DELL Technologies

Die wichtigsten 10 Cybersicherheitsbedenken bei GenAI und LLMs

Erste Schritte

Einführung

Künstliche Intelligenz (KI) revolutioniert die Art und Weise, wie Unternehmen arbeiten, wobei generative KI (Generative Artificial Intelligence, GenAI) und große Sprachmodelle (Large Language Models, LLMs) zu kritischen Workloads in modernen Unternehmensumgebungen werden.

Wie jede andere Workload haben auch diese Anwendungen ihre eigenen Komplexitäten und Schwachstellen, mit denen AnwenderInnen umgehen müssen. KI wird von immer mehr Unternehmen als treibende Kraft für Innovationen, Effizienz und Wettbewerbsvorteile genutzt. Dadurch wird die Sicherheit dieser Anwendungen zu einer grundlegenden Notwendigkeit. Eine gute Cyber-Hygiene ist die Grundlage für die Absicherung jeder Workload. So wie Sie der Sicherheit bei all Ihren Workloads Priorität einräumen, ist es wichtig, auch bei KI eine gute Cyber-Hygiene zu praktizieren. Dazu gehören die Implementierung von Praktiken wie ordnungsgemäßes System-Patching, Multi-Faktor-Authentifizierung, rollenbasierter Zugriff und Netzwerksegmentierung. Diese Maßnahmen sind von grundlegender Bedeutung. Der Schlüssel liegt darin, zu verstehen, wie sie in die spezifische Architektur und Nutzung Ihrer Workload passen.

Wir bei Dell verstehen die Details der KI-Workload und ihrer einzigartigen Sicherheitsherausforderungen. Indem wir herausfinden, wie Bedrohungsakteure diese Workloads angreifen könnten, kann Dell Ihnen helfen, eine robuste Sicherheitsstrategie zu entwickeln. Dazu gehört die Bewältigung von Risiken wie: Vergiftung von Trainingsdaten, Diebstahl oder Manipulation von Modellen, Rekonstruktion von Datensätzen und mehr.

Wir konzentrieren uns auch auf die Bewältigung von Herausforderungen, die mit dem Input für Ihr KI-Modell verbunden sind, wie z. B. die Verhinderung der Offenlegung sensibler Informationen, Maßnahmen gegen Themen mit Sicherheitsrisiken und Vorurteile und die Einhaltung von Vorschriften. Auf der Output-Seite helfen wir dabei, Probleme wie die übermäßige Abhängigkeit vom Modell und Compliance-bezogene Risiken anzugehen.

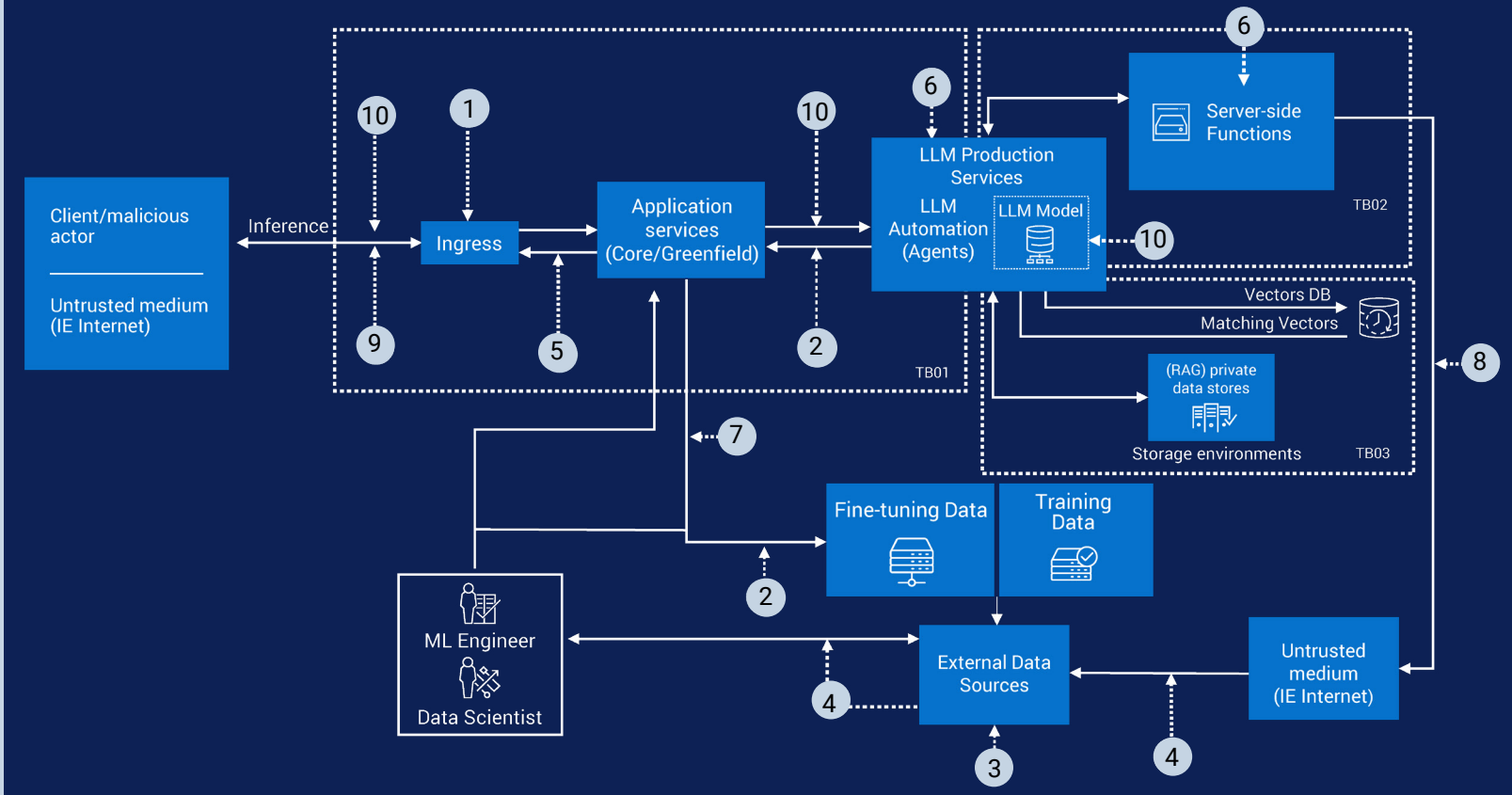
Bei Dell ermöglichen wir Unternehmen, diese Risiken mit ihren bestehenden Cybersicherheitslösungen oder neuen Werkzeugen und Praktiken zur Sicherung ihrer Systeme abzuschwächen. Unser Ziel ist es, sicherzustellen, dass Sicherheit Ihren Innovationen nicht im Weg steht. Durch unser Verständnis von KI-Workloads und ihren Sicherheitsbedrohungen können wir Ihnen helfen, eine stärkere Sicherheitsstruktur aufzubauen, die Ihre Umgebung widerstandsfähiger macht und Ihnen souveräne Innovationen ermöglicht. Mit unserem Fachwissen helfen wir Ihnen, das Potenzial von KI souverän zu nutzen – ohne Abstriche bei der Sicherheit.

Die 10 wichtigsten Bedenken zum Thema Cybersicherheit bei GenAI und LLMs

Dies sind nach Angaben von OWASP die wichtigsten Bedenken beim Schutz von GenAI/LLM-Modellen.

Klicken Sie auf ein Thema, um weitere Informationen zu erhalten:

1. Prompt-Einschleusung
2. Offenlegung sensibler Informationen
3. Lieferkette
4. Vergiftung von Trainingsdaten
5. Unsachgemäße Verarbeitung der Ausgaben
6. Übermäßige Handlungsmacht
7. Lecks von System-Prompts
8. Schwächen bei Vektoren und Einbettung
9. Fehlinformationen
10. Ungebundener Verbrauch

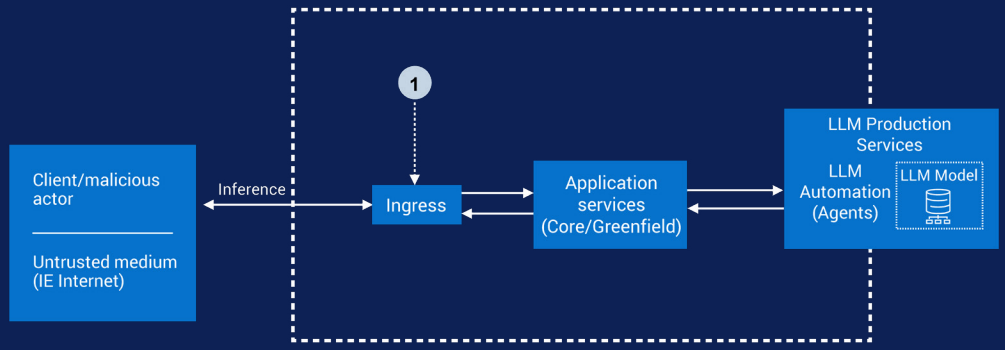


Thema Nr. 1: Prompt-Einschleusung

Strategien zur Verhinderung von Prompt-Einschleusung:

- **Datenbereinigung und Eingabevalidierung:** Überprüfen Sie Nutzereingaben gründlich, um schädliche Inhalte zu entfernen. Verwenden Sie Normalisierung und Codierung, um Missbrauch zu verhindern.
- **Ansätze mit Verarbeitung natürlicher Sprache (Natural Language Processing, NLP) und maschinellem Lernen:** Verwenden Sie NLP und maschinelles Lernen, um manipulierte oder böswillige Prompts zu erkennen und zu blockieren.
- **Klare Ausgabeformatierung und Antwortkontrollen:** Legen Sie strenge Antwortbegrenzungen fest, um sicherzustellen, dass die Ausgaben den beabsichtigten Formaten entsprechen und unbefugte Aktionen zu verhindern. Verwenden Sie Prompt-Filterung und Antwortvalidierung, um die Integrität abzusichern.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfungen für kritische Entscheidungen.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Secure Prompt Engineering:** Verwenden Sie sicheres Prompt-Design und Prompt-Analysen als Teil der allgemeinen Softwaresicherheit, um die Eingabeverarbeitung zu schützen.
- **Modellvalidierung:** Validieren Sie ML-Modelle regelmäßig, um sicherzustellen, dass sie vor dem Einsatz nicht manipuliert wurden, und um ihre Fehlerfreiheit und Integrität zu gewährleisten.
- **Filterung, Einstufung und Antwortvalidierung von Prompts:** Analysieren Sie Prompts und ordnen Sie sie ein, um sicherzustellen, dass nur sichere Eingaben verarbeitet werden. Validieren Sie die Antworten, um Missbrauch zu verhindern.
- **Robustheitsüberprüfungen:** Führen Sie regelmäßige Bewertungen durch, um Schwachstellen zu identifizieren und zu beheben, damit die KI sicher und zuverlässig bleibt.

Zurück zum Architekturdiagramm



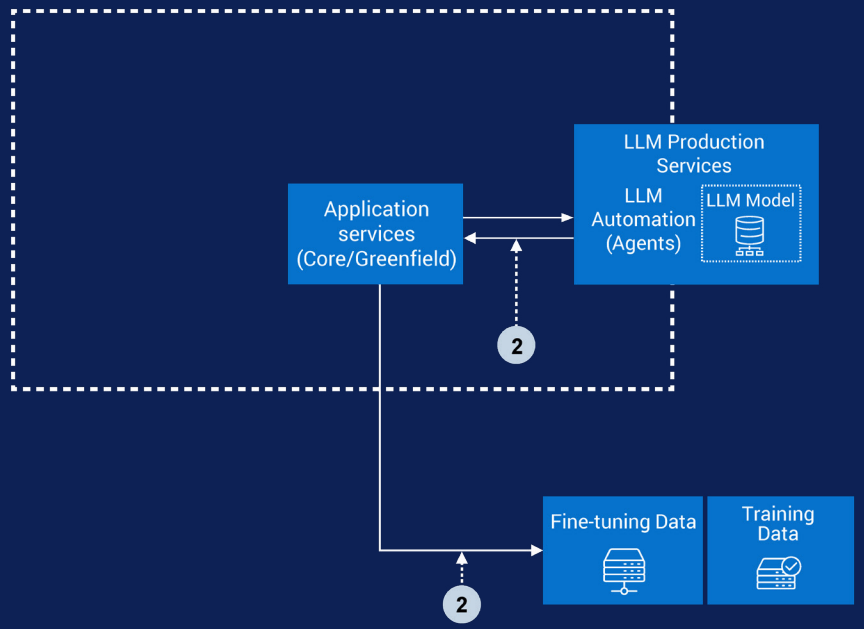
Prompt-Einschleusung ist eine neue Herausforderung in der Welt der generativen KI (GenAI), bei der Eingaben mit böser Absicht erstellt werden, um das Verhalten des Modells zu manipulieren oder seine Integrität zu gefährden. Diese Angriffe nutzen Sicherheitslücken in der Art und Weise aus, wie KI-Systeme Nutzereingaben verarbeiten und darauf reagieren, und können potenziell zu unbefugten Aktionen, Fehlinformationen oder der Offenlegung sensibler Daten führen. Da GenAI zunehmend in kritische Geschäftsworkflows integriert wird, ist die Bewältigung dieser Risiken unerlässlich für Vertrauen und Sicherheit.

Thema Nr. 2: Offenlegung vertraulicher Informationen

Strategien zur Vermeidung der Offenlegung sensibler Informationen:

- **Datenbereinigung und Eingabevalidierung:** Überprüfen Sie Nutzereingaben gründlich, um schädliche Inhalte zu entfernen. Verwenden Sie Normalisierung und Codierung, um Missbrauch zu verhindern.
- **Homomorphe Verschlüsselung:** Verarbeiten Sie sensible Daten sicher, ohne ihren Inhalt preiszugeben. Dadurch wird sichergestellt, dass Daten auch während der Nutzung verschlüsselt und vor Sicherheitsverletzungen geschützt bleiben.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfung für kritische Entscheidungen.
- **Sichere APIs und Systemschnittstellen** für KI-Dateninteraktionen: Überprüfen Sie die Konfigurationen regelmäßig, um die Gefährdung und Angriffsfläche zu minimieren.
- **Sichere Datenerhebung, Storage und Policies:** Setzen Sie umfassende Data-Protection- und Governance-Policies durch, die die Compliance sicherstellen und Datenrisiken minimieren.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Sichere Entwicklung, Konfiguration und Audits:** Sorgen Sie für sichere und aktuelle KI-Systemkonfigurationen. Wenden Sie dazu sichere Codierungsverfahren an, verwenden Sie automatisierte Konfigurationsmanagementtools und führen Sie regelmäßige Überprüfungen, Audits und Updates durch.
- **Nutzerschulung und Sicherheitsbewusstsein:** Bieten Sie NutzerInnen und AdministratorInnen fortlaufende, KI-spezifische Sicherheitsschulungen an, um unsichere Nutzung und versehentliche Datenweitergabe zu vermeiden.

Zurück zum Architekturdiagramm



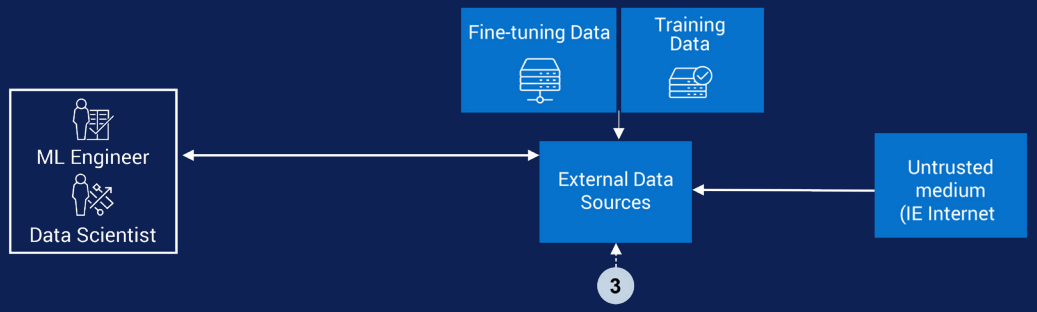
GenAI hat erhebliche Fortschritte gebracht, die aber auch mit erheblichen Risiken einhergehen, insbesondere hinsichtlich der unbeabsichtigten Offenlegung sensibler Informationen. Dieses Risiko betrifft sowohl personenbezogene Daten als auch proprietäre Geschäftsdaten. Der Missbrauch oder die falsche Handhabung von GenAI-Tools kann zu Datenlecks, Nichteinhaltung von Vorschriften oder Rufschädigung führen. Daher ist es für Unternehmen von entscheidender Bedeutung, diese Risiken zu verstehen und proaktiv Maßnahmen zu ergreifen, um eine sichere Implementierung und Nutzung von KI-Systemen sicherzustellen.

Thema Nr. 3: Sicherheitslücken in der Lieferkette

Strategien zur Schließung von Sicherheitslücken in der Lieferkette:

- **Prüfen der Lieferanten, unter anderem auf die Einhaltung sicherer Lieferkettenpraktiken:** Bewerten Sie Lieferanten und treffen Sie Vereinbarungen, die die Sicherheit der Lieferkette priorisieren.
- **Implementieren eines Software Bill of Materials:** Verfolgen und überprüfen Sie die Herkunft von Softwarekomponenten, um Transparenz zu gewährleisten und das Risiko von kompromittiertem Code zu verringern.
- **Modellvalidierung:** Validieren Sie ML-Modelle regelmäßig, um sicherzustellen, dass sie vor dem Einsatz nicht manipuliert wurden, und um ihre Fehlerfreiheit und Integrität zu gewährleisten.
- **Ausführen von Containern und Pods mit den geringsten Berechtigungen:** Dies reduziert die potenziellen Auswirkungen im Falle einer Sicherheitsverletzung und schränkt den unbefugten Zugriff ein.
- **Nutzung von Firewalls:** Blockieren Sie unnötige Netzwerkverbindungen, um die Anfälligkeit für potenzielle Bedrohungen zu verringern und die Möglichkeiten für AngreiferInnen zu begrenzen.
- **Schutz von Daten und Anmerkungen:** Sichern Sie Ihre Daten und die zugehörigen Anmerkungen, um Manipulationen, unbefugten Zugriff und Beschädigungen kritischer Informationen zu verhindern.
- **Sichere Hardware:** Verwenden Sie sicherheitsgeprüfte Hardware, um Schwachstellen zu vermeiden, die durch hardwarebasierte Angriffe entstehen könnten, und sorgen Sie so für eine solide Grundlage für Ihre Infrastruktur.
- **Sichere ML-Softwarekomponenten:** Verwenden Sie vertrauenswürdige und überprüfte ML-Softwarekomponenten, um Schwachstellen zu reduzieren und die Sicherheit Ihrer maschinellen Lernabläufe insgesamt zu verbessern.
- **Sichere Entwicklung, Konfiguration und Audits:** Sorgen Sie für sichere und aktuelle KI-Systemkonfigurationen. Wenden Sie dazu sichere Codierungsverfahren an, verwenden Sie automatisierte Konfigurationsmanagementtools und führen Sie regelmäßige Überprüfungen, Audits und Updates durch.

Zurück zum Architekturdiagramm



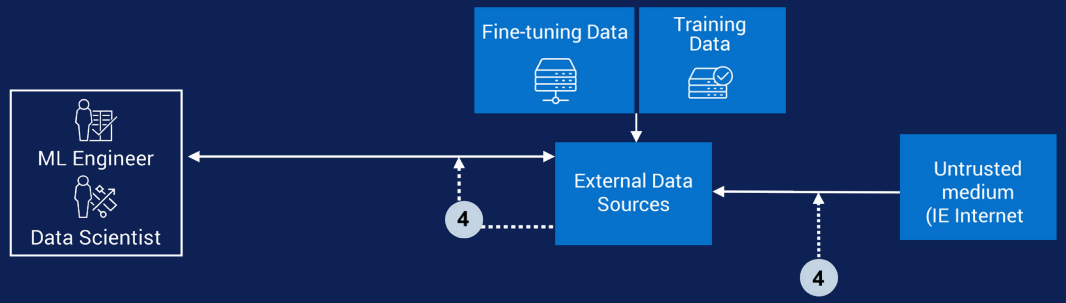
Untersuchen Sie Schwachstellen in der LLM-Lieferkette, die sich auf kritische Komponenten wie die Integrität vortrainierter Modelle und Adapter von Drittanbietern auswirken können. KI-Systeme sind sowohl auf Hardware als auch auf Software angewiesen. Diese können lange vor der Bereitstellung kompromittiert werden. AngreiferInnen können Schwachstellen in verschiedenen Phasen der Lieferkette des maschinellen Lernens ausnutzen und GPU-Hardware, Daten und deren Anmerkungen, Elemente des ML-Software-Stacks oder sogar das Modell selbst manipulieren. Über diese Einzelabschnitte können sie sich einen Einstieg in die Systeme schaffen, wodurch erhebliche Sicherheits- und Integritätsrisiken entstehen. Das Verständnis und die Schließung dieser Sicherheitslücken ist entscheidend für die Entwicklung robuster, sicherer KI-Lösungen.

Thema Nr. 4: Vergiftung von Trainingsdaten

Strategien zur Vermeidung der Vergiftung von Modelldaten:

- **Anomalieerkennung und Datenvalidierung während des Trainings:** Identifizieren Sie dadurch Inkonsistenzen in Daten, beheben Sie sie und stellen Sie so sicher, dass nur bereinigte, hochwertige Daten zum Trainieren des Modells verwendet werden.
- **Isolieren Sie Umgebungen während des Fine-Tunings:** Verhindern Sie so unbefugten Zugriff und Kontaminierung des Modells während kritischer Entwicklungsphasen.
- **Modellvalidierung:** Validieren Sie ML-Modelle regelmäßig, um sicherzustellen, dass sie vor dem Einsatz nicht manipuliert wurden, und um ihre Fehlerfreiheit und Integrität zu gewährleisten.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfungen für kritische Entscheidungen.
- **Datenbereinigung und Eingabevalidierung:** Überprüfen Sie Nutzereingaben gründlich, um schädliche Inhalte zu entfernen. Verwenden Sie Normalisierung und Codierung, um Missbrauch zu verhindern.
- **Sichere Entwicklung, Konfiguration und Audits:** Sorgen Sie für sichere und aktuelle KI-Systemkonfigurationen. Wenden Sie dazu sichere Codierungsverfahren an, verwenden Sie automatisierte Konfigurationsmanagementtools und führen Sie regelmäßige Überprüfungen, Audits und Updates durch.
- **Robustheitsüberprüfungen:** Führen Sie regelmäßige Bewertungen durch, um Schwachstellen zu identifizieren und zu beheben, damit die KI sicher und zuverlässig bleibt.
- **Netzwerksegmentierung:** Schränken Sie den Zugriff auf unsichere Schnittstellen und kritische Systemkomponenten ein.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.

Zurück zum Architekturdiagramm



Die Vergiftung von Modelldaten ist eine Sicherheitsbedrohung im KI-Lebenszyklus, bei der AngreiferInnen absichtlich Trainingsdaten mit beschädigten, irreführenden oder böswilligen Eingaben verunreinigen. Dieses Risiko kann Auswirkungen auf kritische Komponenten haben – von der Erfassung von Rohdaten und Anmerkungen bis hin zur Kuratierung und Integration von Datensätzen, die für maschinelles Lernen oder große Sprachmodelle verwendet werden. Die Zuverlässigkeit von KI-Systemen hängt von der Integrität ihrer Datenquellen ab, die vor dem Training, während der Vorverarbeitung oder über externe Datenpipelines manipuliert werden können.

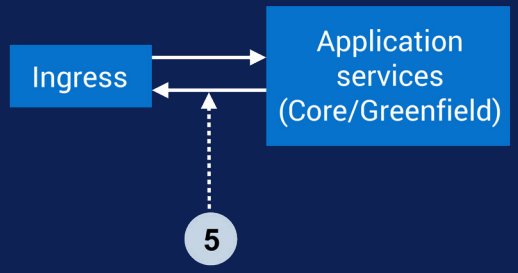
AngreiferInnen nutzen Datenvergiftung, um die Modellgenauigkeit zu mindern, Sicherheitslücken zu schaffen oder schädliche Ausgaben auszulösen. Indem sie auf Schwachstellen in der Datenherkunft, der Qualität der Anmerkungen oder den Prozessen zur Aufnahme von Datensätzen abzielen, können AngreiferInnen die Sicherheit, Vertrauenswürdigkeit und Resilienz untergraben. Die Erkennung und Abwehr dieser datenzentrierten Bedrohungen ist für die Entwicklung robuster, zuverlässiger KI-Lösungen unerlässlich.

Problem Nr. 5: Unsachgemäße Verarbeitung der Ausgaben

Strategien zur Vermeidung unsachgemäßer Ausgabeverarbeitung:

- **Kontextbezogene Ausgabekodierung:** Wenden Sie immer Codierungs- und Escaping-Techniken an, die auf den spezifischen Kontext zugeschnitten sind, in dem die Ausgabe verwendet wird, z. B. HTML-, SQL- oder API-Umgebungen, um Sicherheitslücken, etwa gegenüber Injektionsangriffen, zu schließen.
- **Bereinigung von Ausgaben:** Befolgen Sie strenge Validierungs- und Bereinigungsverfahren für Modellausgaben, die den Richtlinien des Application Security Verification Standard (ASVS) des Open Web Application Security Project (OWASP) entsprechen, um eine sichere Weiterverwendung zu gewährleisten und Sicherheitsrisiken zu minimieren.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Automatisierte Output-Sicherheitstests:** Führen Sie regelmäßige Sicherheitstests mit automatisierten Tools durch, um Risiken in Ausgaben zu identifizieren, wie z. B. Sicherheitslücken mit Cross-Site Scripting (XSS) oder Injektionsschwachstellen, und beheben Sie diese proaktiv.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfung für kritische Entscheidungen.
- **Human-in-the-Loop-Überprüfung:** Bei risikoreichen Anwendungen, wie z. B. im Finanz- oder Gesundheitswesen, ist eine menschliche Aufsicht und Überprüfung der Modellergebnisse erforderlich, um Fehlerfreiheit, Sicherheit und Schutz zu gewährleisten.
- **Datenschutz und Compliance:** Integrieren Sie Techniken zur Wahrung des Datenschutzes in den Ausgabeprozess und sorgen Sie für die Einhaltung einschlägiger Vorschriften und Standards für die sichere Verwendung sensibler Daten.

Zurück zum Architekturdiagramm



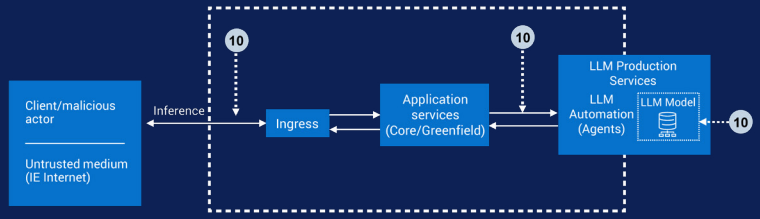
Eine unzureichende Validierung oder Bereinigung der KI-Modellausgabe kann zu schwerwiegenden Sicherheitsrisiken führen, einschließlich der Ausweitung von Zugriffsrechten und Datenschutzverletzungen. Wenn KI-Modelle Ausgaben produzieren, die nicht ordnungsgemäß geprüft oder gefiltert werden, können böswillige Akteure diese Schwachstellen ausnutzen, um sich unbefugten Zugang zu verschaffen oder ihre Rechte innerhalb eines Systems zu erweitern. Dieser Mangel an Aufsicht kann zu kompromittierten Daten, unbefugten Handlungen und erheblichen Sicherheitsverletzungen führen. Dies unterstreicht, wie wichtig es ist, robuste Validierungs- und Bereinigungsprozesse für alle KI-generierten Ausgaben zu implementieren.

Thema Nr. 6: Übermäßige Handlungsmacht

Strategien zur Minderung übermäßiger Handlungsmacht

- **Durchsetzung der geringstmöglichen Berechtigung:** Gewähren Sie LLMs und agentenbasierten Subsystemen nur die minimalen Berechtigungen, die für die Durchführung der geplanten Operationen erforderlich sind, und überprüfen Sie die Zugriffskontrollen regelmäßig.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfungen für kritische Entscheidungen.
- **Festlegen von Einsatzbegrenzungen:** Definieren Sie klar, worauf LLMs/Agents zugreifen oder was sie ausführen können.
- **Human-in-the-Loop-Überprüfung:** Bei risikoreichen Anwendungen, wie z. B. im Finanz- oder Gesundheitswesen, ist eine menschliche Aufsicht und Überprüfung der Modellergebnisse erforderlich, um Fehlerfreiheit, Sicherheit und Schutz zu gewährleisten.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Beschränkung der Autonomie:** Schränken Sie die Funktionen von LLMs ein, um uneingeschränkten Zugriff oder uneingeschränkte Kontrolle zu vermeiden.
- **Sichere Entwicklung, Konfiguration und Audits:** Sorgen Sie für sichere und aktuelle KI-Systemkonfigurationen. Wenden Sie dazu sichere Codierungsverfahren an, verwenden Sie automatisierte Konfigurationsmanagementtools und führen Sie regelmäßige Überprüfungen, Audits und Updates durch.
- **Nutzung von Firewalls:** Blockieren Sie unnötige Netzwerkverbindungen, um die Anfälligkeit für potenzielle Bedrohungen zu verringern und die Möglichkeiten für AngreiferInnen zu begrenzen.
- **Robustheitsüberprüfungen:** Führen Sie regelmäßige Bewertungen durch, um Schwachstellen zu identifizieren und zu beheben, damit die KI sicher und zuverlässig bleibt.

Zurück zum Architekturdiagramm



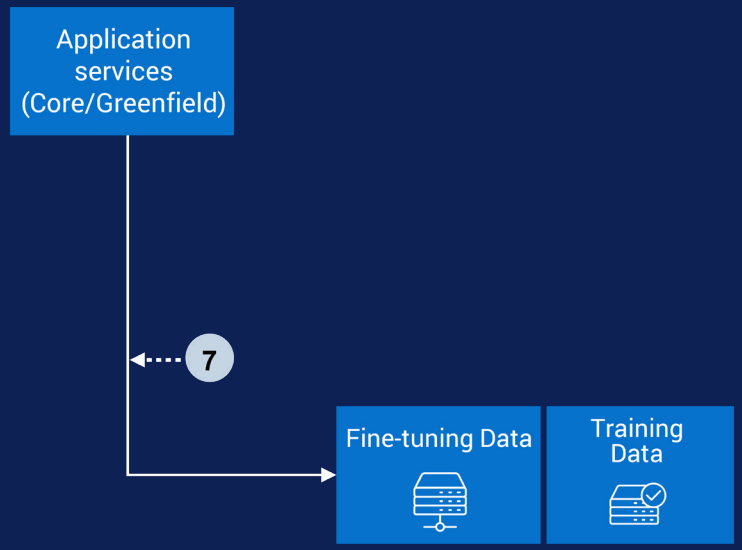
Wenn KI-Agenten oder -Plugins übermäßige Autonomie oder unnötige Funktionen innerhalb von Workflows erhalten, kann dies erhebliche Risiken mit sich bringen. Weiter als erforderlich gefasste Berechtigungen oder Funktionen von KI-Systemen erhöhen das Risiko unerwünschter Folgen. Wenn Systeme auf der Basis großer Sprachmodelle mit zu vielen Berechtigungen entwickelt werden, können sie ungewollt Aktionen durchführen oder auf Informationen zuzugreifen. Eine solche Überschreitung kann zu Fehlern, Datenmissbrauch oder sogar Sicherheitslücken führen. Dies zeigt, wie wichtig es ist, KI-Funktionen sorgfältig zu begrenzen und zu überwachen, um eine sichere und verantwortungsvolle Nutzung zu gewährleisten.

Thema Nr. 7: Lecks von Prompts

Strategien zur Minderung von Lecks von Prompts

- **Vermeiden der Einbettung sensibler Informationen in Prompts:** Erwähnen Sie niemals Zugangsdaten, API-Schlüssel oder proprietäre Logik in Prompts– verwalten Sie diese sicher außerhalb des Systems.
- **Von Prompts separate Sicherheitskontrollen:** Verwalten Sie Authentifizierung, Autorisierung und Sitzungsmanagement in der Anwendungslogik, nicht in Prompts.
- **Validieren von Eingaben und Ausgaben:** Bereinigen Sie Prompts und Antworten mit einer robusten Validierung, um verdächtige Muster oder Manipulationen zu verhindern.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfungen für kritische Entscheidungen.
- **Verschlüsseln und Sichern von Prompts:** Speichern Sie Prompts und Konfigurationen in verschlüsseltem, sicherem Storage, um unbefugten Zugriff zu verhindern.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Regelmäßige Überprüfung der Prompts:** Überprüfen und bereinigen Sie die Prompts regelmäßig, um sensible Daten zu entfernen und die Einhaltung der Sicherheitsvorschriften zu kontrollieren.
- **Testen und Suchen nach Schwachstellen (Red-Teaming):** Führen Sie kontradiktorische Tests durch, um Schwachstellen in der Prompt-Verwaltung oder den Ausgaben zu identifizieren und zu beheben.
- **Isolieren von Prompts von Nutzereingaben:** Entwickeln Sie Systeme, die verhindern, dass Nutzerabfragen Prompts manipulieren oder offenlegen.
- **Durchsetzen von Ratenbeschränkungen:** Beschränken Sie die API-Nutzung, drosseln Sie verdächtige Aktivitäten und blockieren Sie automatische Prompt-Angriffe.

Zurück zum Architekturdiagramm



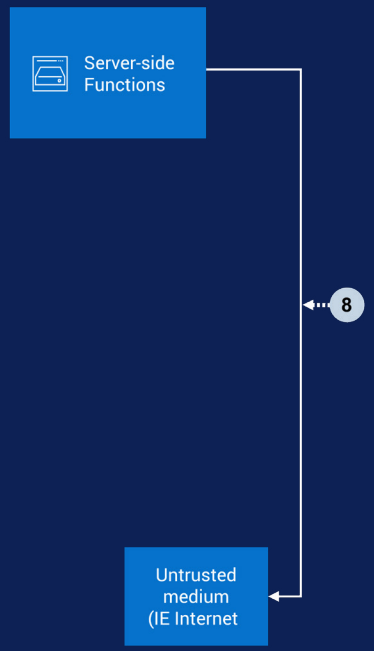
Bei einem System-Prompt-Leckage-Angriff auf ein LLM (Large Language Model) oder ein KI-System extrahieren AngreiferInnen die versteckten Anweisungen – die „System-Prompts“ – die das Verhalten des Modells steuern seine Einsatzbegrenzungen festlegen, oder leiten diese ab. Diese Prompts sollen in der Regel nicht für EndnutzerInnen sichtbar sein, da sie Kernregeln, Einschränkungen und manchmal sensible Betriebslogik enthalten. Durch speziell gestaltete Eingaben oder das Ausnutzen von Sicherheitslücken können AngreiferInnen das LLM dazu bringen, seine System-Prompts entweder ganz oder teilweise offenzulegen. Wenn diese Informationen nach außen dringen, können sie dazu verwendet werden, Einschränkungen rückgängig zu machen, Sicherheitsfilter zu umgehen oder neue gezielte Angriffe zu entwickeln, was letztlich das Risiko von Prompt-Einschleusung, Berechtigungserweiterung oder Missbrauch des Modells und nachgelagerter Systeme, die auf seine Integrität angewiesen sind, erhöht.

Thema Nr. 8: Schwächen bei Vektoren und Einbettung

Strategien zur Minderung von Vektor- und Einbettungsschwächen

- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfungen für kritische Entscheidungen.
- **Verschlüsselung:** Schützen Sie Vektordaten bei der Übertragung und im Ruhezustand mit robusten Verschlüsselungsstandards wie AES.
- **Sichere Konfiguration und Überwachung:** Machen Sie Ihre Systeme weniger angreifbar, konfigurieren Sie sie sicher und überwachen Sie sie kontinuierlich auf Fehlkonfigurationen, unbefugten Zugriff oder Anomalien.
- **Sicherheitslückenmanagement:** Aktualisieren und patchen Sie regelmäßig alle Programme, Abhängigkeiten und Vektorspeicher-Engines, um Sicherheitsrisiken zu vermeiden.
- **Datenbereinigung und Eingabevalidierung:** Überprüfen Sie Nutzereingaben gründlich, um schädliche Inhalte zu entfernen. Verwenden Sie Normalisierung und Codierung, um Missbrauch zu verhindern.
- **Sichere APIs und Systemschnittstellen** für KI-Dateninteraktionen: Überprüfen Sie die Konfigurationen regelmäßig, um die Gefährdung und Angriffsfläche zu minimieren.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Sichere Hardware:** Verwenden Sie sicherheitsgeprüfte Hardware, um Schwachstellen zu vermeiden, die durch hardwarebasierte Angriffe entstehen könnten, und sorgen Sie so für eine solide Grundlage für Ihre Infrastruktur.
- **Sichere Entwicklung, Konfiguration und Audits:** Sorgen Sie für sichere und aktuelle KI-Systemkonfigurationen. Wenden Sie dazu sichere Codierungsverfahren an, verwenden Sie automatisierte Konfigurationsmanagementtools und führen Sie regelmäßige Überprüfungen, Audits und Updates durch.

Zurück zum Architekturdiagramm



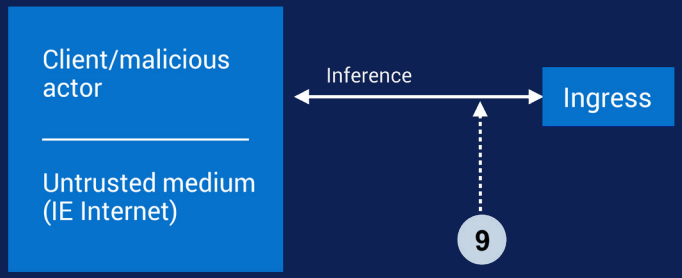
Ein Angriff infolge von Vektor- und Einbettungsschwächen auf ein LLM (Large Language Model) oder ein KI-System – insbesondere solche, die Retrieval Augmented Generation (RAG) verwenden – nutzt Schwachstellen in der Art und Weise aus, wie Informationen als numerische Vektoren und Einbettungen kodiert, gespeichert und abgerufen werden. Schwachstellen in diesen Mechanismen können durch böswillige Aktionen wie Einbettungsinversion (Rekonstruktion sensibler Daten aus Einbettungen), Datenvergiftung (Einspeisung schädlicher oder verzerrter Inhalte zur Manipulation des Modellverhaltens), unbefugten Zugriff auf Vektordatenbanken (was zu Datenlecks führt) oder Manipulation der Abrufergebnisse ausgenutzt werden. Diese Angriffe bedrohen Datenschutz, Integrität und Zuverlässigkeit, indem sie es AngreiferInnen ermöglichen, sensible Informationen offenzulegen, Ausgaben zu ändern oder das Vertrauen der NutzerInnen in KI-gesteuerte Anwendungen zu untergraben. Ordnungsgemäße Zugriffskontrollen, Datenvalidierung, Verschlüsselung und fortlaufendes Monitoring sind für die Abwehr dieser sich entwickelnden Bedrohungen von entscheidender Bedeutung.

Thema Nr. 9: Fehlinformationen

Strategien zur Vermeidung von Fehlinformationen

- **Retrieval-Augmented Generation (RAG) mit maßgeblichen Quellen:** Verwenden Sie RAG, um Informationen aus verifizierten, vertrauenswürdigen Datenbanken und Wissensbeständen abzurufen und zu integrieren und so Halluzinationen zu reduzieren.
- **Modelltuning und Ausgabekalibrierung:** Optimieren Sie Modelle mit vielfältigen Datensätzen und minimieren Sie Vorurteile und Fehlinformationen mit dazu entwickelten Techniken.
- **Automatisierte Faktenüberprüfung:** Vergleichen Sie Ausgaben mit zuverlässigen Quellen und kennzeichnen Sie automatisch falsche Informationen.
- **Überwachung von Unsicherheiten:** Markieren Sie Antworten, bei denen das Modell nicht sicher ist, damit sie in kritischen Fällen von Menschen geprüft werden können.
- **Human-in-the-Loop-Überprüfung:** Bei risikoreichen Anwendungen, wie z. B. im Finanz- oder Gesundheitswesen, ist eine menschliche Aufsicht und Überprüfung der Modellergebnisse erforderlich, um Fehlerfreiheit, Sicherheit und Schutz zu gewährleisten.
- **Nutzerfeedback:** Ermöglichen Sie es NutzerInnen, Fehler zu melden, um das Modell kontinuierlich zu verbessern und Fehlinformationen schnell zu korrigieren.
- **Zugriffsbeschränkungen und menschliche Aufsicht:** Wenden Sie rollenbasierte Zugriffskontrolle (RBAC), Multifaktor-Authentifizierung (MFA) und Identitätsmanagement an, um den Zugriff zu begrenzen. Nutzen Sie menschliche Überprüfungen für kritische Entscheidungen.
- **Sichere Entwicklung, Konfiguration und Audits:** Sorgen Sie für sichere und aktuelle KI-Systemkonfigurationen. Wenden Sie dazu sichere Codierungsverfahren an, verwenden Sie automatisierte Konfigurationsmanagementtools und führen Sie regelmäßige Überprüfungen, Audits und Updates durch.
- **Risikokommunikation:** Informieren Sie NutzerInnen über KI-Einschränkungen und ermutigen Sie sie dazu, die Richtigkeit der Ergebnisse unabhängig zu prüfen.
- **Bewusstes UI- und API-Design:** Markieren Sie KI-generierte Inhalte und leiten Sie NutzerInnen zu einer verantwortungsvollen Nutzung an.

Zurück zum Architekturdiagramm



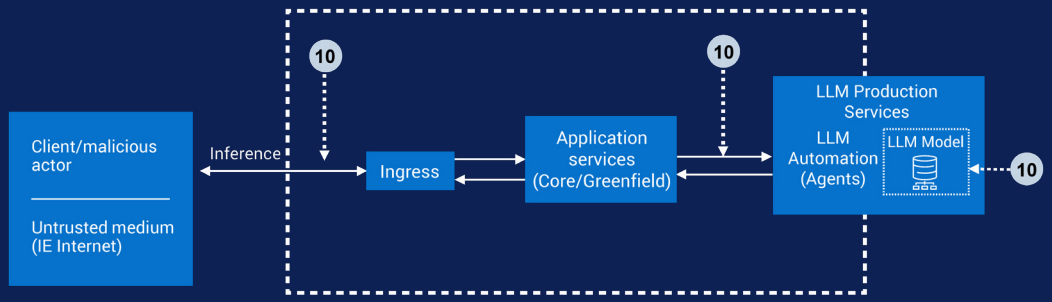
Ein Fehlinformationsangriff auf ein LLM- oder KI-System ist ein absichtlicher Versuch, das Modell dazu zu bringen, falsche, irreführende oder scheinbar glaubwürdige – aber inkorrekte– Informationen als Ausgaben zu generieren oder zu verbreiten. Diese Anfälligkeit ist auf mehrere Faktoren zurückzuführen: die Neigung des Modells zu „Halluzinationen“ (die Generierung erfundener, aber plausibel klingender Inhalte), Vorurteilen oder Lücken in den Trainingsdaten und der Einfluss von mit böswilliger Absicht formulierten Prompts. Zu Halluzinationen kommt es, weil LLMs Text generieren, der in ein statistisches Muster passt, anstatt die Fakten wirklich zu verstehen, was zu Antworten führt, die selbstbewusst erscheinen, aber in Wirklichkeit keine Grundlage haben. Zu den Risiken solcher Angriffe gehören Sicherheitsverletzungen, Rufschädigung und sogar rechtliche Haftung, insbesondere in Umgebungen, in denen sich Benutzer zu sehr auf LLM-Antworten verlassen, ohne deren Richtigkeit oder Anwendbarkeit zu überprüfen, wodurch Fehler oder Fehlinformationen in wichtige Entscheidungen und Prozesse einfließen können.

Thema Nr. 10: Ungebundener Verbrauch

Strategien gegen ungebundenen Verbrauch

- **Durchsetzen von Ratenbeschränkungen und Nutzerquoten:** Legen Sie strenge Beschränkungen für Anfragen, Token oder Daten pro NutzerIn, API-Schlüssel oder App fest, um Missbrauch zu verhindern.
- **Verpflichtende Authentifizierung und Nutzersegmentierung:** Verwenden Sie eine starke Authentifizierung (z. B. API-Schlüssel, OAuth) und weisen Sie Rollen oder Tiers zu, um nur autorisierte Anfragen zu verarbeiten.
- **Eingabevalidierung und Größenbeschränkungen:** Validieren Sie die Größe und Struktur der Prompts und blockieren oder kürzen Sie große oder fehlerhafte Abfragen.
- **Anwenden von Verarbeitungs-Timeouts und Ressourcen-Drosselung:** Setzen Sie Timeouts und Ressourcenobergrenzen für jede Anfrage fest, um langwierige Operationen und Ressourcenverbrauch zu vermeiden.
- **Einsetzen von intelligentem Caching und Deduplizierung:** Legen Sie Antworten auf doppelte oder ähnliche Abfragen im Cache ab, um unnötige Verarbeitung zu vermeiden.
- **Monitoring, Protokollierung und Erkennung von Anomalien:** Überwachen und protokollieren Sie die Aktivitäten des KI-Systems kontinuierlich mit Lösungen wie MDR/XDR/SIEM, um unbefugte Zugriffe, Anomalien oder Datenlecks schnell zu erkennen, zu untersuchen und darauf zu reagieren.
- **Budgetverfolgung und Ausgabenkontrollen:** Verwenden Sie Dashboards und Warnmeldungen, um die Kosten zu überwachen und die Nutzung bei Budgetschwellen zu blockieren.
- **Sandboxing und Isolationstechniken:** Führen Sie Workloads in isolierten Umgebungen mit eingeschränkten Berechtigungen aus, um Risiken zu reduzieren.
- **Begrenzen von Gesprächstiefe und Sprecherwechseln:** Legen Sie Einschränkungen für rekursive Abfragen oder Gesprächsschritte fest, um eine Ausnutzung zu verhindern.
- **Anwenden von Modellen mit Tiers oder Ressourcenzuweisung:** Leiten Sie Anfragen mit hoher Priorität zu Premium-Modellen und Datenverkehr mit niedriger Priorität zu kostengünstigen Modellen um.

Zurück zum Architekturdiagramm



Eine Bedrohung durch unbeschränkten Verbrauch in einem LLM- oder KI-System ist eine Sicherheitslücke, bei der die Anwendung es NutzerInnen ohne wirksame Ratenbegrenzung, Authentifizierung oder Nutzungsbeschränkungen erlaubt – mit böser Absicht oder in gutem Glauben – exzessive, unkontrollierte Inferenzanfragen oder Aufforderungen einzureichen. Da LLM-Inferenzen sehr rechenintensiv sind, kann dieser Mangel an Kontrolle auf verschiedene Weise ausgenutzt werden: AngreiferInnen können Denial-of-Service (DoS) verursachen, indem sie die Systemressourcen überlasten, unvorhergesehene wirtschaftliche Verluste in Pay-per-Use- oder Cloud-gehosteten Implementierungen erzeugen oder das Modell systematisch abfragen, um sein Verhalten zu klonen und geistiges Eigentum zu stehlen. Zu den Folgen gehören Serviceunterbrechungen, Leistungseinbußen für andere NutzerInnen, finanzielle Belastungen und ein erhöhtes Risiko, dass sensible Modelldaten nach außen dringen. Wenn die Ressourcennutzung nicht ordnungsgemäß geregelt ist, kommt es zu einer unbegrenzten Nutzung, wodurch LLM-basierte Anwendungen sowohl versehentlich als auch vorsätzlich ausgenutzt werden können.

Gründe für die Auswahl von Dell für KI-Sicherheit

Dell unterstützt Unternehmen durch einen umfassenden Ansatz, der Hardware, Software und Managed Services umfasst, bei der Sicherung von KI-Modellen und LLMs. Sicherheit ist von der Lieferkette bis zu den Geräten, der Infrastruktur, den Daten und den Anwendungen eingebettet und orientiert sich am Zero-Trust-Prinzip. Die Lösungen von Dell sind so konzipiert, dass sie die Cyber-Hygiene mit Funktionen wie MFA, RBAC, minimalen Rechten und kontinuierlicher Überprüfung verbessern. Dieser umfassende „Sicher per Design"-Ansatz stellt sicher, dass Unternehmen KI und LLMs souverän für Innovationen nutzen und das Risiko von Modelldiebstahl, Datenlecks, feindlichen Angriffen und anderen fortschrittlichen Cyber-Bedrohungen minimieren können.

Lieferkette

Die sichere Lieferkette von Dell bietet grundlegenden Schutz für KI-Modelle und LLMs, indem sie Sicherheit in jede Phase der Produktentwicklung, Fertigung und Lieferung integriert. Durch kryptografisch signierte BIOS- und Firmware-Updates, Secured Component Verification, einen auf KI ausgerichteten Software Bill of Materials (SBOM), Verfolgung der Datensatzherkunft, integrierte Sicherheitssoftware und -konfiguration sowie strenge, an globalen Standards ausgerichtete Risikobewertungen von Anbietern minimiert Dell die Risiken von Manipulationen, unbefugtem Zugriff und Angriffen auf die Lieferkette und stellt damit sicher, dass Unternehmen vertrauenswürdige, widerstandsfähige KI-Workloads mit vollständiger Transparenz, Integrität und Einhaltung gesetzlicher Vorschriften bereitstellen können.

KI-PCs

Dell bietet grundlegende Sicherheit für KI-Workloads auf dem Gerät. Dell Trusted Devices – die sichersten KI-PCs der Welt* – sind auf Sicherheit ausgelegt. Die Lieferkettensicherheit minimiert das Risiko von Produktsicherheitslücken und Manipulationen. Einzigartige Abwehrmaßnahmen, die direkt in Hardware und Firmware integriert sind, schützen den PC und die EndnutzerInnen bei der Nutzung. Dell SafeBIOS bietet umfassende Transparenz und Manipulationserkennung auf BIOS-Ebene, während Dell SafeID die Sicherheit von Zugangsdaten verbessert und eine Authentifizierung ohne Kennwort ermöglicht. Partnersoftware bietet erweiterten Schutz für Endpunkt-, Netzwerk- und Cloud-Umgebungen.

Ausfallsicherheit bei Cyberangriffen

Die PowerProtect-Lösungen von Dell für die Ausfallsicherheit bei Cyberangriffen sichern KI-Daten mit verschlüsselten, unveränderlichen Backups, schneller Wiederherstellung und isolierten Cyber Recovery Vaults.
Diese Funktionen verhindern Zerstörung, mindern die Auswirkungen böswilliger Updates und unterstützen Compliance und Recovery nach einem Angriff.

Server

PowerEdge-Server bieten vertrauliches Computing zur Isolierung und Sicherung von KI-/LLM-Prompts und -Einbettungen, vertrauenswürdige Retrieval-Augmented-Generation-Lösungen (RAG), die in vertrauenswürdigen Quellen verankert sind, sowie MFA, RBAC, Silicon Root of Trust, signierte Firmware und kontinuierliche Überwachung zum Schutz kritischer KI-Workloads.

Storage

Das Storage-Portfolio von Dell gewährleistet den sicheren, verschlüsselten Storage sensibler KI-Daten mit robuster AES-256-Verschlüsselung für Data at Rest und während der Übertragung. Eine erweiterte Verschlüsselung, die gegen künftige Quantenbedrohungen resistent ist, ist für ausgewählte Angebote verfügbar. Das Portfolio umfasst Hochgeschwindigkeits-NVMe-Performance, FIPS-konforme Verschlüsselungsmodule zum Schutz von Daten – einschließlich der in KI-Workloads verwendeten Daten –, unveränderliche Snapshots und Air-Gap-Cyber-Recovery-Vaults zur Abwehr von Ransomware-Angriffen. Zero-Trust-Architektur, Lieferkettensicherheit und manipulationssichere Auditfunktionen verbessern die Governance. Integrierte Anomalieerkennung und AIOps-ML-Modelle schützen Workloads, ohne Kundendaten für das Training zu verwenden, und minimieren so die Risiken Input-basierter Angriffe.

AIOps

Dell AIOps bietet automatisiertes, kontinuierliches Monitoring zur Erkennung von Fehlkonfigurationen und Sicherheitslücken (einschließlich CVEs) und unterstützt das Bewusstsein für Risiken, die sich auf KI-/LLM-Workloads auswirken, in der Lieferkette. CVE-Scans in Echtzeit, intelligente Warnmeldungen und KI-gestützte Dashboards ermöglichen eine schnelle Intervention, indem sie Anomalien markieren und Lösungsworkflows verfolgen. Integrierte Compliancefunktionen, rollenbasierte Zugriffskontrollen und automatisiertes Reporting tragen dazu bei, einen sicheren Betrieb über Workloads hinweg aufrechtzuerhalten, während die nahtlose EDR/XDR-Integration und KI-gesteuerte betriebliche Einblicke – einschließlich generativer Funktionen in unterstützten Lösungen – die IT-Effizienz weiter verbessern.

Networking

Die Netzwerklösungen von Dell schützen KI-/LLM-Umgebungen durch eine robuste Netzwerksegmentierung, die laterale Bewegungen minimiert. Verschlüsselte Netzwerkpfade und integrierte Firewallkontrollen blockieren unbefugten Zugriff auf KI-Daten.

KI-Services für Sicherheit und Resilienz

Die KI-Services für Sicherheit und Resilienz von Dell sind darauf ausgelegt, neue Risiken im Zusammenhang mit der Integration von KI in Ihr Unternehmen zu bewältigen. Unsere Services wurden für die Zusammenarbeit mit Ihren Teams entwickelt, während Sie KI so schnell wie möglich integrieren. Sie bieten Fachwissen für strategische Planung, Lösungsimplementierung und Managed Security Services, um den betrieblichen Aufwand zu verringern, damit Sie mit KI sicher Innovationen schaffen können. Jeder ist darauf zugeschnitten, Unternehmen dabei zu unterstützen, sich entwickelnde KI-Risiken zu bewältigen und sichere KI-Bereitstellungen zu optimieren.

Dell AI Factory

Ein integriertes Portfolio von speziell entwickelten Sicherheitslösungen wie die sichere Lieferkette von Dell, Zero-Trust-Funktionen zur Durchsetzung der geringstmöglichen Berechtigungen und KI-MDR-Lösungen, die dafür sorgen, dass Ihr Modell sicher und geschützt ist.

* Basierend auf einer internen Analyse von Dell, Oktober 2024 (Intel) und März 2025 (AMD). Gilt für PCs mit Intel und AMD-Prozessoren. Nicht alle Funktionen sind bei allen PCs verfügbar. Einige Funktionen müssen zusätzlich erworben werden. Von Principled Technologies validierte auf Intel Technologie basierende PCs, Juli 2025.

Fazit

Um robuste KI-Frameworks zu entwickeln, ist eine Zusammenarbeit zwischen Organisationen und Sicherheitsfachleuten von entscheidender Bedeutung. Da KI und LLMs die Branchen weiter umgestalten, ist es wichtig, sich mit den Risiken auseinanderzusetzen, die sie mit sich bringen, einschließlich Datensicherheit, Modellintegrität und Compliance-Herausforderungen. Unternehmen müssen proaktiven Strategien den Vorzug geben, die Sicherheit in jede Phase ihrer KI-Reise integrieren.

Dell Technologies ist ein vertrauenswürdiger Partner bei dieser Aufgabe und bietet eine durchgängige GenAI-Anpassung, Sicherheitsberatung und integrierte Lösungen, die auf Ihre individuellen Anforderungen zugeschnitten sind. Durch den Einsatz der robusten Cybersecurity-Lösungen von Dell können Unternehmen KI- und LLM-Risiken effektiv mindern und gleichzeitig das Potenzial ihrer bestehenden Sicherheitsinvestitionen maximieren. Dell ermöglicht es Unternehmen, ihre KI-Infrastruktur zu schützen, indem sie erweiterte Sicherheit nahtlos in ihre aktuellen Frameworks integrieren und so eine zukunftsfähige, sichere Umgebung sicherstellen.

Erfahren Sie, wie die umfassenden KI-Lösungen von Dell Ihre GenAI- und LLM-Umgebungen schützen können: Dell.com/CyberSecurityMonth