DELL Technologies

10 principales préoccupations en matière de cybersécurité relatives à l'IA générative et aux LLM

Démarrer

Introduction

L'intelligence artificielle (IA) révolutionne le mode de fonctionnement des entreprises. L'IA générative (GenAI) et les grands modèles de langage (LLM) deviennent des charges applicatives critiques dans les environnements d'entreprise modernes.

À l'instar de toute autre charge applicative, ces applications présentent leurs propres complexités et vulnérabilités. À mesure que les entreprises continuent d'adopter l'IA pour stimuler l'innovation, l'efficacité et l'avantage concurrentiel, garantir la sécurité de ces applications devient une nécessité fondamentale. La protection de toute charge applicative passe par une bonne hygiène sur le plan de la cybersécurité. Et de la même manière que la sécurité de toutes les charges applicatives vous semblerait importante, ces bonnes pratiques d'hygiène s'appliquent aussi à l'IA. Cela inclut la mise en œuvre de pratiques telles que l'application de correctifs système appropriés, l'authentification multifacteur, l'accès basé sur les rôles et la segmentation du réseau. Ces mesures sont fondamentales, mais la clé consiste à comprendre comment ces fonctionnalités s'intègrent à l'architecture spécifique et à l'utilisation de votre charge applicative.

Chez Dell, nous avons une connaissance approfondie de la charge applicative liée à l'IA et des défis de sécurité uniques auxquels elle est confrontée. En identifiant les façons dont les acteurs de la menace peuvent cibler ces charges applicatives, Dell peut vous aider à créer une stratégie de sécurité robuste. Pour ce faire, il faut s'intéresser aux risques suivants : corruption des données d'entraînement, vol ou manipulation des modèles, reconstruction des jeux de données, etc.

Nous faisons également particulièrement attention aux difficultés associées aux données saisies dans votre modèle d'IA, en prévenant la divulgation de données sensibles, en atténuant les sujets dangereux ou les biais, et en assurant la conformité aux réglementations. En ce qui concerne les résultats, nous aidons à résoudre des problèmes tels que la dépendance excessive vis-à-vis du modèle et les risques liés à la conformité.

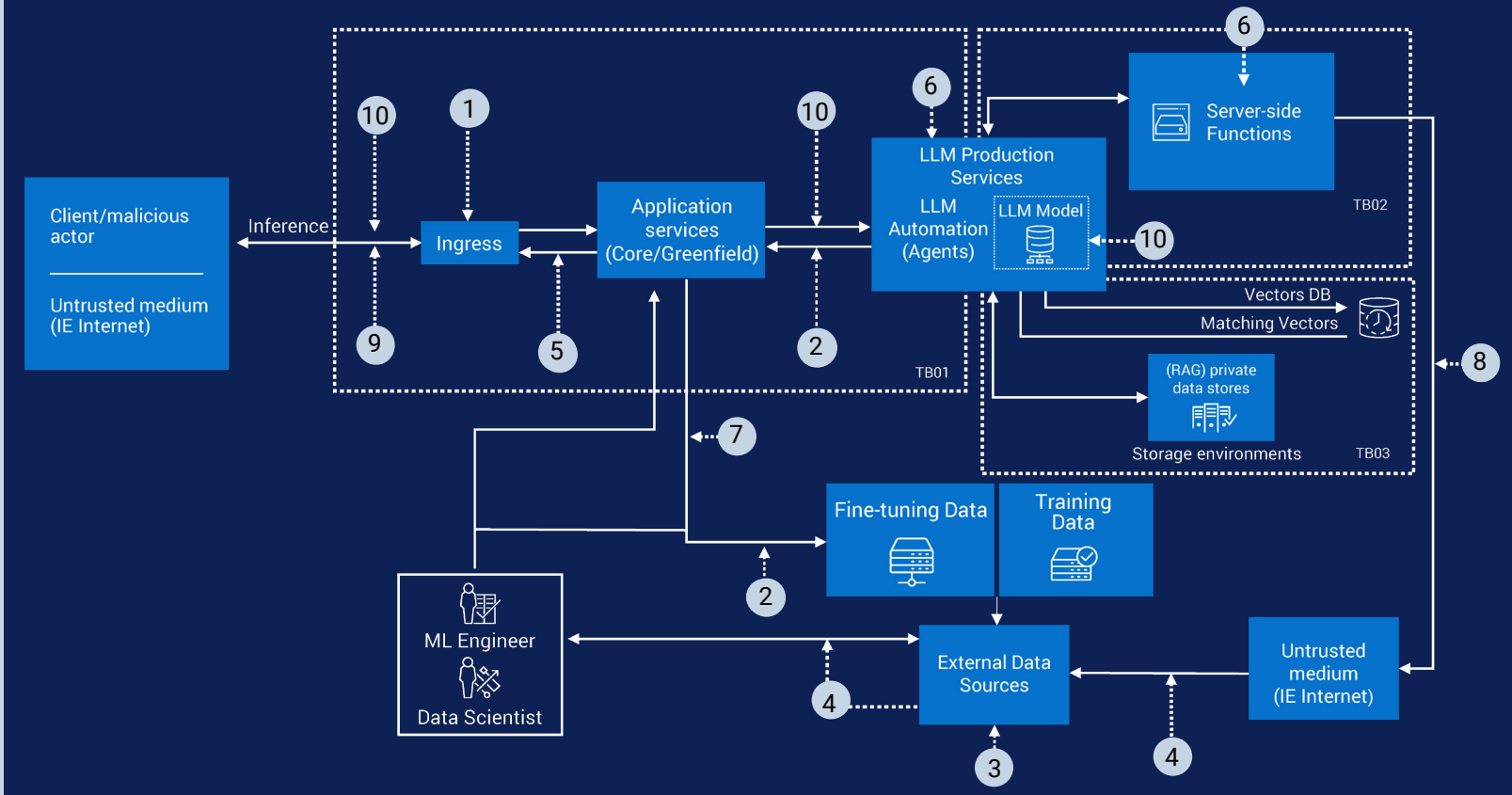
Chez Dell, nous donnons aux entreprises les moyens de limiter ces risques en tirant parti de leurs solutions de cybersécurité existantes ou en explorant de nouveaux outils et pratiques pour protéger leurs systèmes. Notre objectif est de nous assurer que la sécurité n'entrave pas votre innovation. En comprenant le fonctionnement des charges applicatives de l'IA et les menaces de sécurité auxquelles elles sont confrontées, nous pouvons vous aider à renforcer votre posture de sécurité, à rendre votre environnement plus résilient tout en vous permettant d'innover en toute confiance. Forts de notre expertise, nous vous aidons à exploiter en toute confiance le potentiel de l'IA tout en maintenant une sécurité robuste.

10 principales préoccupations en matière de cybersécurité relatives à l'IA générative et aux LLM

Ce sont les principales préoccupations en matière de protection des modèles d'IA générative/LLM, comme l'indique l'OWASP.

Cliquez sur une préoccupation pour en savoir plus :

1. Infiltration de requête
2. Divulgation de données sensibles
3. Chaîne d'approvisionnement
4. Corruption des données du modèle
5. Mauvaise manipulation des résultats
6. Confiance excessive
7. Fuite de prompt système
8. Faiblesses vectorielles et d'intégration
9. Désinformation
10. Consommation débridée

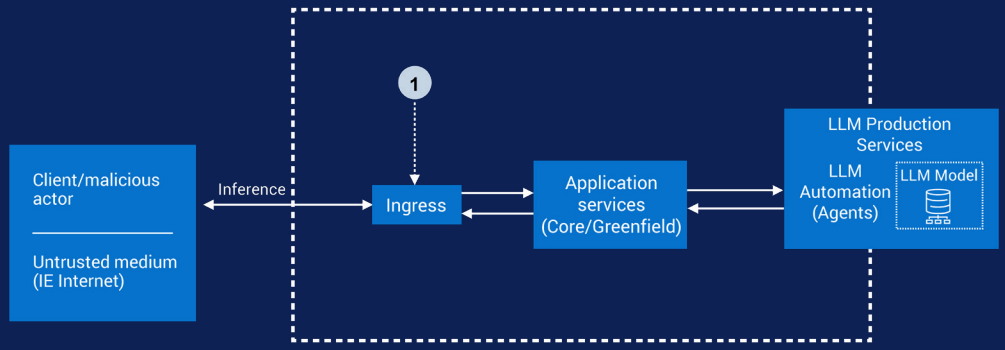


Préoccupation n° 1 : Infiltration de requête

Stratégies d'atténuation de l'infiltration de requête :

- **Nettoyage des données et validation des entrées :** effectuez un contrôle minutieux des entrées utilisateur pour supprimer les contenus nuisibles. Utilisez la normalisation et l'encodage pour éviter toute utilisation abusive.
- **Traitement du langage naturel et approches basées sur l'apprentissage automatique :** utilisez le traitement du langage naturel et l'apprentissage automatique pour détecter et bloquer les prompts manipulateurs ou malveillants.
- **Formatage des résultats et contrôle des réponses clairs :** posez des limites de réponse strictes pour vous assurer que les résultats respectent les formats prévus et empêcher les actions non autorisées. Utilisez le filtrage des prompts et la validation des réponses pour en préserver l'intégrité.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Conception sécurisée de prompts :** utilisez un modèle de prompt et d'analyse sécurisé dans le cadre de la sécurité logicielle globale pour protéger le traitement des données saisies.
- **Validation des modèles :** validez régulièrement les modèles ML pour garantir qu'ils n'ont pas été altérés avant le déploiement, ce qui garantit leur précision et leur intégrité.
- **Filtrage des prompts, classement et validation des réponses :** analysez et classez les prompts pour vous assurer que seules les demandes sécurisées sont traitées. Validez les réponses pour éviter toute utilisation abusive.
- **Contrôles de robustesse :** effectuez des évaluations régulières pour identifier et corriger les vulnérabilités, afin de garantir la sécurité et la fiabilité de l'IA.

Retour au diagramme d'architecture



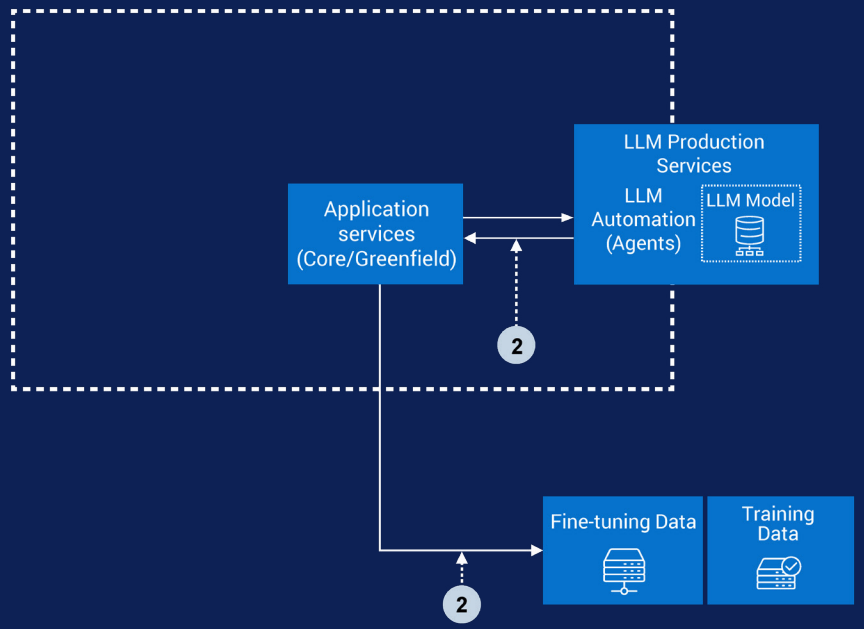
L'infiltration de requête fait partie des nouveaux défis avec lesquels le monde de l'IA générative (GenAI) doit composer, et qui consiste à concevoir des prompts malveillants visant à manipuler le comportement du modèle ou à compromettre son intégrité. Ces attaques exploitent les vulnérabilités présentes dans la manière dont les systèmes d'IA traitent et réagissent aux prompts des utilisateurs, ce qui peut entraîner des actions non autorisées, de la désinformation ou la divulgation de données sensibles. À mesure que l'IA générative s'intègre de plus en plus dans les workflows stratégiques de l'entreprise, il est essentiel de gérer ces risques pour maintenir la confiance et la sécurité

Préoccupation n° 2 : Divulgation de données sensibles

Stratégies d'atténuation de la divulgation de données sensibles :

- **Nettoyage des données et validation des entrées :** effectuez un contrôle minutieux des entrées utilisateur pour supprimer les contenus nuisibles. Utilisez la normalisation et l'encodage pour éviter toute utilisation abusive.
- **Utilisez le chiffrement homomorphe** pour traiter les données sensibles en toute sécurité sans exposer leur contenu. Même lorsque les données sont en cours d'utilisation, elles restent ainsi chiffrées et protégées contre les violations de sécurité.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Utilisez des API et des interfaces système sécurisées** pour interagir avec les données d'IA, en examinant régulièrement les configurations afin de limiter l'exposition et la surface d'attaque.
- **Sécurisez la collecte, le stockage et les stratégies des données** et appliquez des stratégies complètes de gouvernance et de protection des données qui garantissent la conformité aux réglementations et minimisent les risques liés aux données.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Développement, configuration et audits sécurisés :** appliquez des pratiques de codage sécurisées, utilisez des outils de gestion de la configuration automatisés et effectuez régulièrement des examens, des audits et des mises à jour pour assurer la sécurité et la mise à jour des configurations des systèmes d'IA.
- **Formation des utilisateurs et sensibilisation à la sécurité :** proposez une formation continue de sensibilisation à la sécurité spécifique à l'IA aux utilisateurs et aux administrateurs afin de réduire les utilisations dangereuses et la divulgation accidentelle de données.

Retour au diagramme d'architecture



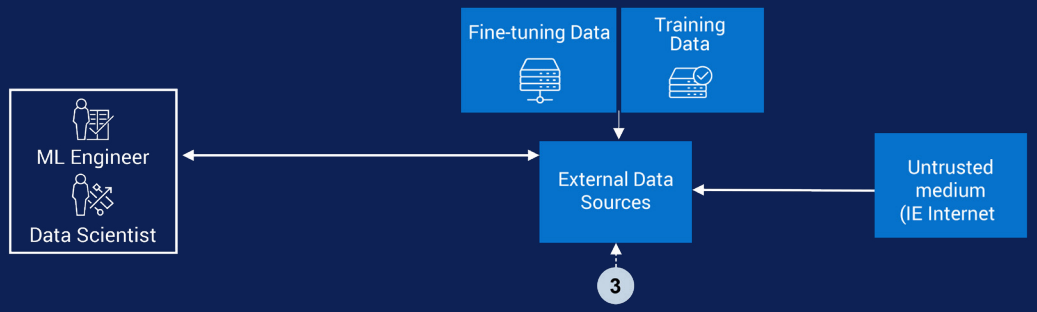
L'IA générative a apporté des avancées incroyables, mais elle comporte également des risques importants, en particulier l'exposition involontaire de données sensibles. Qu'il s'agisse d'informations personnelles identifiables ou de données commerciales propriétaires, l'utilisation abusive ou la mauvaise manipulation des outils d'IA générative peut entraîner des fuites de données, une non-conformité réglementaire ou une atteinte à la réputation. Il est donc essentiel pour les entreprises de comprendre ces risques et de les traiter de manière proactive afin de garantir une mise en œuvre et une utilisation sécurisées des systèmes d'IA.

Préoccupation n° 3 : Vulnérabilités de la chaîne logistique

Stratégies d'atténuation des vulnérabilités de la chaîne logistique :

- **Contrôle des fournisseurs et garantie de la conformité aux pratiques sécurisées de la chaîne logistique** : évaluez les fournisseurs et instaurez des contrats qui mettent l'accent sur la sécurité de la chaîne logistique.
- **Mise en œuvre d'une nomenclature logicielle :** Tsuivez et vérifiez l'origine des composants logiciels pour plus de transparence et pour réduire les risques de compromission du code.
- **Validation des modèles :** validez régulièrement les modèles ML pour garantir qu'ils n'ont pas été altérés avant le déploiement, ce qui garantit leur précision et leur intégrité.
- **Adoption du principe du moindre privilège dans les conteneurs et les pods :** ce principe limite l'impact d'une éventuelle compromission et restreint les accès non autorisés.
- **Déploiement des pare-feu :** bloquez la connectivité réseau inutile, réduisant ainsi l'exposition aux menaces potentielles et limitant les possibilités d'attaque.
- **Protection des données et des annotations :** sécurisez vos données et les annotations associées afin d'éviter toute altération, tout accès non autorisé et toute corruption des informations critiques.
- **Matériel sécurisé :** utilisez du matériel validé pour la sécurité afin de prévenir les vulnérabilités qui pourraient survenir suite à des attaques matérielles, garantissant ainsi une base solide à votre infrastructure.
- **Composants logiciels ML sécurisés :** utilisez des composants logiciels ML fiables et approuvés pour réduire les vulnérabilités et renforcer la sécurité globale de vos workflows d'apprentissage automatique.
- **Développement, configuration et audits sécurisés :** appliquez des pratiques de codage sécurisées, utilisez des outils de gestion de la configuration automatisés et effectuez régulièrement des examens, des audits et des mises à jour pour assurer la sécurité et la mise à jour des configurations des systèmes d'IA.

Retour au diagramme d'architecture



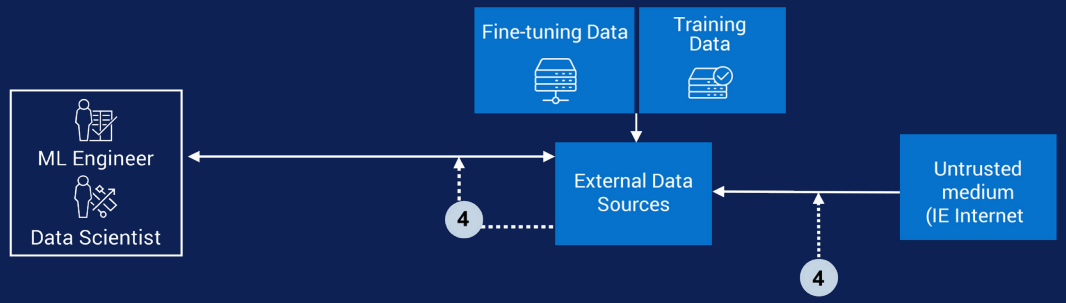
Étudiez les vulnérabilités de la chaîne logistique du LLM, qui peuvent affecter des composants critiques tels que l'intégrité des modèles préentraînés et les adaptateurs tiers. Les systèmes d'intelligence artificielle reposent sur des composants matériels et logiciels qui peuvent être compromis bien avant leur déploiement. Les différentes étapes de la chaîne logistique de l'apprentissage automatique sont à la merci des adversaires, qui ciblent les processeurs graphiques, les données et leurs annotations, les éléments de l'infrastructure logicielle ML ou même le modèle lui-même. En compromettant des portions du système, les attaquants mettent un pied dans la porte, ce qui présente des risques importants de sécurité et d'intégrité. Comprendre et atténuer ces vulnérabilités est essentiel pour créer des solutions d'IA robustes et sécurisées.

Préoccupation n° 4 : Corruption des données du modèle

Stratégies d'atténuation de la corruption des données du modèle :

- **Utilisez la détection des anomalies et la validation des données pendant la formation** pour identifier et corriger les incohérences dans les données et s'assurer que seules des données propres et de haute qualité sont utilisées pour entraîner le modèle.
- **Isolez les environnements pendant les phases de réglage fin** afin d'éviter tout accès non autorisé ou toute contamination du modèle pendant les étapes critiques du développement.
- **Validation des modèles :** validez régulièrement les modèles ML pour garantir qu'ils n'ont pas été altérés avant le déploiement, ce qui garantit leur précision et leur intégrité.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Nettoyage des données et validation des entrées :** effectuez un contrôle minutieux des entrées utilisateur pour supprimer les contenus nuisibles. Utilisez la normalisation et l'encodage pour éviter toute utilisation abusive.
- **Développement, configuration et audits sécurisés :** appliquez des pratiques de codage sécurisées, utilisez des outils de gestion de la configuration automatisés et effectuez régulièrement des examens, des audits et des mises à jour pour assurer la sécurité et la mise à jour des configurations des systèmes d'IA.
- **Contrôles de robustesse :** effectuez des évaluations régulières pour identifier et corriger les vulnérabilités, afin de garantir la sécurité et la fiabilité de l'IA.
- **Déployez la segmentation du réseau** pour limiter l'accès aux interfaces non sécurisées et aux composants système critiques.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.

Retour au diagramme d'architecture



La corruption des données du modèle est une menace de sécurité dans le cycle de vie de l'IA, qui consiste à contaminer intentionnellement les données d'entraînement via des prompts corrompus, trompeurs ou malveillants. Ce risque peut affecter des composants critiques, allant de la collecte de données brutes et de l'annotation à la conservation et à l'intégration de jeux de données utilisés pour l'apprentissage automatique ou les grands modèles de langage. La fiabilité des systèmes d'IA dépend de l'intégrité de leurs sources de données, qui peuvent être exposées à des manipulations avant l'entraînement, pendant le prétraitement ou via des pipelines de données externes.

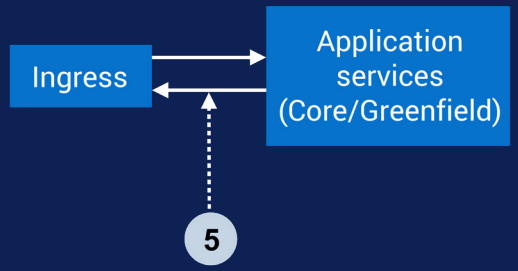
Les attaquants ont recours à la corruption des données pour nuire à la précision des modèles, introduire des vulnérabilités ou générer des résultats néfastes. En ciblant les faiblesses dans la provenance des données, la qualité des annotations ou les processus d'ingestion des jeux de données, les attaquants peuvent compromettre la sécurité, la fiabilité et la résilience. Reconnaître et atténuer ces menaces centrées sur les données est essentiel pour créer des solutions d'IA robustes et fiables.

Préoccupation n° 5 : Mauvaise manipulation des résultats

Stratégies d'atténuation de la mauvaise manipulation des résultats :

- **Encodage des résultats sensible au contexte :** appliquez toujours des techniques d'encodage et d'échappement adaptées au contexte spécifique dans lequel le résultat sera utilisé, comme les environnements HTML, SQL ou API, afin de prévenir les attaques par injection, par exemple.
- **Nettoyage des résultats :** suivez des pratiques strictes de validation et de nettoyage pour les résultats du modèle, conformément aux directives Application Security Verification Standard (ASVS) d'Open Web Application Security Project (OWASP) afin de garantir une utilisation en aval sécurisée et de limiter les risques de sécurité.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Tests automatisés de sécurité des résultats :** effectuez des tests de sécurité réguliers à l'aide d'outils automatisés pour identifier les risques liés aux résultats, comme le cross-site scripting (XSS) ou l'infiltration, et y répondre de manière proactive.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Intervention humaine :** les applications à haut risque, comme la finance ou la santé, nécessitent une surveillance humaine et une vérification des résultats du modèle afin d'en garantir la précision, la sécurité et la sûreté.
- **Confidentialité et conformité :** intégrez des techniques de protection de la confidentialité dans le processus de sortie et assurez-vous de la conformité aux réglementations et normes applicables pour une utilisation sécurisée des données sensibles.

Retour au diagramme d'architecture



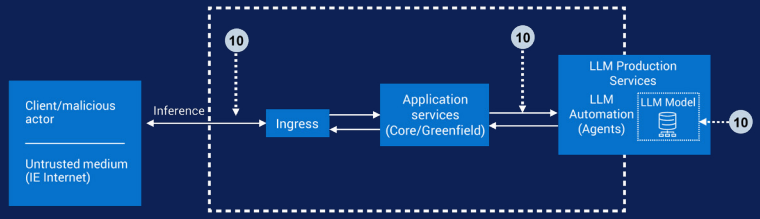
Une validation ou un nettoyage insuffisants des résultats du modèle d'IA peut entraîner de graves risques de sécurité, notamment une escalade des privilèges et des violations de données. Lorsque les modèles d'IA produisent des résultats qui ne sont pas correctement vérifiés ou filtrés, des acteurs malveillants peuvent exploiter ces vulnérabilités pour obtenir un accès non autorisé ou augmenter leurs privilèges au sein d'un système. Ce manque de supervision peut entraîner la compromission des données, des actions non autorisées et des violations de la sécurité importantes, ce qui souligne l'importance de mettre en œuvre des processus de validation et de nettoyage robustes pour tous les résultats générés par l'IA.

Préoccupation n° 6 : Confiance excessive

Stratégies d'atténuation de la confiance excessive

- **Adoption du principe du moindre privilège :** n'accordez aux sous-systèmes des LLM et agentiques que les autorisations minimales requises pour effectuer les opérations prévues et vérifiez régulièrement les contrôles d'accès.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Limites opérationnelles :** définissez clairement les sections auxquelles les LLM/agents peuvent accéder ou qu'ils peuvent exécuter.
- **Intervention humaine :** les applications à haut risque, comme la finance ou la santé, nécessitent une surveillance humaine et une vérification des résultats du modèle afin d'en garantir la précision, la sécurité et la sûreté.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Limites de l'autonomie :** restreignez les capacités du LLM pour éviter un accès ou un contrôle illimité.
- **Développement, configuration et audits sécurisés :** appliquez des pratiques de codage sécurisées, utilisez des outils de gestion de la configuration automatisés et effectuez régulièrement des examens, des audits et des mises à jour pour assurer la sécurité et la mise à jour des configurations des systèmes d'IA.
- **Déploiement des pare-feu** : bloquez la connectivité réseau inutile, réduisant ainsi l'exposition aux menaces potentielles et limitant les possibilités d'attaque.
- **Contrôles de robustesse :** effectuez des évaluations régulières pour identifier et corriger les vulnérabilités, afin de garantir la sécurité et la fiabilité de l'IA.

Retour au diagramme d'architecture



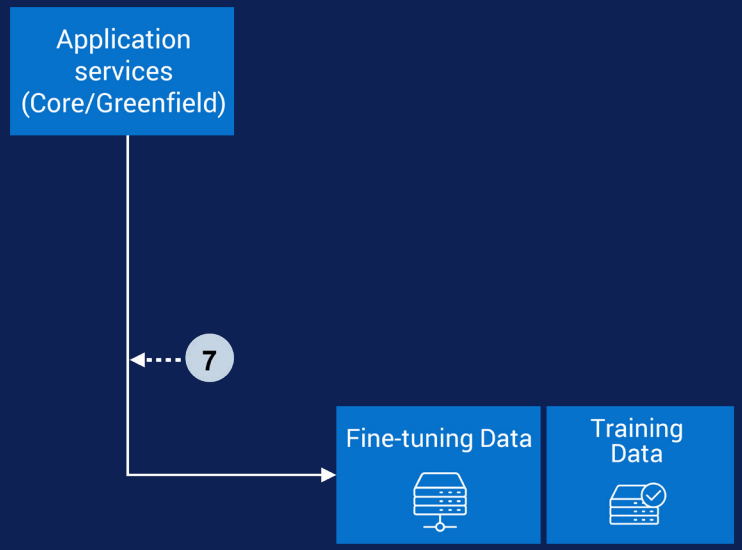
Accorder aux agents ou plug-ins IA une autonomie excessive ou des fonctionnalités inutiles au sein des workflows peut présenter des risques importants. Lorsque les privilèges ou les capacités d'un système d'IA sont plus poussés que nécessaire, la probabilité de conséquences involontaires augmente. Cela peut se produire lorsque les systèmes basés sur un grand modèle de langage (LLM) sont conçus avec des autorisations excessives, ce qui leur permet de prendre des mesures ou d'accéder à des informations auxquelles ils ne devraient pas accéder. Une telle ingérence peut entraîner des erreurs, une mauvaise utilisation des données ou même des failles de sécurité, ce qui souligne l'importance de limiter et de surveiller soigneusement les capacités d'IA pour garantir une utilisation sûre et responsable.

Préoccupation n° 7 : Fuite de prompt

Stratégies d'atténuation des fuites de prompt

- **Prévention de l'inclusion des données sensibles dans les prompts** : n'incluez jamais d'informations d'identification, de clés API ou de logique propriétaire dans les prompts : gérez ces éléments en toute sécurité en dehors du système.
- **Séparation des contrôles de sécurité des prompts** : gérez l'authentification, l'autorisation et la gestion des sessions dans la logique d'application, et non dans les prompts.
- **Validation des prompts et des résultats** : nettoyez les prompts et les réponses grâce à une validation robuste, capable de bloquer les schémas ou manipulations suspects.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Prompts chiffrés et sécurisés :** stockez les prompts et les configurations dans un espace de stockage sécurisé et chiffré pour empêcher tout accès non autorisé.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Contrôle régulier des prompts** : passez régulièrement en revue et nettoyez les prompts afin de supprimer les données sensibles et de garantir la conformité de la sécurité.
- **Tests d'intrusion** : réalisez des tests de piratage pour identifier et corriger les vulnérabilités dans la gestion des prompts ou les résultats.
- **Isolement des prompts des entrées utilisateurs** : créez des systèmes qui empêchent les requêtes des utilisateurs de manipuler ou d'exposer les prompts.
- **Définition des limites d'utilisation** : limitez l'utilisation des API, restreignez les activités suspectes et bloquez les attaques automatisées par prompt.

Retour au diagramme d'architecture



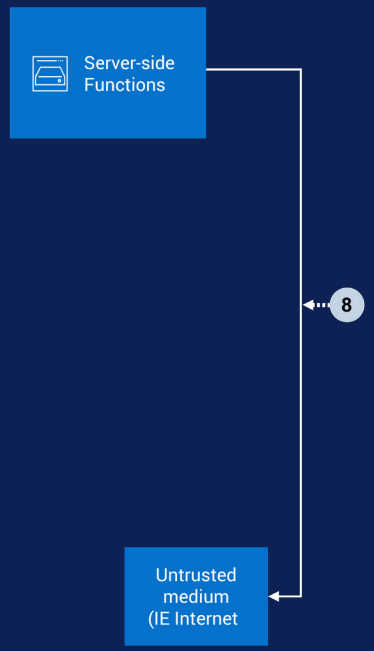
Une attaque par fuite de prompt système sur un grand modèle de langage (LLM) ou un système d'IA se produit lorsqu'un attaquant est en mesure d'extraire ou de deviner les instructions cachées (les « prompts système ») qui guident le comportement du modèle et définissent ses limites opérationnelles. Ces prompts ne sont généralement pas destinées à être visibles par les utilisateurs finaux, car elles contiennent des règles essentielles, des limitations et parfois une logique opérationnelle sensible. À l'aide de prompts très spécifiques ou en exploitant ses vulnérabilités, un attaquant peut inciter le LLM à révéler une partie ou l'intégralité de son prompt système. En cas de fuite, ces informations peuvent être utilisées pour analyser les restrictions par rétro-ingénierie, contourner les filtres de sécurité ou développer de nouvelles attaques ciblées, augmentant à terme le risque d'infiltration de requête, d'escalade des privilèges ou d'utilisation abusive du modèle et des systèmes en aval qui dépendent de son intégrité.

Préoccupation n° 8 : Faiblesses vectorielles et d'intégration

Stratégies d'atténuation des faiblesses vectorielles et d'intégration

- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Chiffrement** : sécurisez les données vectorielles en transit et au repos à l'aide de normes de chiffrement robustes telles que AES.
- **Configuration et surveillance sécurisées** : renforcez les systèmes, configurez-les en toute sécurité et surveillez en permanence les erreurs de configuration, les accès non autorisés ou les anomalies.
- **La gestion des vulnérabilités** met à jour et corrige régulièrement tous les logiciels, dépendances et moteurs de stockage vectoriel pour faire face aux risques de sécurité.
- **Nettoyage des données et validation des entrées :** effectuez un contrôle minutieux des entrées utilisateur pour supprimer les contenus nuisibles. Utilisez la normalisation et l'encodage pour éviter toute utilisation abusive.
- **Utilisez des API et des interfaces système sécurisées** pour interagir avec les données d'IA, en examinant régulièrement les configurations afin de limiter l'exposition et la surface d'attaque.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Matériel sécurisé :** utilisez du matériel validé pour la sécurité afin de prévenir les vulnérabilités qui pourraient survenir suite à des attaques matérielles, garantissant ainsi une base solide à votre infrastructure.
- **Développement, configuration et audits sécurisés :** appliquez des pratiques de codage sécurisées, utilisez des outils de gestion de la configuration automatisés et effectuez régulièrement des examens, des audits et des mises à jour pour assurer la sécurité et la mise à jour des configurations des systèmes d'IA.

Retour au diagramme d'architecture



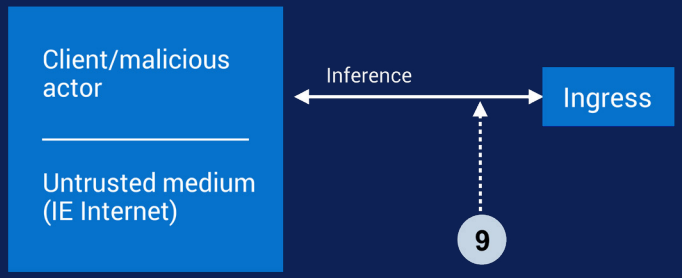
Les attaques ciblant les faiblesses vectorielles et d'intégration sur un grand modèle de langage (LLM) ou un système d'IA, en particulier ceux utilisant la RAG (Retrieval-Augmented Generation), ciblent les vulnérabilités d'encodage, de stockage et de récupération des informations sous la forme de vecteurs numériques et d'intégrations. Les faiblesses de ces mécanismes peuvent être exploitées par des actions malveillantes, comme l'inversion d'intégration (reconstruction de données sensibles à partir d'intégrations), la corruption des données (injection de contenu nuisible ou biaisé pour manipuler le comportement du modèle), l'accès non autorisé à des bases de données vectorielles (entraînant des fuites de données) ou la manipulation des résultats de récupération. Ces attaques menacent la confidentialité, l'intégrité et la fiabilité en permettant aux attaquants de divulguer des données sensibles, d'altérer les résultats ou de saper la confiance des utilisateurs dans les applications basées sur l'IA. Des contrôles d'accès appropriés, la validation des données, le chiffrement et la surveillance continue sont essentiels pour se défendre contre ces menaces en constante évolution.

Préoccupation n° 9 : Désinformation

Stratégies pour atténuer la désinformation

- **RAG (Retrieval-Augmented Generation) avec des sources faisant autorité :** utilisez la RAG pour récupérer et intégrer des informations à partir de bases de données et de référentiels de connaissances vérifiés et fiables, qui réduisent le nombre d'hallucinations.
- **Modèle et étalonnage de sortie :** ajustez avec précision les modèles avec divers jeux de données et appliquez des techniques pour minimiser les biais et la désinformation.
- **Vérification automatisée des faits** : comparez les résultats avec des sources fiables et signalez automatiquement les informations erronées.
- **Surveillance des incertitudes :** signalez les réponses peu fiables pour les faire vérifier par un humain dans les cas les plus critiques.
- **Intervention humaine :** les applications à haut risque, comme la finance ou la santé, nécessitent une surveillance humaine et une vérification des résultats du modèle afin d'en garantir la précision, la sécurité et la sûreté.
- **Commentaires des utilisateurs :** permettez aux utilisateurs de signaler les erreurs pour améliorer continuellement le modèle et corriger rapidement les biais menant à la désinformation.
- **Restrictions des accès et supervision humaine :** appliquez le contrôle d'accès basé sur les rôles (RBAC), l'authentification multifacteur (MFA) et la gestion des identités pour limiter l'accès. Faites intervenir un humain en cas de prise de décision critique.
- **Développement, configuration et audits sécurisés :** appliquez des pratiques de codage sécurisées, utilisez des outils de gestion de la configuration automatisés et effectuez régulièrement des examens, des audits et des mises à jour pour assurer la sécurité et la mise à jour des configurations des systèmes d'IA.
- **Communication sur les risques :** éduquez les utilisateurs sur les limites de l'IA et encouragez la vérification indépendante.
- **Conception intentionnelle de l'interface utilisateur et de l'API :** mettez en évidence le contenu généré par l'IA et orientez les utilisateurs vers une utilisation responsable.

Retour au diagramme d'architecture



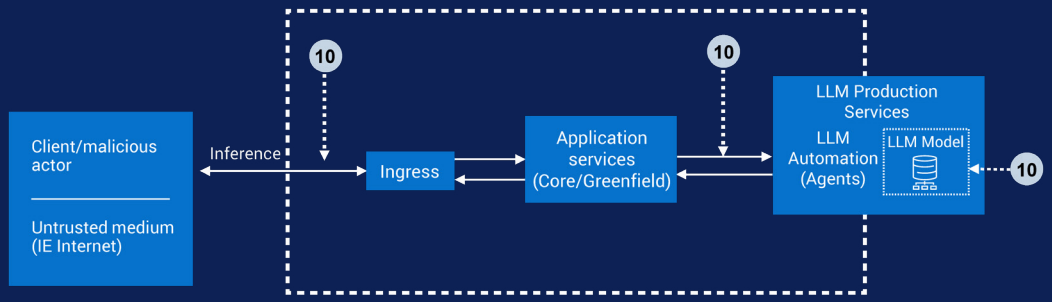
Une attaque par désinformation sur un LLM ou un système d'IA est un effort intentionnel visant à pousser le modèle à générer ou diffuser des informations fausses, trompeuses ou crédibles en apparence, mais incorrectes, dans ses résultats. Cette vulnérabilité provient de plusieurs facteurs : la tendance du modèle à « halluciner » (c'est-à-dire à générer de toutes pièces du contenu pourtant plausible), les biais ou les lacunes présents dans les données d'entraînement et l'influence des prompts contradictoires. Les hallucinations surviennent parce que les LLM génèrent statistiquement un texte qui correspond à un modèle plutôt que de comprendre réellement les faits, ce qui peut générer des réponses qui semblent fiables mais qui ne reposent sur aucune source fondée. Ces attaques présentent des risques de violations de la sécurité, d'atteinte à la réputation et même de responsabilité juridique, en particulier dans les environnements où les utilisateurs accordent une confiance aveugle aux réponses du LLM sans en vérifier l'exactitude ou la validité, ce qui peut entraîner des erreurs ou de la désinformation dans les décisions et processus critiques.

Préoccupation n° 10 : Consommation débridée

Stratégies pour une consommation débridée

- **Mise en place d'une limite de débit et de quotas par utilisateur** : définissez des limites strictes de demandes, de jetons ou de données par utilisateur, clé API ou application afin d'éviter tout abus.
- **Authentification et segmentation des utilisateurs** : utilisez une authentification forte (par exemple, clés API, OAuth) et assignez des rôles ou des niveaux pour traiter uniquement les demandes autorisées.
- **Validation des prompts et restrictions de taille** : validez la taille et la structure des prompts pour bloquer ou réduire les requêtes volumineuses ou mal formulées.
- **Gestion des délais de traitement et limitation des ressources** : définissez des délais d'expiration et des plafonds de ressources pour chaque demande afin d'éviter les opérations interminables et la surconsommation de ressources.
- **Déployez des réponses intelligentes de mise en cache et de déduplication** pour les requêtes dupliquées ou similaires afin de réduire les traitements inutiles.
- **Surveillance, journalisation et détection des anomalies :** surveillez et consignez en continu les activités des systèmes d'IA à l'aide de solutions MDR/XDR/SIEM, afin de détecter rapidement les accès non autorisés, les anomalies ou les fuites de données, d'enquêter et d'y répondre.
- **Suivi du budget et contrôle des dépenses** : utilisez des tableaux de bord et des alertes pour surveiller les coûts et bloquer l'utilisation aux seuils budgétaires.
- **Bacs à sable et techniques d'isolation** : exécutez des charges applicatives dans des environnements isolés avec des permissions limitées, ce qui permet de réduire les risques.
- **Limite de la profondeur des appels et des échanges** : limitez les appels récursifs ou mettez en place des étapes de conversation afin d'éviter toute exploitation.
- **Application d'un modèle hiérarchisé ou allocation de ressources** : acheminez les demandes prioritaires vers les modèles avancés et dirigez le trafic peu prioritaire vers les modèles plus économiques.

Retour au diagramme d'architecture



La menace d'une consommation débridée sur un système LLM ou IA fait référence à une faille de sécurité dans laquelle l'application permet aux utilisateurs, malveillants ou non, d'envoyer des demandes ou des prompts d'inférence excessifs et non contrôlés, sans aucune limite, authentification ou restriction d'utilisation efficaces. L'inférence des LLM étant très gourmande en ressources de calcul, ce manque de contrôle peut être exploité de plusieurs façons : les attaquants peuvent provoquer une attaque par déni de service (DoS) en submergeant les ressources système, générer des pertes économiques imprévues dans les déploiements basés sur le paiement à l'utilisation ou hébergés dans le Cloud, ou interroger systématiquement le modèle pour dupliquer son comportement et voler la propriété intellectuelle. Conséquences : interruption des services, dégradation des performances pour les autres utilisateurs, pression financière et risque accru de fuite de modèles sensibles. Dans les faits, la consommation débridée se produit lorsque l'utilisation des ressources n'est pas correctement gérée, ce qui expose les applications basées sur le LLM à une exploitation accidentelle et délibérée.

Pourquoi choisir Dell pour la sécurité de l'IA

Dell aide les entreprises à protéger les modèles d'IA et les LLM grâce à une approche complète qui couvre le matériel, les logiciels et les services managés. La sécurité est intégrée de la chaîne logistique aux appareils, aux infrastructures, aux données et aux applications, le tout en conformité avec les principes du Zero-Trust. Dans l'ensemble de la gamme, les solutions Dell sont conçues pour améliorer l'hygiène en matière de cybersécurité, avec des fonctionnalités comme la MFA, le RBAC, le moindre privilège et la vérification continue. Cette approche complète, sécurisée dès la conception, garantit aux entreprises la possibilité d'innover en toute confiance avec l'IA et les LLM, réduisant ainsi les risques de vol de modèles, de fuite de données, d'attaques contradictoires et d'autres cybermenaces avancées.

Chaîne d'approvisionnement

La chaîne logistique sécurisée de Dell fournit une protection fondamentale pour les modèles d'IA et les LLM, en intégrant la sécurité à chaque étape du développement, de la fabrication et de la livraison. Grâce à des mises à jour du BIOS et du firmware signées de manière chiffrée, à la vérification sécurisée des composants, à la nomenclature logicielle (SBOM) axée sur l'IA, au suivi de la traçabilité des jeux de données, à la configuration et aux logiciels de sécurité intégrés et à des évaluations rigoureuses des risques fournisseurs conformes aux normes internationales, Dell limite les risques liés au sabotage, aux accès non autorisés et aux attaques de la chaîne logistique. Chaque entreprise est ainsi en mesure de déployer des charges applicatives d'IA fiables et résilientes, avec une transparence, une intégrité et une conformité réglementaire complètes.

PC IA

Dell offre une sécurité de base pour les charges applicatives d'IA intégrées. Dell Trusted Devices, les PC d'IA professionnels les plus sécurisés au monde*, avec la sécurité intégrée dès la conception. La sécurité de la chaîne logistique minimise le risque de vulnérabilité et de falsification des produits. Des défenses uniques intégrées directement au matériel et au firmware protègent le PC et l'utilisateur final lors de son utilisation. Dell SafeBIOS offre une visibilité approfondie au niveau du BIOS et des détections de falsification, tandis que Dell SafeID améliore la sécurité des informations d'identification et permet une authentification sans mot de passe. Les logiciels partenaires offrent une protection avancée sur les points de terminaison, le réseau et les environnements Cloud.

Cyber-résilience

Les solutions de cyberrésilience PowerProtect de Dell sécurisent les données d'IA avec des sauvegardes chiffrées et immuables, une restauration rapide et des coffres-forts isolés de cyberrécupération. Ces fonctionnalités empêchent la destruction, atténuent l'impact des mises à jour malveillantes et prennent en charge la conformité et la récupération après une attaque.

Serveurs

Les serveurs PowerEdge intègrent un calcul confidentiel pour isoler et sécuriser les prompts et les intégrations IA/LLM, des solutions de RAG fiables ancrées dans des sources faisant autorité, ainsi que la MFA, le RBAC, une racine de confiance en silicium, des firmwares signés et une surveillance continue pour protéger les charges applicatives critiques de l'IA.

Stockage

La gamme de solutions de stockage Dell garantit un stockage sécurisé et chiffré pour les données d'IA sensibles avec un chiffrement AES-256 robuste pour les données au repos et en transit. Une technologie de chiffrement avancée, conçue pour résister aux menaces quantiques futures, est disponible sur certaines offres. La gamme comprend des performances NVMe haut débit, des modules de chiffrement conformes à la norme FIPS pour sécuriser les données, y compris celles utilisées dans les charges applicatives d'IA, des snapshots immuables et des coffres-forts de cyberrécupération isolés pour neutraliser les attaques par ransomware. L'architecture Zero-Trust, la sécurité de la chaîne logistique et les fonctionnalités d'audit inviolables améliorent la gouvernance. La détection intégrée des anomalies et les modèles ML AIOps protègent les charges applicatives sans utiliser les données des clients pour l'entraînement, ce qui limite les risques d'attaques liées aux prompts.

AIOps

Dell AIOps assure une surveillance continue et automatisée pour détecter les erreurs de configuration et les vulnérabilités (y compris les CVE), et contribue à la sensibilisation aux risques liés à la chaîne logistique qui peuvent affecter les charges applicatives d'IA/de LLM. L'analyse CVE en temps réel, les alertes intelligentes et les tableaux de bord optimisés par l'IA permettent une intervention rapide en signalant les anomalies et en suivant les workflows de résolution. Des fonctionnalités de conformité intégrées, des contrôles d'accès basés sur les rôles et des rapports automatisés contribuent à sécuriser les opérations sur l'ensemble des charges applicatives, tandis que l'intégration transparente EDR/XDR et les informations opérationnelles basées sur l'IA, y compris les fonctionnalités génératives des solutions prises en charge, améliorent encore l'efficacité IT.

Gestion de réseau

Les solutions Dell Networking protègent les environnements IA/LLM grâce à une segmentation robuste du réseau, qui limite les mouvements latéraux. Des chemins réseau chiffrés et des contrôles de pare-feu intégrés bloquent l'accès non autorisé aux données d'IA.

Services de sécurité et de résilience de l'IA

Les services de sécurité et de résilience de l'IA de Dell sont conçus pour répondre aux nouveaux risques associés à l'adoption de l'IA au sein de votre entreprise. Par défaut, nos services collaborent avec vos équipes pour intégrer l'IA le plus rapidement possible. Notre expertise vous oriente tout au long dans la planification stratégique, de la mise en œuvre des solutions et des services de sécurité gérés afin de réduire les charges opérationnelles et de favoriser l'innovation sécurisée avec l'IA. Chaque service est personnalisé pour aider les entreprises à faire face à l'évolution des risques liés à l'IA et à optimiser les déploiements sécurisés de l'IA.

Dell AI Factory

Une gamme intégrée de solutions de sécurité spécialisées telles que la chaîne logistique sécurisée de Dell, les fonctionnalités Zero-Trust pour appliquer le principe du moindre privilège et les solutions MDR d'IA conçues pour assurer la sécurité de votre modèle.

* D'après une analyse interne réalisée par Dell en octobre 2024 (Intel) et mars 2025 (AMD). S'applique aux PC équipés de processeurs Intel et AMD. Toutes les fonctionnalités ne sont pas disponibles sur tous les PC. Certaines fonctionnalités sont vendues séparément. PC Intel validés par Principled Technologies juillet 2025.

Conclusion

Pour créer des cadres d'IA résilients, une approche collaborative entre les entreprises et les experts en sécurité est essentielle. À mesure que l'IA et les LLM continuent de refaçonner les secteurs, il est essentiel d'aborder les risques qu'ils entraînent, notamment en matière de sécurité des données, d'intégrité des modèles et de conformité. Les entreprises doivent donner la priorité à des stratégies proactives qui intègrent la sécurité à chaque étape de leur parcours vers l'IA.

En tant que partenaire de confiance dans cette mission, Dell Technologies propose une personnalisation complète de l'IA générative, des conseils de sécurité et des solutions intégrées adaptées à vos besoins uniques. En tirant parti des solides solutions de cybersécurité de Dell, les entreprises peuvent atténuer efficacement les risques liés à l'IA et aux LLM tout en optimisant le potentiel de leurs investissements en sécurité existants. Dell permet aux entreprises de protéger leur infrastructure d'IA en intégrant de manière transparente une sécurité avancée dans leurs cadres actuels, garantissant ainsi un environnement sécurisé et prêt pour l'avenir.

Découvrez comment les solutions complètes d'IA de Dell peuvent protéger vos environnements d'IA générative et de LLM : Dell.com/CyberSecurityMonth