DELL Technologies

# 生成AIとLLMのサイバーセキュリティに関する懸念事項上位10件

開始する

# はじめに

人工知能(AI)は組織の運用方法に革命をもたらしており、生成AI (GenAI)と大規模言語モデル(LLM)は現代のエンタープライズ環境における重要なワークロードとなりつつあります。

他のワークロードと同様、これらのアプリケーションにも対処すべき固有の複雑さと脆弱性があります。企業がAIの導入を進めてイノベーションを促進し、効率性や競争優位性を高めていくにつれ、これらのアプリケーションのセキュリティを確保することは基本的な必須事項となります。優れたサイバー ハイジーンは、あらゆるワークロードを保護するための基盤です。すべてのワークロードでセキュリティに優先的に取り組むのと同じように、AIの優れたサイバー ハイジーンを実践することも不可欠です。これには、システムへの適切なパッチの適用、多要素認証、ロールベースのアクセス制御、ネットワークの区分化などのプラクティスを導入することなどがあります。これらの対策は基本的なものですが、重要なのは、これらの機能がワークロードの特定のアーキテクチャと使用方法にどのように適合するかを理解することです。

Dellは、AIワークロードとそれが直面するセキュリティ上の独自の課題に関する深い知識を有しています。Dellは、攻撃者がこれらのワークロードを標的とする場合に取る可能性のある方法を特定することで、堅牢なセキュリティ戦略の策定を支援できます。これには、トレーニング データのポイズニング、モデル窃盗やモデル改ざん、データセットの再構築などのリスクへの対処が含まれます。

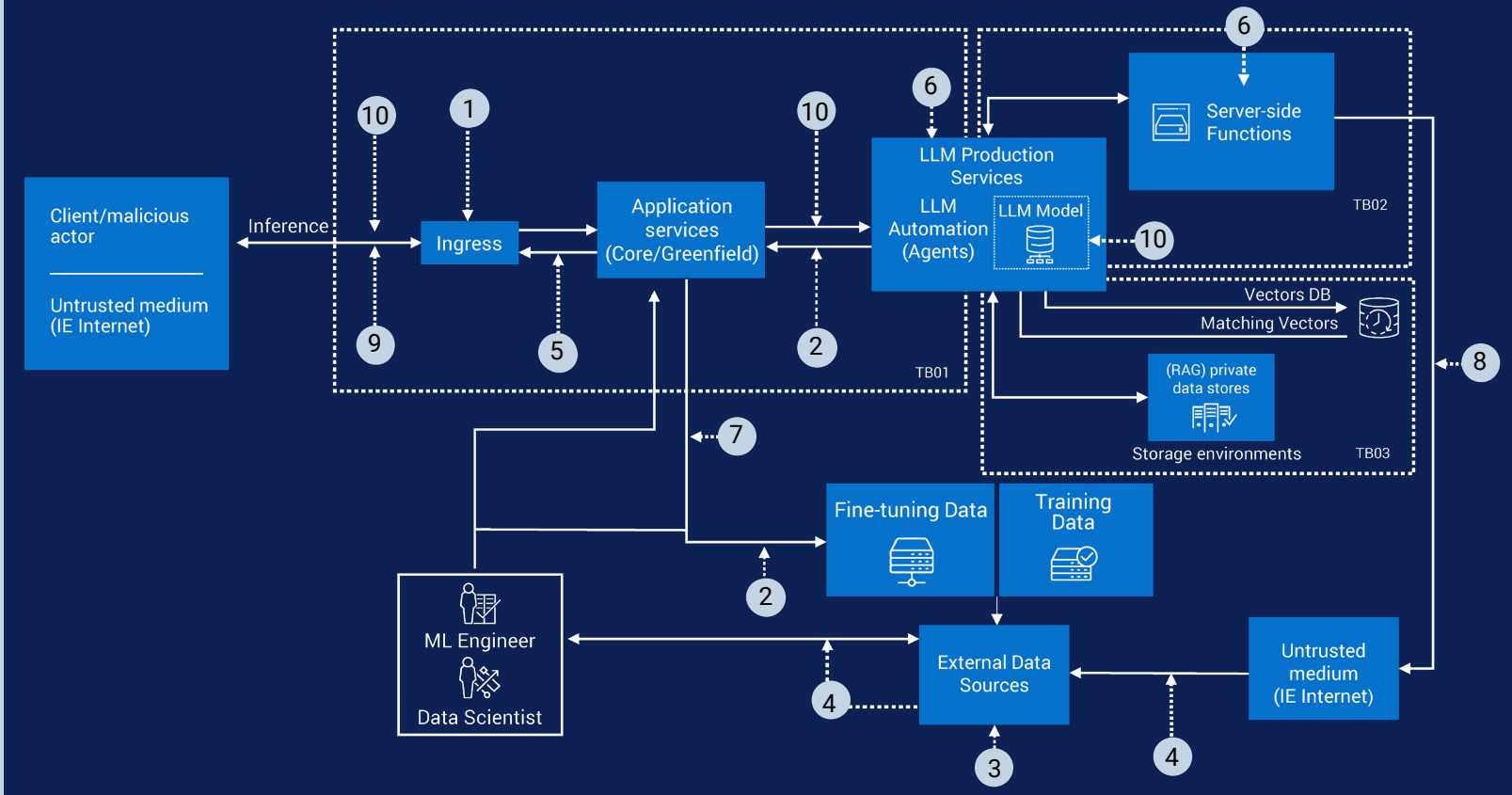
またDellは、AIモデルへの入力に関連する課題に対処することにも重点的に取り組んでいます。そうした課題には、機密情報の露出防止、安全でないトピックやバイアスの軽減、規制遵守の確保などがあります。出力の面では、モデルへの過度な依存やコンプライアンスに関するリスクなどの問題への対処を支援します。

Dellでは、既存のサイバーセキュリティ ソリューションを活用したり、システムを保護するための新しいツールやプラクティスを検討したりすることで、企業がこれらのリスクを軽減できるよう支援しています。当社の目標は、セキュリティがイノベーションの妨げにならないようにすることです。AIワークロードが機能する仕組みと、AIワークロードが直面するセキュリティの脅威を理解していることで、Dellはお客様がセキュリティ体制を強化し、環境のレジリエンスを高めながら、自信を持ってイノベーションを推進できるよう支援できます。当社の専門知識を活用し、お客様が堅牢なセキュリティを維持しながら、AIの可能性を確実に活かせるようサポートいたします。

# 生成AIとLLMのサイバーセキュリティに関する懸念事項上位10件

以下は、生成AIやLLMのモデルの保護に関してOWASPが示している主な懸念事項です。
**詳細については、各懸念事項をクリックしてください。**

1. プロンプト インジェクション
2. 機密情報の露出
3. サプライ チェーン
4. モデル データ ポイズニング
5. 不適切な出力処理
6. 過剰なエージェンシー
7. システム プロンプトの漏洩
8. ベクトルと埋め込みの弱点
9. 誤情報
10. 無制限の消費

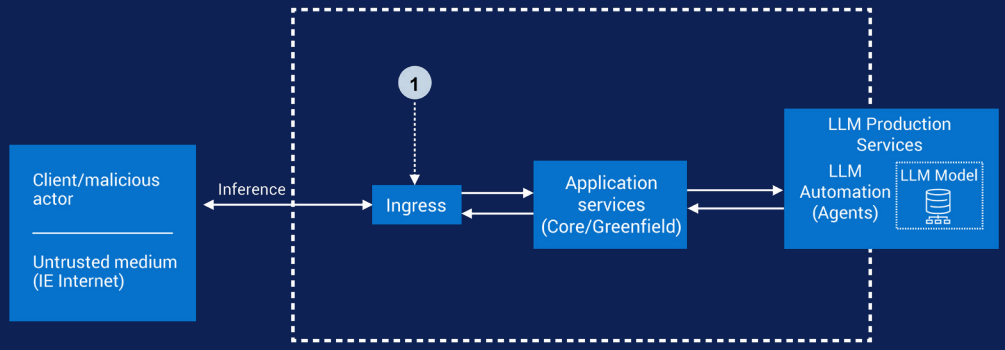


# 懸念事項1：プロンプト インジェクション

### プロンプト インジェクションを軽減するための戦略

- **データ サニタイズと入力の検証：**ユーザー入力を徹底的にスクリーニングして有害なコンテンツを排除します。正規化と符号化を使用して、悪用を防止します。
- **自然言語処理(NLP)と機械学習ベースのアプローチ：**NLPと機械学習を使用して、改ざんされたプロンプトや悪意のあるプロンプトを検出してブロックします。
- **明確な出力形式の設定と応答の制御：**厳格な応答の境界を設定し、出力が意図した形式に従うようにして、不正なアクションを防止します。プロンプトのフィルタリングと応答の検証を行って、整合性を維持します。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **安全なプロンプト エンジニアリング：**ソフトウェア セキュリティ全体の一環として、安全なプロンプト設計と分析を活用し、入力処理を保護します。
- **モデルの検証：**MLモデルを定期的に検証して、展開前に改ざんされていないことを確認し、正確性と整合性を保護します。
- **プロンプトのフィルタリング、ランキング、応答の検証：**プロンプトを分析して順位付けし、安全な入力のみが処理されるようにします。応答を検証して悪用を防止します。
- **堅牢性のチェック：**定期的な評価を行い、脆弱性を特定して修正し、AIの安全性と信頼性を維持します。

アーキテクチャ図に戻る



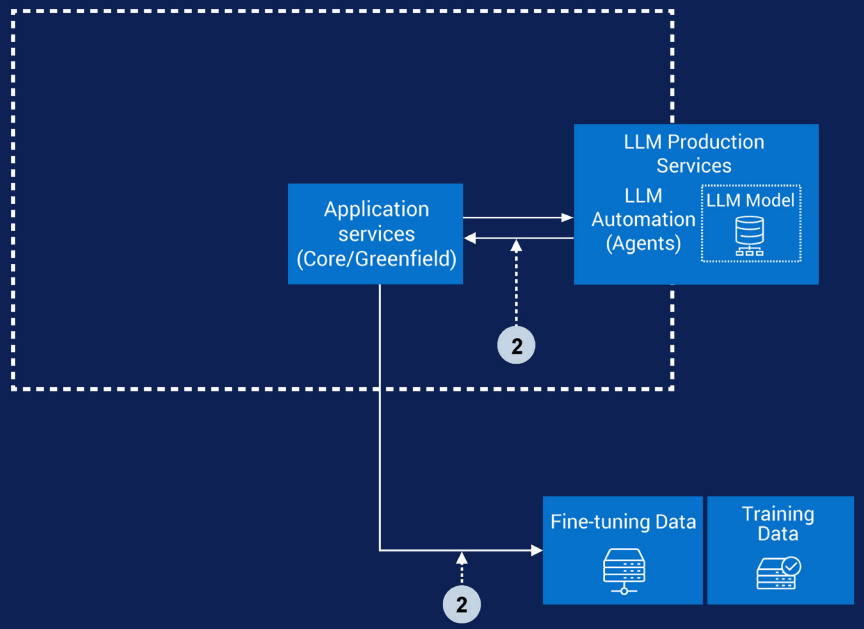

プロンプト インジェクションとは、生成AI (GenAI)の世界における新たな課題で、悪意のある入力を作成してモデルの動作を改ざんしたり、その整合性を損なったりするものです。この攻撃は、AIシステムがユーザー入力を処理して応答する方法の脆弱性を悪用するもので、不正なアクション、誤情報、機密データの露出につながる可能性があります。生成AIの重要なビジネス ワークフローへの統合が進む中、こうしたリスクに対処することは信頼とセキュリティの維持に不可欠です。

# 懸念事項2：機密情報の露出

### 機密情報の露出を軽減するための戦略

- **データ サニタイズと入力の検証：**ユーザー入力を徹底的にスクリーニングして有害なコンテンツを排除します。正規化と符号化を使用して、悪用を防止します。
- **準同型暗号を使用して**、内容を露出することなく、機密データを安全に処理します。これにより、使用中のデータであっても暗号化され、侵害から保護された状態を維持できます。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **安全なAPIとシステム インターフェイスを活用して**AIデータのやり取りを行い、構成を定期的に確認して、露出や攻撃対象領域を最小限に抑えます。
- **データの収集、保管、ポリシーを保護し**、データの保護とガバナンスの包括的なポリシーを適用して、法令遵守を確保し、データリスクを最小限に抑えます。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **安全な開発、構成、監査：**安全なコーディング プラクティスを適用し、自動構成管理ツールを使用するとともに、定期的なレビュー、監査、アップデートを実施して、AIシステムの構成を安全かつ最新の状態に保ちます。
- **ユーザー教育とセキュリティ意識の向上：**ユーザーと管理者にAIに特化した継続的なセキュリティ意識向上トレーニングを実施し、危険な使用や偶発的なデータの露出を減らします。

アーキテクチャ図に戻る



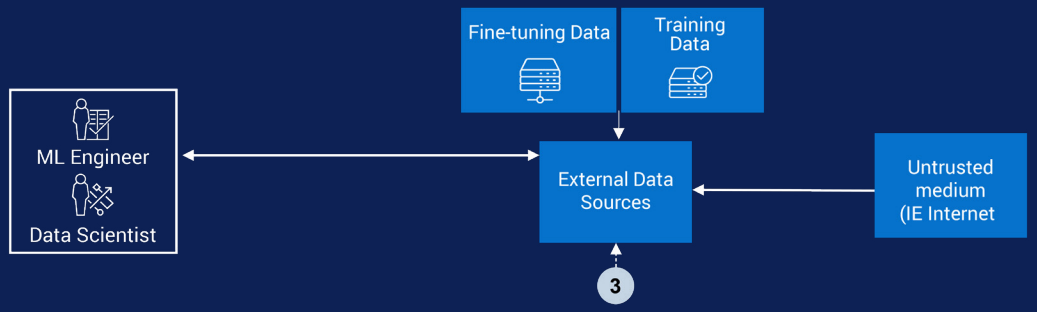

生成AIは驚異的な進歩をもたらしましたが、そこには重大なリスクも伴います。機密情報の意図しない露出はその代表例です。個人識別情報(PII)であっても、専有ビジネス データであっても、生成AIツールの誤用や取り扱いミスは、データ漏洩、法令遵守違反、評判の低下につながる可能性があります。そのため、組織がこうしたリスクを理解し、プロアクティブに対処して、AIシステムを安全に実装して使用することが重要です。

# 懸念事項3：サプライ チェーンの脆弱性

### サプライ チェーンの脆弱性を軽減するための戦略

- **サプライヤーを精査し、安全なサプライ チェーン プラクティスに準拠：**サプライヤーを評価し、サプライ チェーンのセキュリティに高い優先順位を置く契約を締結します。
- **ソフトウェア部品表を実装：**ソフトウェア コンポーネントの出所を追跡し検証して、透明性を確保し、コードの侵害のリスクを軽減します。
- **モデルの検証：**MLモデルを定期的に検証して、展開前に改ざんされていないことを確認し、正確性と整合性を保護します。
- **コンテナやポッドを最小権限で実行：**侵害が発生した場合の潜在的な影響を軽減し、不正アクセスを制限します。
- **ファイアウォールを導入：**不要なネットワーク接続をブロックして、潜在的な脅威につながる露出を減らし、攻撃者の侵入経路を制限します。
- **データとアノテーションを保護：**データとそれに関連するアノテーションを保護して、重要な情報に対する改ざん、不正アクセス、破損を防止します。
- **ハードウェアを保護：**セキュリティ検証済みのハードウェアを使用して、ハードウェアベースの攻撃の対象となり得る脆弱性を防止し、インフラストラクチャの強固な基盤を確保します。
- **MLソフトウェア コンポーネントを保護：**信頼性の高い精査済みのMLソフトウェア コンポーネントを使用して、脆弱性を軽減し、機械学習ワークフローの全体的なセキュリティを強化します。
- **安全な開発、構成、監査：**安全なコーディング プラクティスを適用し、自動構成管理ツールを使用するとともに、定期的なレビュー、監査、アップデートを実施して、AIシステムの構成を安全かつ最新の状態に保ちます。

アーキテクチャ図に戻る



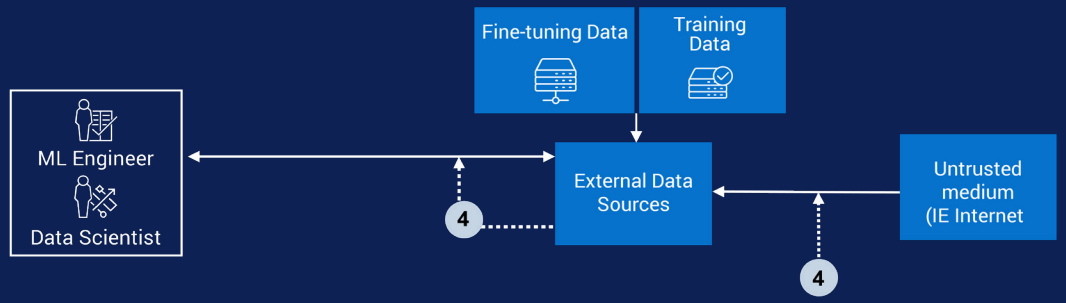

事前トレーニング済みモデルの整合性やサードパーティー製アダプターなどの重要なコンポーネントに影響を与える可能性がある、LLMサプライ チェーンの脆弱性を調査します。AIシステムが依存するハードウェアとソフトウェアの両方が、導入のかなり前の段階で侵害されている可能性があります。攻撃者は、機械学習サプライ チェーンのさまざまな段階の弱点を悪用する可能性があり、その標的はGPUハードウェア、データとそのアノテーション、MLソフトウェア スタックの構成要素、さらにはモデル自体にまで及びます。これらの固有の部分を侵害することで、攻撃者はシステムへの初期アクセスを取得し、セキュリティと整合性に重大なリスクをもたらす可能性があります。堅牢で安全なAIソリューションを構築するには、こうした脆弱性を理解して軽減することが不可欠です。

# 懸念事項4：モデル データ ポイズニング

### モデル データ ポイズニングを軽減するための戦略

- **トレーニング中に異常の検出とデータの検証を行い**、データの不整合を特定して対処し、クリーンで高品質のデータのみがモデルのトレーニングに使用されるようにします。
- **微調整フェーズ中には環境を分離して**微調整を行い、開発の重要な段階におけるモデルへの不正アクセスや汚染を防止します。
- **モデルの検証：**MLモデルを定期的に検証して、展開前に改ざんされていないことを確認し、正確性と整合性を保護します。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **データ サニタイズと入力の検証：**ユーザー入力を徹底的にスクリーニングして有害なコンテンツを排除します。正規化と符号化を使用して、悪用を防止します。
- **安全な開発、構成、監査：**安全なコーディング プラクティスを適用し、自動構成管理ツールを使用するとともに、定期的なレビュー、監査、アップデートを実施して、AIシステムの構成を安全かつ最新の状態に保ちます。
- **堅牢性のチェック：**定期的な評価を行い、脆弱性を特定して修正し、AIの安全性と信頼性を維持します。
- **ネットワーク セグメンテーションを実装して**、安全でないインターフェイスや重要なシステム コンポーネントへのアクセスを制限します。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。

アーキテクチャ図に戻る




モデル データ ポイズニングとは、AIライフサイクルにおけるセキュリティ上の脅威で、攻撃者が、間違いを含む入力、誤った結果に誘導する入力、悪意のある入力によって、トレーニング データを意図的に汚染するというものです。このリスクは、RAWデータの収集やアノテーションから、機械学習や大規模言語モデルに使用されるデータセットのキュレーションや統合まで、重要なコンポーネントに影響を与える可能性があります。AIシステムの信頼性は、データ ソースの整合性によって決まります。データ ソースは、トレーニング前や前処理中、または外部のデータ パイプラインを介して、改ざんされる危険があります。

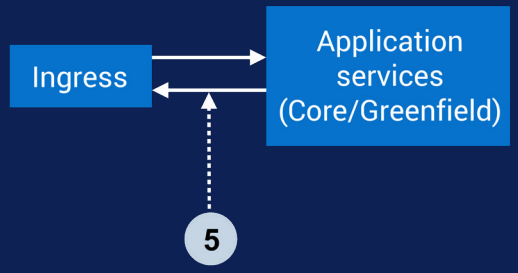
攻撃者は、データ ポイズニングを利用してモデルの精度を低下させたり、脆弱性を生じさせたり、有害な出力を引き起こしたりします。また、データの出所、アノテーション品質、データセットの取り込みプロセスの弱点を標的にして、セキュリティ、信頼性、レジリエンスを弱体化させる可能性があります。堅牢で信頼できるAIソリューションを構築するには、こうしたデータ起因の脅威を認識して軽減することが必須です。

# 懸念事項5：不適切な出力処理

### 不適切な出力処理を軽減するための戦略

- **コンテキスト認識型の出力エンコーディング：**HTML、SQL、API環境など、出力が使用される特定のコンテキストに合わせてカスタマイズされたエンコーディングやエスケープの手法を常に適用して、インジェクション攻撃などの脆弱性を防止します。
- **出力のサニタイズ：**Open Web Application Security Project (OWASP) Application Security Verification Standard (ASVS)のガイドラインに準拠して、モデル出力に対し厳格な検証とサニタイズ処理を実施することで、ダウンストリームでの安全な使用を確保し、セキュリティリスクを軽減します。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **出力セキュリティ テストの自動化：**自動化されたツールを使用して定期的なセキュリティ テストを実施し、クロスサイト スクリプティング(XSS)やインジェクションの脆弱性などの出力のリスクを特定して、プロアクティブに対処します。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **ヒューマンインザループのレビュー：**金融や医療などのリスクの高いアプリケーションでは、モデルの出力を人間が監視しレビューして、正確性、セキュリティ、安全性を確保する必要があります。
- **プライバシーとコンプライアンス：**プライバシー保護の手法を出力プロセスに組み込み、機密情報の安全な使用に関連する規制と基準を確実に遵守します。

アーキテクチャ図に戻る



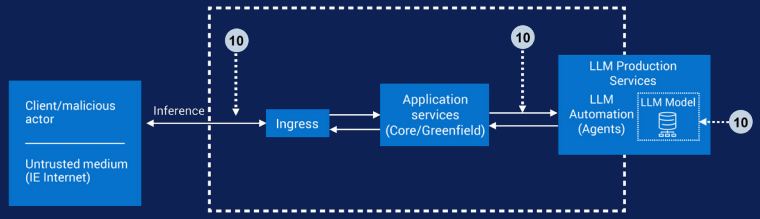

AIモデルの出力の検証やサニタイズが不十分な場合、権限昇格やデータ侵害など、深刻なセキュリティリスクを引き起こすおそれがあります。AIモデルが生成する出力に適切なチェックやフィルタリングが行われていないと、悪意のある攻撃者がそれらの脆弱性を悪用して不正なアクセスをしたり、システム内で自分の権限を昇格したりする可能性があります。こうした管理の欠如は、データ侵害、不正なアクション、重大なセキュリティ侵害につながる場合があり、AIが生成するあらゆる出力に対して堅牢な検証とサニタイズ プロセスを実施することの重要性が強調されます。

# 懸念事項6：過剰なエージェンシー

### 過剰なエージェンシーを軽減するための戦略

- **最小権限を適用：**LLMやエージェント型サブシステムには、目的の操作を実行するために必要な最小限の権限のみを付与し、アクセス制御を定期的に確認します。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **運用の範囲を設定：**LLMやエージェントがアクセスしたり実行したりできる範囲を明確に定義します。
- **ヒューマンインザループのレビュー：**金融や医療などのリスクの高いアプリケーションでは、モデルの出力を人間が監視しレビューして、正確性、セキュリティ、安全性を確保する必要があります。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **自律性を制限：**LLMの機能を制限して、無制限のアクセスや制御ができないようにします。
- **安全な開発、構成、監査：**安全なコーディング プラクティスを適用し、自動構成管理ツールを使用するとともに、定期的なレビュー、監査、アップデートを実施して、AIシステムの構成を安全かつ最新の状態に保ちます。
- **ファイアウォールを導入：**不要なネットワーク接続をブロックして、潜在的な脅威につながる露出を減らし、攻撃者の侵入経路を制限します。
- **堅牢性のチェック：**定期的な評価を行い、脆弱性を特定して修正し、AIの安全性と信頼性を維持します。

アーキテクチャ図に戻る



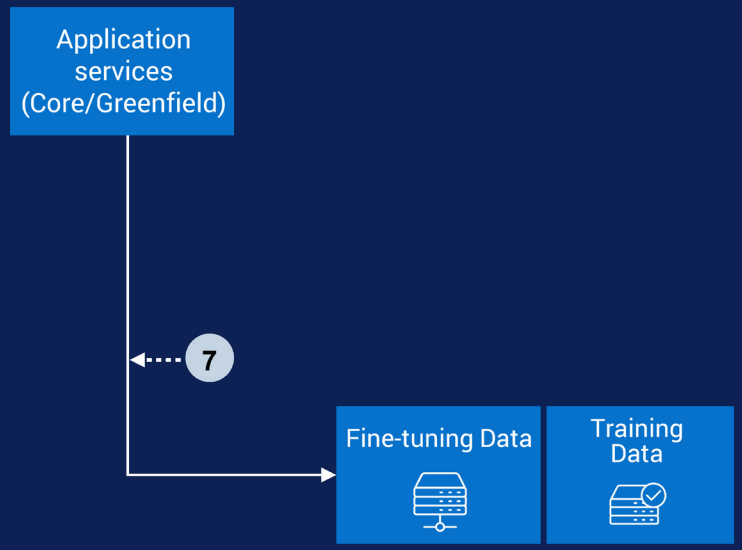

ワークフロー内でAIエージェントやプラグインに過度の自律性や不要な機能を与えると、重大なリスクを引き起こすおそれがあります。AIシステムに必要以上の権限や機能が付与されると、意図しない結果が生じる可能性が高まります。これは、大規模言語モデル(LLM)ベースのシステムに設計上過度の権限が与えられている場合に起こり得ることで、システムが本来認められるべきでない動作や情報アクセスを行えるようになります。こうした過剰な機能は、エラー、データの悪用、セキュリティの脆弱性につながる可能性があるため、AIの機能を慎重に制限し監視して、安全で責任ある使用を確保することの重要性が強調されます。

# 懸念事項7：プロンプトの漏洩

### プロンプトの漏洩を軽減するための戦略

- **プロンプトに機密情報を埋め込むことを避ける：**プロンプトに、認証情報、APIキー、独自ロジックを決して含めないでください。これらの情報はシステム外で安全に管理します。
- **プロンプトからセキュリティ制御を分離：**認証、承認、セッション管理は、プロンプトではなく、アプリケーション ロジックで処理します。
- **入力と出力を検証：**堅牢な検証によってプロンプトや応答をサニタイズし、疑わしいパターンや改ざんをブロックします。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **プロンプトを暗号化して保護：**プロンプトと構成は暗号化された安全なストレージに保存し、不正アクセスを防止します。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **プロンプトを定期的にレビュー：**プロンプトを定期的にレビューしてサニタイズすることで、機密データを削除してセキュリティ コンプライアンスを確保します。
- **弱点のテストとレッドチーミングを実施：**攻撃的なテストを実施し、プロンプト管理や出力の脆弱性を特定して修正します。
- **プロンプトをユーザー入力から分離：**ユーザー クエリーによってプロンプトが改ざんされたり、開示されたりしないようにシステムを設計します。
- **レート制限を適用：**APIの使用を制限して、不審なアクティビティを抑制し、自動的なプロンプト攻撃をブロックします。

アーキテクチャ図に戻る



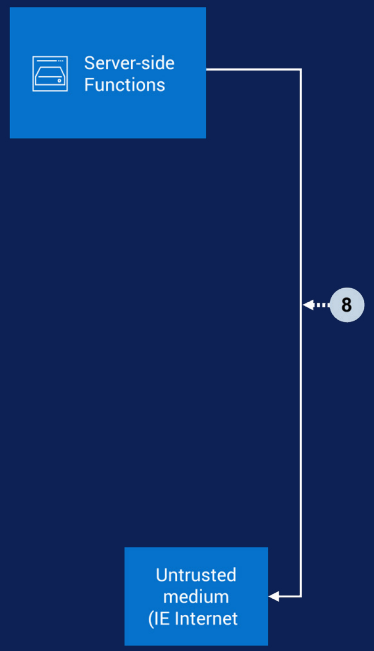

大規模言語モデル(LLM)やAIシステムに対するシステム プロンプト漏洩攻撃は、攻撃者が、モデルの動作をガイドし、運用上の境界を設定する隠れた命令（「システム プロンプト」）を抽出したり推測したりできる場合に発生します。こうしたプロンプトは、中核的なルールや制限事項、場合によっては機密性の高い運用ロジックを含んでいるため、通常はエンド ユーザーに表示されないようになっています。攻撃者は、特別に細工した入力や脆弱性の悪用によって、LLMを欺き、システム プロンプトの全体または一部を開示させることがあります。この情報は、漏洩した場合、制限をリバースエンジニアリングしたり、安全フィルターをバイパスしたり、新たな標的型攻撃を開発したりすることに使用でき、最終的には、プロンプト インジェクション、権限昇格、その情報の整合性に依存するモデルやダウンストリーム システムの悪用などのリスクが高まります。

# 懸念事項8：ベクトルと埋め込みの弱点

### ベクトルと埋め込みの弱点を軽減するための戦略

- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **暗号化：**転送中や保存中のベクトル データを、AESなどの堅牢な暗号化標準を使用して保護します。
- **安全な構成と監視：**システムを強化し、安全に構成するとともに、構成ミス、不正アクセス、異常を継続的に監視します。
- **脆弱性の管理：**すべてのソフトウェア、依存関係、ベクトル ストア エンジンを定期的にアップデートしてパッチを適用することで、セキュリティリスクに対処します。
- **データ サニタイズと入力の検証：**ユーザー入力を徹底的にスクリーニングして有害なコンテンツを排除します。正規化と符号化を使用して、悪用を防止します。
- **安全なAPIとシステム インターフェイスを活用して**AIデータのやり取りを行い、構成を定期的に確認して、露出や攻撃対象領域を最小限に抑えます。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **ハードウェアを保護：**セキュリティ検証済みのハードウェアを使用して、ハードウェアベースの攻撃の対象となり得る脆弱性を防止し、インフラストラクチャの強固な基盤を確保します。
- **安全な開発、構成、監査：**安全なコーディング プラクティスを適用し、自動構成管理ツールを使用するとともに、定期的なレビュー、監査、アップデートを実施して、AIシステムの構成を安全かつ最新の状態に保ちます。

アーキテクチャ図に戻る



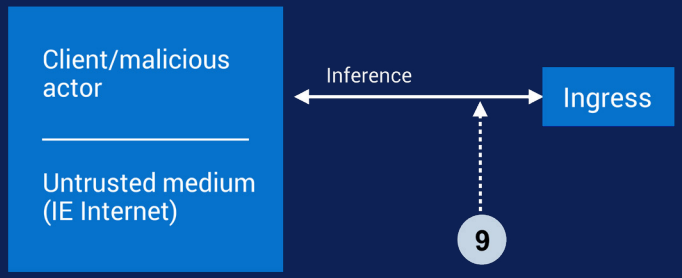

大規模言語モデル(LLM)やAIシステム、特に検索拡張生成(RAG)を使用するシステムに対するベクトルと埋め込みの弱点を突いた攻撃は、情報が数値ベクトルや埋め込みとしてエンコード、保存、取得される方法の脆弱性を標的にします。これらの仕組みの弱点は、埋め込みの反転（埋め込みからの機密データの再構築）、データ ポイズニング（有害なコンテンツやバイアスを含むコンテンツの注入によるモデルの動作の改ざん）、ベクトル データベースへの不正アクセス（データ漏洩につながる）、取得した出力の改ざんなどの悪意のある行為によって悪用される可能性があります。これらの攻撃は、攻撃者が機密情報を開示したり、出力を改造したり、AI主導型のアプリケーションに対するユーザーの信頼を損なったりすることで、プライバシー、整合性、信頼性を脅かします。こうした進化する脅威を防ぐには、適切なアクセス制御、データの検証、暗号化、継続的な監視が不可欠です。

# 懸念事項9：誤情報

### 誤情報を軽減するための戦略

- **信頼できるソースを使用した検索拡張生成(RAG)：**RAGを使用して、検証済みの信頼できるデータベースやナレッジ リポジトリーから情報を取得して統合し、ハルシネーションを軽減します。
- **モデルのチューニングと出力のキャリブレーション：**多様なデータセットを使用してモデルを微調整し、バイアスや誤情報を最小限に抑える手法を適用します。
- **自動ファクトチェック：**信頼性の高いソースを使用して出力を相互参照し、虚偽の情報に自動的にフラグを設定します。
- **不確実性の監視：**重要なケースについては、信頼性の低い応答にフラグを設定し、人間がレビューします。
- **ヒューマンインザループのレビュー：**金融や医療などのリスクの高いアプリケーションでは、モデルの出力を人間が監視しレビューして、正確性、セキュリティ、安全性を確保する必要があります。
- **ユーザーからのフィードバック：**ユーザーがエラーを報告できるようにして、モデルを継続的に改善し、誤った情報経路を迅速に修正します。
- **アクセス制限と人間による監視：**ロールベースのアクセス制御(RBAC)、多要素認証(MFA)、ID管理を適用して、アクセスを制限します。重要な決定出力には、人間によるレビューを行います。
- **安全な開発、構成、監査：**安全なコーディング プラクティスを適用し、自動構成管理ツールを使用するとともに、定期的なレビュー、監査、アップデートを実施して、AIシステムの構成を安全かつ最新の状態に保ちます。
- **リスク コミュニケーション：**AIの制限事項をユーザーに理解させ、ユーザー自身による検証を推奨します。
- **UIとAPIの意図的な設計：**AIが生成したコンテンツを強調表示し、ユーザーが責任ある使用をするようガイドします。

アーキテクチャ図に戻る



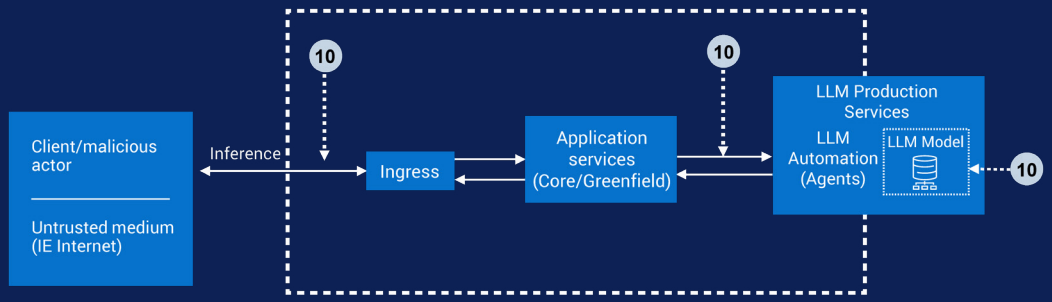

LLMやAIシステムに対する偽情報攻撃とは、誤った情報、誤解を招く情報、または一見信頼できそうな誤った情報を、モデルが出力を通じて生成し拡散するように仕向ける意図的な試みのことです。この脆弱性はいくつかの要素に起因します。モデルの「ハルシネーション」（捏造されたもっともらしいコンテンツの生成）を起こす傾向、トレーニング データに存在するバイアスやギャップ、敵対的プロンプトの影響などがその一例です。ハルシネーションは、LLMが事実を真に理解するのではなく、統計的にパターンに合うテキストを生成する結果、もっともらしく見えるが実際には根拠のない回答が導き出されることで、発生します。このような攻撃のリスクには、セキュリティ侵害や評判への悪影響、さらには法的責任までもが含まれます。これは、ユーザーが正確性や妥当性を検証せずにLLMの回答に過度に依存しているような環境では特に深刻で、重要な意思決定やプロセスにエラーや誤った情報が埋め込まれてしまう可能性があります。

# 懸念事項10：無制限の消費

### 無制限の消費に対する戦略

- **レート制限やユーザー クォータを適用：**ユーザー、APIキー、アプリケーションごとに要求、トークン、データの厳格な制限を設定して、乱用を防止します。
- **認証のユーザーの区分化を必須に設定：**強力な認証（APIキー、OAuthなど）を使用し、ロールや階層を割り当てて許可された要求以外は処理できないようにします。
- **入力の検証とサイズの制限：**プロンプトのサイズや構造を検証し、長大なクエリーや不正な形式のクエリーをブロックまたはトリミングします。
- **処理のタイムアウトやリソースのスロットリングを適用：**各要求のタイムアウトやリソースの上限を設定することで、長時間の操作やリソースの枯渇を回避します。
- **スマート キャッシュと重複排除を導入：**重複したクエリーや類似したクエリーへの応答をキャッシュに入れておき、不要な処理を削減します。
- **監視、ログ記録、異常検出：**AIシステムのアクティビティを継続的に監視しログを記録します。MDR、XDR、SIEMなどのソリューションを使用し、不正アクセス、異常、データ漏洩を迅速に検出して調査し、対応を行います。
- **予算の追跡と支出の制御：**ダッシュボードとアラートを使用して費用を監視し、予算のしきい値で使用をブロックします。
- **サンドボックス化と分離技術：**権限が制限されている隔離された環境でワークロードを実行することで、リスクを軽減します。
- **呼び出しの深さと会話のやり取り回数を制限：**再帰的な会話や会話ステップ数に制限を課して、悪用を防ぎます。
- **階層型モデルやリソース割り当てを適用：**優先度の高い要求はプレミアム モデルにルーティングし、優先度の低いトラフィックはコスト効率の良いモデルにルーティングします。

アーキテクチャ図に戻る




LLMやAIシステムの無制限の消費の脅威とは、悪意のあるユーザーであるかそうでないかに関わらず、過剰な推論要求やプロンプトを制御することなく送信することをアプリケーションが許可しており、有効なレート制限、認証、使用制限を課していない場合のセキュリティの脆弱性を指します。LLMの推論は大きな計算コストがかかるため、こうした制御の欠如はさまざまな方法で悪用される可能性があります。たとえば、攻撃者がシステム リソースを過剰に消費することでサービス拒否(DoS)を引き起こす、従量課金制またはクラウドホスト型の導入環境で予期しない経済的損失を引き起こす、モデルに体系的に問い合わせを行って動作を模倣し、知的財産を窃取するといった方法です。その結果、サービスの中断、他のユーザーに対するパフォーマンスの低下、財務的な負担、機密性が高いモデルの漏洩リスクの増大などが発生します。基本的に、無制限の消費は、リソース使用量が適切に管理されていない場合に発生し、LLMベースのアプリケーションが偶発的な悪用と意図的な悪用の両方の危険にさらされます。

# AIセキュリティにDellを選ぶ理由

Dellは、ハードウェア、ソフトウェア、マネージド サービスにまたがる包括的なアプローチを通じて、組織がAIモデルとLLMを保護できるよう支援します。サプライ チェーンからデバイス、インフラストラクチャ、データ、アプリケーションにいたるすべてにわたって、ゼロ トラストの原則に沿ったセキュリティが組み込まれています。Dellのソリューションは、MFA、RBAC、最小権限、継続的な検証などの機能を使用して、ポートフォリオ全体でサイバー ハイジーンを推進するように構築されています。この包括的な「セキュリティ バイ デザイン」アプローチにより、モデル窃盗、データ漏洩、敵対的攻撃など、さまざまな高度なサイバー脅威によるリスクを最小限に抑え、組織は自信を持ってAIとLLMを活用したイノベーションを進めることができます。

## サプライ チェーン

Dellのセキュアなサプライ チェーンは、製品開発、製造、デリバリーのあらゆる段階にセキュリティを組み込むことで、AIモデルとLLMを根本から保護します。Dellは、暗号署名付きのBIOSとファームウェアのアップデート、Secured Component Verification、AIに重点を置いたソフトウェア部品表(SBOM)、データセットの系統追跡、統合されたセキュリティ ソフトウェアと構成、グローバル規格に沿った厳格なベンダー リスク アセスメントを通じて、改ざん、不正アクセス、サプライ チェーン攻撃のリスクを最小限に抑え、完全な透明性、整合性、法令遵守を備えた、信頼性が高くレジリエンスに優れたAIワークロードを導入できるよう支援します。

## AI PC

Dellは、オンデバイスAIワークロードの基盤となるセキュリティを確立します。Dell Trusted Devicesは、世界で最も安全なビジネス向けAI PC*で、セキュリティを念頭に置いて設計されています。サプライ チェーンのセキュリティで、製品の脆弱性や改ざんのリスクを最小限に抑えています。ハードウェアとファームウェアに直接組み込まれた独自の防御機能が、パソコンとエンドユーザーを使用中に保護します。Dell SafeBIOSは、BIOSレベルの詳細な可視性と改ざん検出機能を提供。Dell SafeIDが認証情報のセキュリティを強化し、パスワード不要の認証を可能にします。パートナー ソフトウェアにより、エンドポイント、ネットワーク、クラウド環境全体で高度な保護を利用できます。

## サイバー レジリエンス

DellのPowerProtectサイバー レジリエンス ソリューションは、暗号化されたイミュータブル バックアップ、迅速なリストア、隔離されたサイバー リカバリー ヴォールトによってAIデータを保護します。これらの機能は、破壊を防ぎ、悪意のあるアップデートによる影響を軽減して、攻撃後のコンプライアンスとリカバリーをサポートします。

## サーバー

PowerEdgeサーバーは、機密コンピューティングによりAIやLLMのプロンプトと埋め込みを分離して保護します。確かなソースに支えられた信頼性の高い検索拡張生成(RAG)ソリューションに加え、MFA、RBAC、シリコン ルート オブ トラスト、署名済みファームウェアを備えており、継続的な監視により重要なAIワークロードを保護します。

## ストレージ

Dellのストレージ ポートフォリオは、保存データと転送中データの堅牢なAES-256暗号化により、機密AIデータを安全に暗号化して保存します。一部の製品は、将来的な量子脅威に対するレジリエンスを備えた高度な暗号化に対応しています。このポートフォリオには、高速NVMeパフォーマンス、データ（AIワークロードで使用されているデータも含む）を保護するFIPS準拠の暗号化モジュール、不変スナップショット、ランサムウェア攻撃に対抗するエアギャップされたサイバー リカバリー ヴォールトが含まれています。ゼロ トラスト アーキテクチャ、サプライ チェーン セキュリティ、改ざん防止監査機能により、ガバナンスを強化。組み込みの異常検出とAIOps MLモデルは、お客様のデータをトレーニングに使用することなくワークロードを保護するため、入力ベースの攻撃リスクを最小限に抑えられます。

## AIOps

Dell AIOpsは、自動化された継続的な監視によって構成ミスや脆弱性（CVEを含む）を検出し、AIやLLMのワークロードに影響を与えるサプライ チェーン リスクの認識をサポートします。リアルタイムのCVEスキャン、スマート アラート、AIを活用したダッシュボードにより、異常にフラグ付けを行って解決のワークフローを追跡することで、迅速な介入を支援します。組み込みのコンプライアンス機能、ロールベースのアクセス制御、自動レポート作成により、ワークロード全体で安全な運用を維持できるようにし、EDRとXDRのシームレスな統合と、対応ソリューションにおける生成機能を含むAI主導型の運用インサイトにより、ITの効率がさらに向上されます。

## ネットワーキング

Dell Networkingソリューションは、堅牢なネットワーク セグメンテーションを通じてAIやLLMの環境を保護し、ラテラル ムーブメントを最小限に抑えます。暗号化されたネットワーク パスと統合されたファイアウォール制御により、AIデータへの不正アクセスをブロックします。

## AIのセキュリティとレジリエンスに関するサービス

DellのAIセキュリティおよびレジリエンスに関するサービスは、AIの組織への統合に結び付いた新たなリスクに対処するように設計されています。Dellのサービスは、お客様のチームと連携して、AIをできる限り迅速にオンボーディングできるように構築されており、戦略計画、ソリューションの実装、マネージド セキュリティ サービスをガイドする専門知識を提供し、運用上の負担を軽減して、AIを活用して安全にイノベーションを実現します。そのそれぞれが、組織が進化するAIリスクに対処し、安全なAI導入を最適化できるようにカスタマイズされます。

## Dell AI Factory

Dellの安全なサプライ チェーン、ゼロ トラスト機能による最小権限の適用、モデルを安全に保護するよう設計されたAI MDRソリューションなど、目的に特化したセキュリティの統合ポートフォリオです。

*2024年10月（インテル）および2025年3月(AMD)のDell社内分析に基づきます。インテルとAMDのプロセッサー搭載のパソコンに当てはまります。一部のパソコンでは利用できない機能があります。一部の機能については追加購入が必要です。インテルベースのパソコンについてはPrincipled Technologiesによって検証されています（2025年7月）。

# まとめ

レジリエンスに優れたAIフレームワークを構築するには、組織とセキュリティ専門家による協働的なアプローチが不可欠です。AIとLLMが業界を再構築し続ける中、データ セキュリティ、モデルの整合性、コンプライアンスの課題など、それらがもたらすリスクに対処することが極めて重要です。組織は、AI導入のあらゆる段階にセキュリティを統合するプロアクティブな戦略を優先させる必要があります。

デル・テクノロジーズは、このミッションにおいて信頼できるパートナーとして、エンドツーエンドでのGenAIのカスタマイズ、セキュリティ コンサルティング、統合ソリューションを、お客様固有のニーズに最適化して提供します。Dellの堅牢なサイバーセキュリティ ソリューションを活用することで、既存のセキュリティ投資の可能性を最大限に引き出しながら、AIとLLMのリスクを効果的に軽減することができます。Dellは、高度なセキュリティを現在のフレームワークにシームレスに統合し、未来に対応した安全な環境を確保することで、組織がAIインフラストラクチャを保護できるよう支援します。

Dellの包括的なAIソリューションで生成AI環境とLLM環境を保護する方法をDell.com/CyberSecurityMonthでご確認ください