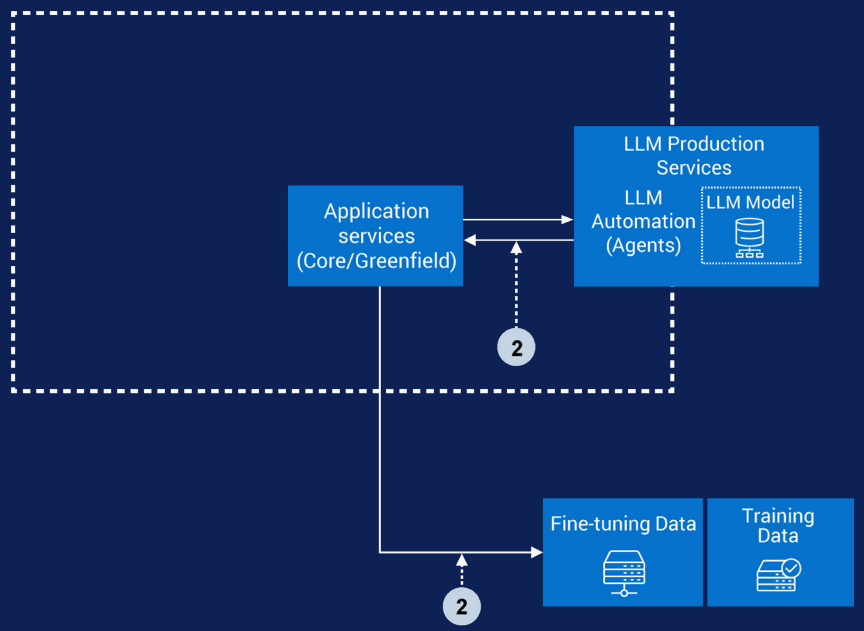
A injeção de prompt é um desafio emergente no mundo da IA generativa (GenAI), em que entradas mal-intencionadas são criadas para manipular o comportamento ou comprometer a integridade do modelo. Esses ataques exploram vulnerabilidades na forma como os sistemas de IA processam e respondem às entradas do usuário, o que pode levar a ações não autorizadas, à desinformação ou à exposição de dados confidenciais. À medida que a IA generativa se torna cada vez mais integrada aos fluxos de trabalho essenciais dos negócios, lidar com esses riscos é essencial para manter a confiança e a segurança

Problema nº 2: divulgação de informações confidenciais

Estratégias para atenuar a divulgação de informações confidenciais:

- **Limpeza de dados e validação de entrada:** analise cuidadosamente as entradas do usuário para remover conteúdo prejudicial. Use normalização e codificação para evitar o uso indevido.
- **Utilizar criptografia homomórfica** para processar dados confidenciais com segurança sem expor seu conteúdo. Isso garante que, mesmo enquanto os dados estiverem em uso, eles permaneçam criptografados e protegidos contra violações.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Utilizar APIs seguras e interfaces de sistema** para interações de dados de IA, analisando rotineiramente as configurações a fim de minimizar a exposição e a superfície de ataque.
- **Proteger a coleta, o armazenamento e as políticas** de dados e aplicar políticas abrangentes de governança e proteção de dados que garantam a conformidade normativa e minimizem o risco de dados.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Desenvolvimento, configuração e auditorias seguros:** aplique práticas seguras de codificação, use ferramentas automatizadas de gerenciamento de configurações e realize revisões, auditorias e atualizações regulares para manter as configurações do sistema de IA seguras e atualizadas.
- **Educação do usuário e conscientização sobre segurança:** ofereça treinamento contínuo de conscientização sobre segurança específica para IA a usuários e administradores para reduzir o uso inseguro e a divulgação acidental de dados.

Voltar ao Diagrama de arquitetura



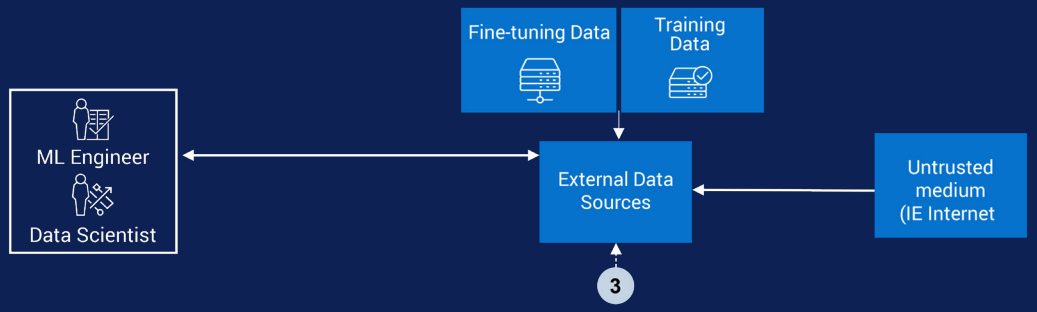
A IA generativa trouxe avanços incríveis, mas também traz riscos significativos, especialmente a exposição não intencional de informações confidenciais. Sejam informações de identificação pessoal (PII) ou dados comerciais proprietários, o uso indevido ou o manuseio incorreto de ferramentas de IA generativa pode levar a vazamentos de dados, a não conformidade regulatória ou danos à reputação. É fundamental que as organizações entendam esses riscos e os enfrentem de maneira proativa para garantir a implementação e o uso seguros de sistemas de IA.

Problema nº 3: vulnerabilidades da cadeia de suprimentos

Estratégias para reduzir vulnerabilidades da cadeia de suprimentos:

- **Avaliar fornecedores e garantir a conformidade com práticas seguras na cadeia de suprimentos** Avalie fornecedores e estabeleça acordos que priorizem a segurança da cadeia de suprimentos.
- **Implementar uma lista de materiais de software** Monitore e verifique as origens dos componentes de software, garantindo a transparência e reduzindo o risco de comprometimento do código.
- **Validação de modelos:** valide regularmente os modelos de ML para garantir que eles não tenham sido adulterados antes da implementação, protegendo sua precisão e integridade.
- **Executar contêineres e pods com o mínimo de privilégios** Isso reduz o impacto potencial em caso de comprometimento e limita o acesso não autorizado.
- **Implementar firewalls:** bloqueie a conectividade de rede desnecessária, reduzindo a exposição a possíveis ameaças e limitando os caminhos para invasores.
- **Proteger dados e anotações** Proteja seus dados e anotações associadas para evitar adulteração, acesso não autorizado e corrupção de informações essenciais.
- **Hardware seguro:** use hardware validado para segurança a fim de evitar vulnerabilidades que possam surgir de ataques baseados em hardware, garantindo uma base sólida para sua infraestrutura.
- **Componentes de software de ML seguros** Use componentes de software de ML confiáveis e verificados para reduzir vulnerabilidades e aprimorar a segurança geral de seus fluxos de trabalho de aprendizado de máquina.
- **Desenvolvimento, configuração e auditorias seguros:** aplique práticas seguras de codificação, use ferramentas automatizadas de gerenciamento de configurações e realize revisões, auditorias e atualizações regulares para manter as configurações do sistema de IA seguras e atualizadas.

Voltar ao Diagrama de arquitetura



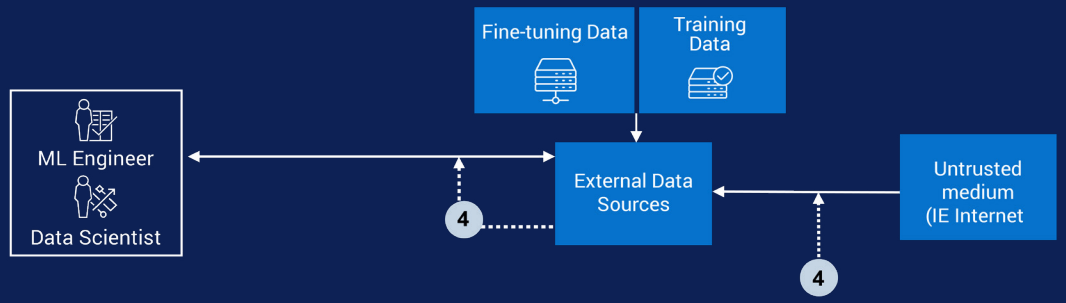
Explore as vulnerabilidades na cadeia de suprimentos do LLM que podem afetar componentes essenciais, como integridade do modelo pré-treinado e adaptadores de terceiros. Os sistemas de IA dependem de hardware e software que podem ser comprometidos muito antes da implementação. Os adversários podem explorar os pontos fracos em diferentes estágios da cadeia de suprimentos de aprendizado de máquina, visando o hardware da GPU, os dados e suas anotações, os elementos da pilha de software de ML ou até mesmo o próprio modelo. Ao comprometer essas partes exclusivas, os invasores podem obter acesso inicial aos sistemas, o que representa riscos significativos à segurança e à integridade. Entender e mitigar essas vulnerabilidades é crucial para criar soluções de IA robustas e seguras.

Problema nº 4: envenenamento de dados do modelo

Estratégias para reduzir o envenenamento de dados do modelo:

- **Usar a detecção de anomalias e a validação de dados durante o treinamento** para identificar e resolver inconsistências nos dados e garantir que apenas dados limpos e de alta qualidade sejam usados para treinar o modelo.
- **Isolar ambientes durante as fases de ajuste** para evitar o acesso não autorizado ou a contaminação do modelo durante estágios críticos de desenvolvimento.
- **Validação de modelos:** valide regularmente os modelos de ML para garantir que eles não tenham sido adulterados antes da implementação, protegendo sua precisão e integridade.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Limpeza de dados e validação de entrada:** analise cuidadosamente as entradas do usuário para remover conteúdo prejudicial. Use normalização e codificação para evitar o uso indevido.
- **Desenvolvimento, configuração e auditorias seguros:** aplique práticas seguras de codificação, use ferramentas automatizadas de gerenciamento de configurações e realize revisões, auditorias e atualizações regulares para manter as configurações do sistema de IA seguras e atualizadas.
- **Verificações de robustez:** realize avaliações regulares para identificar e corrigir vulnerabilidades, mantendo a IA segura e confiável.
- **Implementar a segmentação de rede** para limitar o acesso a interfaces inseguras e componentes essenciais do sistema.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.

Voltar ao Diagrama de arquitetura



O envenenamento de dados de modelos é uma ameaça à segurança no ciclo de vida da IA, em que os adversários contaminam intencionalmente os dados de treinamento com entradas corrompidas, enganosas ou mal-intencionadas. Esse risco pode afetar componentes essenciais, desde a coleta e anotação de dados brutos até a curadoria e integração de conjuntos de dados usados para aprendizado de máquina ou grandes modelos de linguagem. A confiabilidade dos sistemas de IA depende da integridade de suas fontes de dados, que podem ser expostas à manipulação antes do treinamento, durante o pré-processamento ou por meio de pipelines de dados externos.

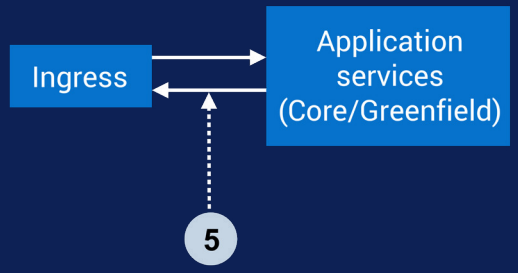
Os invasores aproveitam o envenenamento de dados para degradar a precisão do modelo, introduzir vulnerabilidades ou acionar resultados prejudiciais. Ao visar pontos fracos na proveniência de dados, qualidade de anotações ou processos de ingestão de conjuntos de dados, os adversários podem prejudicar a segurança, a confiabilidade e a resiliência. É essencial reconhecer e atenuar essas ameaças centradas em dados para criar soluções robustas e confiáveis de IA.

Questão 5: tratamento inapropriado de saídas

Estratégias para mitigar o tratamento inapropriado de saídas:

- **Codificação de saída sensível ao contexto:** sempre aplique técnicas de codificação e escape adaptadas ao contexto específico em que a saída será usada, como ambientes HTML, SQL ou API, para evitar vulnerabilidades como ataques de injeção.
- **Higienização de saídas:** siga as práticas rígidas de validação e limpeza de saídas de modelos em conformidade com as diretrizes do Padrão de Verificação de Segurança de Aplicativos (ASVS) do Open Web Application Security Project (OWASP) para garantir o uso seguro downstream e reduzir os riscos de segurança.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Teste automatizado de segurança de saída:** realize testes regulares de segurança usando ferramentas automatizadas para identificar riscos nos resultados, como cross-site scripting (XSS) ou vulnerabilidades de injeção, e resolva-os proativamente.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Revisão com intervenção humana:** para aplicações de alto risco, como as de finanças ou saúde, exija supervisão e análise humanas das saídas de modelos para garantir precisão, segurança e proteção.
- **Privacidade e conformidade:** integre técnicas de preservação de privacidade ao processo de saída e garanta a conformidade com normas e padrões relevantes para o uso seguro de informações confidenciais.

Voltar ao Diagrama de arquitetura



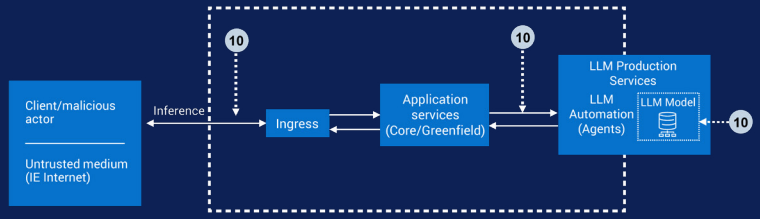
A validação ou higienização insuficiente do resultado do modelo de IA pode levar a sérios riscos de segurança, incluindo elevação de privilégios e violações de dados. Quando os modelos de IA produzem resultados que não são verificados ou filtrados corretamente, agentes mal-intencionados podem explorar essas vulnerabilidades para obter acesso não autorizado ou ampliar seus privilégios em um sistema. Essa falta de supervisão pode resultar em comprometimento de dados, ações não autorizadas e violações significativas de segurança, destacando a importância de implementar processos robustos de validação e higienização para quaisquer resultados gerados por IA.

Questão nº 6: excesso de agências

Estratégias para reduzir o excesso de agências

- **Implementar o princípio de privilégio mínimo:** conceda aos LLMs e subsistemas agênticos apenas as permissões mínimas necessárias para realizar as operações pretendidas e revise regularmente os controles de acesso.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Definir limites operacionais:** defina claramente quais LLMs/agentes podem acessar ou executar.
- **Revisão com intervenção humana:** para aplicações de alto risco, como as de finanças ou saúde, exija supervisão e análise humanas das saídas de modelos para garantir precisão, segurança e proteção.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Limitar a autonomia:** restrinja os recursos do LLM para evitar acesso ou controle irrestrito.
- **Desenvolvimento, configuração e auditorias seguros:** aplique práticas seguras de codificação, use ferramentas automatizadas de gerenciamento de configurações e realize revisões, auditorias e atualizações regulares para manter as configurações do sistema de IA seguras e atualizadas.
- **Implementar firewalls:** bloqueie a conectividade de rede desnecessária, reduzindo a exposição a possíveis ameaças e limitando os caminhos para invasores.
- **Verificações de robustez:** realize avaliações regulares para identificar e corrigir vulnerabilidades, mantendo a IA segura e confiável.

Voltar ao Diagrama de arquitetura



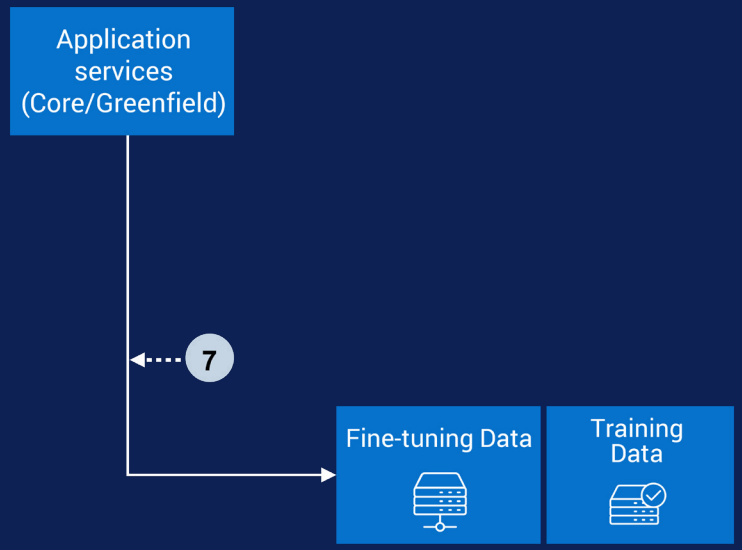
Conceder a agentes ou plug-ins de IA autonomia excessiva ou funcionalidade desnecessária dentro dos fluxos de trabalho pode representar riscos significativos. Quando um sistema de IA recebe privilégios ou recursos além do necessário, isso aumenta a probabilidade de consequências não intencionais. Isso pode acontecer quando sistemas baseados em grandes modelos de linguagem (LLM) são projetados com permissões excessivas, permitindo que eles tomem ações ou acessem informações que não deveriam. Esse excesso de alcance pode levar a erros, uso indevido de dados ou até mesmo vulnerabilidades de segurança, enfatizando a importância de limitar e monitorar cuidadosamente os recursos de IA para garantir um uso seguro e responsável.

Problema nº 7: vazamento de prompts

Estratégias para reduzir o vazamento de prompts

- **Evitar incorporar informações confidenciais em prompts:** nunca inclua credenciais, chaves de API ou lógica proprietária nos prompts. Gerencie-os com segurança fora do sistema.
- **Separar os controles de segurança dos prompts:** gerencie autenticação, autorização e gerenciamento de sessão na lógica da aplicação, não nos prompts.
- **Validar entradas e saídas:** higienize prompts e respostas com validação robusta para bloquear padrões ou manipulações suspeitas.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Criptografar e proteger prompts:** armazene prompts e configurações em armazenamento criptografado e seguro para impedir o acesso não autorizado.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Revisar prompts com regularidade:** revise e limpe os prompts periodicamente para remover dados confidenciais e garantir a conformidade de segurança.
- **Executar testes e simulações de ataque para identificar vulnerabilidades:** realize testes adversariais para identificar e corrigir vulnerabilidades no gerenciamento ou na saída de prompts.
- **Isolar prompts de entradas de usuários:** projete sistemas para evitar que as consultas do usuário manipulem ou exponham prompts.
- **Aplicar limites de taxa:** limite o uso de APIs, acelere atividades suspeitas e bloqueie ataques automatizados de prompts.

Voltar ao Diagrama de arquitetura



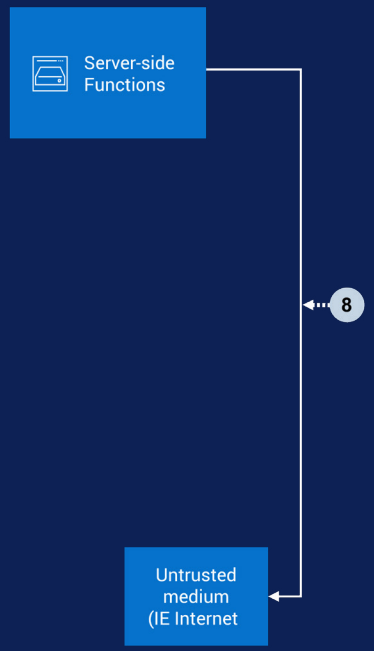
Um ataque por vazamento de prompts do sistema em grandes modelos de linguagem (LLMs) ou sistemas de IA ocorre quando um invasor consegue extrair ou inferir as instruções ocultas — "prompts do sistema" — que orientam o comportamento dos modelos e definem limites operacionais. Normalmente, esses prompts não devem ficar visíveis para os usuários finais, pois contêm regras essenciais, limitações e, às vezes, lógica operacional confidencial. Por meio de entradas especialmente criadas ou da exploração de vulnerabilidades, um invasor pode induzir os LLMs a revelarem os prompts do sistema, no todo ou em parte. Se vazadas, essas informações podem ser usadas para fazer engenharia reversa de restrições, ignorar filtros de segurança ou desenvolver novos ataques direcionados, aumentando, em última análise, o risco de injeção imediata, elevação de privilégios ou uso indevido do modelo e dos sistemas downstream que dependem de sua integridade.

Problema nº 8: vulnerabilidades de incorporações e vetores

Estratégias para reduzir vulnerabilidades de incorporações e vetores

- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Criptografia:** protege dados vetoriais em trânsito e em repouso usando padrões de criptografia robustos, como AES.
- **Configuração e monitoramento de segurança:** reforce os sistemas, configure com segurança e verifique continuamente se há configurações incorretas, acesso não autorizado ou anomalias.
- **Gerenciamento de vulnerabilidades:** atualize e corrija regularmente todos os mecanismos de armazenamento de software, dependências e vetores para lidar com os riscos de segurança.
- **Limpeza de dados e validação de entrada:** analise cuidadosamente as entradas do usuário para remover conteúdo prejudicial. Use normalização e codificação para evitar o uso indevido.
- **Utilizar APIs seguras e interfaces de sistema** para interações de dados de IA, analisando rotineiramente as configurações a fim de minimizar a exposição e a superfície de ataque.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Hardware seguro:** use hardware validado para segurança a fim de evitar vulnerabilidades que possam surgir de ataques baseados em hardware, garantindo uma base sólida para sua infraestrutura.
- **Desenvolvimento, configuração e auditorias seguros:** aplique práticas seguras de codificação, use ferramentas automatizadas de gerenciamento de configurações e realize revisões, auditorias e atualizações regulares para manter as configurações do sistema de IA seguras e atualizadas.

Voltar ao Diagrama de arquitetura



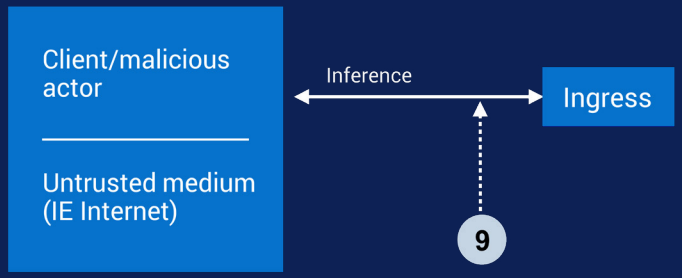
Um ataque por vulnerabilidades de incorporações e vetores contra grandes modelos de linguagem (LLMs) ou sistemas de IA, particularmente aqueles que usam geração aumentada por recuperação (RAG), exploram vulnerabilidades nos mecanismos de codificação, armazenamento e recuperação de informações representadas como vetores numéricos e incorporações. As vulnerabilidades desses mecanismos podem ser exploradas por meio de ações mal-intencionadas, como incorporação de inversão (reconstrução de dados confidenciais a partir de incorporações), envenenamento de dados (injeção de conteúdo prejudicial ou tendencioso para manipular o comportamento dos modelos), acesso não autorizado a bancos de dados vetoriais (levando a vazamentos de dados) ou manipulação de saídas de recuperação. Esses ataques ameaçam a privacidade, a integridade e a confiabilidade, permitindo que invasores divulguem informações confidenciais, alterem resultados ou minem a confiança do usuário em aplicativos orientados por IA. Os controles de acesso adequados, a validação de dados, a criptografia e o monitoramento contínuo são essenciais na defesa contra essas ameaças em evolução.

Problema nº 9: desinformação

Estratégias para reduzir a desinformação

- **Geração aumentada por recuperação (RAG) com fontes autorizadas:** use a RAG para recuperar e integrar informações de bancos de dados confiáveis e verificados e repositórios de conhecimento, reduzindo alucinações.
- **Ajuste de modelos e calibração de saída:** ajuste modelos com diversos conjuntos de dados e aplique técnicas para minimizar o viés e a desinformação.
- **Verificação automatizada de fatos:** faça referência cruzada de saídas com fontes confiáveis e sinalize informações falsas automaticamente.
- **Monitoramento de incertezas:** sinalize respostas de baixa confiança para análise humana em casos críticos.
- **Revisão com intervenção humana:** para aplicações de alto risco, como as de finanças ou saúde, exija supervisão e análise humanas das saídas de modelos para garantir precisão, segurança e proteção.
- **Feedback do usuário:** permita que os usuários relatem erros para melhoria contínua do modelo e correção rápida de caminhos de desinformação.
- **Restrições de acesso e supervisão humana:** aplique controle de acesso baseado em função (RBAC), autenticação baseada em vários fatores (MFA) e gerenciamento de identidades para limitar o acesso. Use a análise humana para tomar decisões críticas.
- **Desenvolvimento, configuração e auditorias seguros:** aplique práticas seguras de codificação, use ferramentas automatizadas de gerenciamento de configurações e realize revisões, auditorias e atualizações regulares para manter as configurações do sistema de IA seguras e atualizadas.
- **Comunicação de riscos:** instrua os usuários sobre as limitações da IA e incentive a verificação independente.
- **Design intencional de IU e API:** destaque o conteúdo gerado por IA e oriente os usuários sobre o uso responsável.

Voltar ao Diagrama de arquitetura



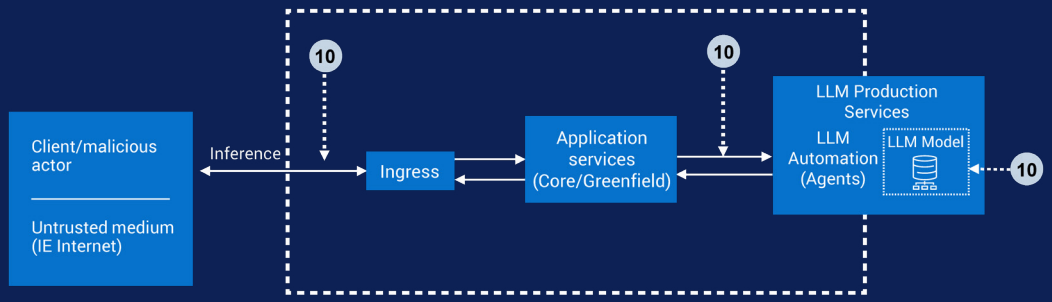
Um ataque por desinformação em um sistema de LLM ou IA é um esforço intencional para fazer com que o modelo gere ou espalhe informações falsas, enganosas ou aparentemente confiáveis, mas incorretas, por meio de seus resultados. Essa vulnerabilidade decorre de vários fatores: tendência de os modelos terem "alucinações" (gerando conteúdo fabricado, mas plausível), vieses ou lacunas presentes nos dados de treinamento e influência de prompts adversários. As alucinações ocorrem porque os LLMs geram estatisticamente texto que se encaixa em um padrão, em vez de realmente entender os fatos, levando a respostas que parecem confiáveis, mas na verdade são infundadas. Os riscos de tais ataques incluem violações de segurança, danos à reputação e até mesmo responsabilidade legal, especialmente em ambientes em que os usuários dependem excessivamente das respostas do LLM sem verificar sua exatidão ou validade, potencialmente incorporando erros ou desinformações em decisões e processos essenciais.

Preocupação nº 10: consumo ilimitado

Estratégias para consumo ilimitado

- **Aplicar limites de taxa e cotas de usuário:** defina limites rígidos para solicitações, tokens ou dados por usuário, chave de API ou aplicativo para evitar abusos.
- **Exigir autenticação e segmentação de usuários:** use autenticação forte (por exemplo, chaves de API, OAuth) e atribua funções ou níveis para processar apenas solicitações autorizadas.
- **Validação de entradas e restrições de tamanho:** valide o tamanho e a estrutura de prompts, bloqueando ou reduzindo consultas grandes ou malformadas.
- **Aplicar limites de tempo de processamento e controle de recursos:** defina tempos limite e restrições de recursos para cada solicitação a fim de evitar operações de longa execução e drenagem de recursos.
- **Implementar cache inteligente e respostas de cache com desduplicação** para consultas duplicadas ou semelhantes a fim de reduzir o processamento desnecessário.
- **Monitoramento, registro e detecção de anomalias:** monitore e registre continuamente as atividades do sistema de IA, usando soluções como MDR/XDR/SIEM, para detectar, investigar e responder rapidamente a acesso não autorizado, anomalias ou vazamentos de dados.
- **Monitoramento de orçamento e controles de gastos:** use painéis e alertas para monitorar custos e bloquear o uso de acordo com os limites de orçamento.
- **Técnicas de sandbox e isolamento:** execute cargas de trabalho em ambientes isolados com permissões limitadas para reduzir riscos.
- **Limitar profundidade de chamadas e etapas de conversação:** imponha limites em chamadas recursivas ou etapas de conversação para evitar exploração.
- **Aplicar alocação de recursos e modelos em camadas:** direcione solicitações de alta prioridade para modelos premium e tráfego de baixa prioridade para modelos econômicos.

Voltar ao Diagrama de arquitetura



Uma ameaça por consumo ilimitado em LLMs ou sistemas de IA se refere a uma vulnerabilidade de segurança em que o aplicativo permite que usuários, mal-intencionados ou não, enviem solicitações ou prompts excessivos e descontrolados de inferência, sem limites de taxa, autenticações ou restrições de uso eficazes. Como a inferência de LLMs é cara do ponto de vista computacional, essa falta de controle pode ser explorada de várias maneiras: os invasores podem causar negação de serviço (DoS) sobrecarregando os recursos do sistema, gerar perdas econômicas imprevistas em implementações hospedadas na nuvem ou de pagamento conforme o uso ou consultar sistematicamente o modelo para clonar seu comportamento e roubar a propriedade intelectual. As consequências incluem interrupção do serviço, redução do desempenho para outros usuários, tensão financeira e maior risco de vazamento de modelos confidenciais. Em essência, o consumo ilimitado ocorre quando o uso de recursos não é controlado adequadamente, deixando as aplicações baseadas em LLM expostas a exploração acidental e deliberada.

Por que escolher a Dell para garantir a segurança de IA?

A Dell ajuda as organizações a proteger modelos de IA e LLMs por meio de uma abordagem completa que abrange hardware, software e serviços gerenciados. A segurança é incorporada desde a cadeia de suprimentos até o dispositivo, a infraestrutura, os dados e os aplicativos, tudo alinhado aos princípios Zero Trust. Em todo o portfólio, as soluções da Dell são criadas para promover a higiene cibernética com recursos como MFA, RBAC, privilégio mínimo e verificação contínua. Essa abordagem abrangente e "segura por padrão" garante que as organizações possam inovar com confiança ao usarem a IA e os LLMs, minimizando o risco de roubo de modelos, vazamento de dados, ataques adversários e outras ameaças cibernéticas avançadas.

Cadeia de suprimentos

A cadeia de suprimentos segura da Dell oferece proteção básica para modelos de IA e LLMs, incorporando a segurança em todos os estágios de desenvolvimento, fabricação e entrega de produtos. Por meio de atualizações de BIOS e firmware assinadas criptograficamente, Secured Component Verification, lista técnica de software (SBOM) focada em IA, rastreamento de linhagem de conjuntos de dados, configuração e software de segurança integrados e rigorosas avaliações de riscos de fornecedores, alinhadas com os padrões globais, a Dell minimiza os riscos de adulteração, acesso não autorizado e ataques à cadeia de suprimentos, garantindo que as organizações possam implementar cargas de trabalho de IA confiáveis e resilientes com total transparência, integridade e conformidade normativa.

AI PCs

A Dell oferece segurança de base para cargas de trabalho de IA no dispositivo. Os Dell Trusted Devices — os AI PCs comerciais mais seguros do mundo* — foram projetados pensando na segurança. A segurança da cadeia de suprimentos minimiza o risco de vulnerabilidades e adulteração do produto. As defesas exclusivas integradas diretamente no hardware e no firmware mantêm o PC e o usuário final protegidos durante o uso. O Dell SafeBIOS oferece visibilidade profunda no nível do BIOS e detecções de violação, enquanto o Dell SafeID aprimora a segurança de credenciais e permite a autenticação sem senha. O software de parceiros oferece proteção avançada em ambientes de endpoint, rede e nuvem.

Resiliência cibernética

As soluções de resiliência cibernética PowerProtect da Dell protegem os dados de IA com backups criptografados e imutáveis, restauração rápida e cofres isolados de recuperação cibernética. Esses recursos impedem a destruição, reduzem o impacto de atualizações mal-intencionadas e dão suporte à conformidade e à recuperação após um ataque.

Servidores

Os servidores PowerEdge apresentam computação confidencial para isolar e proteger prompts e incorporações de IA/LLM, soluções confiáveis de geração aumentada por recuperação (RAG) ancoradas em fontes autorizadas, juntamente com MFA, RBAC, raiz de confiança de silício, firmware assinado e monitoramento contínuo, para proteger cargas de trabalho de IA essenciais.

Armazenamento

O portfólio Dell Storage garante armazenamento seguro e criptografado para dados confidenciais de IA com criptografia AES-256 robusta para dados em repouso e em trânsito. A criptografia avançada projetada para ser resiliente contra futuras ameaças quânticas está disponível em ofertas selecionadas. O portfólio inclui desempenho de NVMe de alta velocidade, módulos de criptografia em conformidade com o FIPS para proteger dados, incluindo aqueles utilizados em cargas de trabalho de IA, snapshots imutáveis e cofres de recuperação cibernética com air gap para combater ataques de ransomware. A arquitetura Zero Trust, a segurança da cadeia de suprimentos e os recursos de auditoria à prova de adulterações aprimoram a governança. Os modelos integrados de detecção de anomalias e AIOps ML protegem as cargas de trabalho sem usar dados do cliente para treinamento, minimizando assim os riscos de ataque baseados em entradas.

AIOps

O Dell AIOps oferece monitoramento automatizado e contínuo para detectar configurações incorretas e vulnerabilidades (incluindo CVEs) e oferece suporte à conscientização sobre riscos da cadeia de suprimentos que afetam as cargas de trabalho de IA/LLM. A verificação de CVE em tempo real, os alertas inteligentes e os painéis de indicadores com IA permitem uma intervenção rápida ao sinalizarem anomalias e rastrearem fluxos de trabalho de resolução. Os recursos de conformidade integrados, os controles de acesso baseados em função e os relatórios automatizados ajudam a manter as operações seguras em todas as cargas de trabalho, enquanto a integração contínua de EDR/XDR e os insights operacionais orientados por IA, incluindo recursos generativos em soluções compatíveis, aumentam ainda mais a eficiência da TI.

Sistema de rede

As soluções Dell Networking protegem ambientes de IA/LLM por meio de segmentação de rede robusta, minimizando o movimento lateral. Os caminhos de rede criptografados e os controles integrados de firewall bloqueiam o acesso não autorizado aos dados de IA.

Serviços de segurança e resiliência da IA

Os serviços de segurança e resiliência de IA da Dell são projetados para lidar com os novos riscos associados à integração da IA em sua organização. Criados para trabalhar com suas equipes à medida que você integra a IA o mais rápido possível, nossos serviços oferecem a experiência para orientar no planejamento estratégico, na implementação de soluções e nos serviços de segurança gerenciados para aliviar as cargas operacionais, para que você possa inovar com segurança com a IA. Cada um deles é personalizado para ajudar as organizações a lidar com os riscos de IA em evolução e otimizar implementações seguras de IA.

Dell AI Factory

Um portfólio integrado de segurança para fins específicos, como cadeia de suprimentos segura da Dell, recursos Zero Trust para impor privilégios mínimos e soluções de MDR de IA projetadas para manter seu modelo seguro e protegido.

*Com base em uma análise interna da Dell, de outubro de 2024 (Intel) e março de 2025 (AMD). Aplicável a PCs com processadores Intel e AMD. Nem todos os recursos estão disponíveis em todos os PCs. Compra adicional necessária para alguns recursos. PCs com tecnologia Intel, validados pela Principled Technologies, julho de 2025.

Conclusão

Para criar estruturas de IA resilientes, é essencial uma abordagem colaborativa entre organizações e especialistas em segurança. À medida que a IA e os LLMs continuam a remodelar os setores, é essencial lidar com os riscos que eles trazem, incluindo segurança de dados, integridade do modelo e desafios de conformidade. As organizações devem priorizar estratégias proativas que integrem a segurança em cada estágio de sua jornada de IA.

A Dell Technologies se destaca como um parceiro confiável nessa missão, oferecendo personalização completa da IA generativa, consultoria de segurança e soluções integradas adaptadas às suas necessidades exclusivas. Ao aproveitar as soluções robustas de segurança cibernética da Dell, as empresas podem reduzir com eficiência os riscos de IA e LLM e maximizar o potencial de seus investimentos em segurança existentes. A Dell capacita as organizações a proteger sua infraestrutura de IA integrando perfeitamente a segurança avançada às estruturas atuais, garantindo um ambiente seguro e pronto para o futuro.

Saiba como as soluções abrangentes de IA da Dell podem proteger seus ambientes de IA generativa e LLM: Dell.com/CyberSecurityMonth