資料４-２

Microsoft Security

MicrosoftのAIサービスにおける セキュリティ対策

日本マイクロソフト株式会社

Microsoft 365 Copilot における攻撃対策

- AIサービスに対する攻撃の影響軽減するために、Microsoft は多層防御戦略を適用します。この戦略には「データ漏えいベクトル」「敵対的な誤用」「未承認のアクセスパターンの継続的な監視」が含まれます
- Microsoft 365 Copilot によって生成されたコンテンツは、他の Microsoft 365 コンテンツと同じアクセス制御とコンプライアンスポリシーによって管理されます。つまり、ユーザーのアクセス許可、秘密度ラベル、条件付きアクセス ポリシーは、コンテンツの生成とアクセスの時点で適用されます

https://learn.microsoft.com/ja-jp/copilot/microsoft-365/microsoft-365-copilot-ai-security

Microsoft 365 Copilot における攻撃対策

インジェクション防御

Microsoft は、迅速な挿入のリスクを軽減するために、Microsoft 365 Copilot プロンプト フロー全体で多層防御戦略を採用しています。 既定でアクティブであり、セットアップを必要としない保護機能の例を次に示します。

- ユーザーがループ内設計を使用すると、ユーザーは AI によって生成されたコンテンツを確認、変更、または拒否できます
- スパム、詐欺、不審なコンテンツ フィルタリングは、プロンプトで悪意のある指示、フィッシング詐欺、不正な資料をブロックするのに役立ちます
- Microsoft 365 Copilot は、迷惑メールや信頼されていないMicrosoft Teamsチャット (外部連絡先からのチャットなど) を無視します
- Microsoft 365 Copilot では、web ブロックBing尊重して、Web 検索中にアダルト サイト、低機関サイト、悪意のあるサイトを除外します
- Microsoft 365 Copilot はステートレス LLM アーキテクチャを使用して動作します。 要求は、テナント スコープの セマンティック インデックス作成を 使用してリアルタイムで処理され、データ アクセスと関連性がユーザーの組織のコンテキストに厳密に限定されるようにします

https://learn.microsoft.com/ja-jp/copilot/microsoft-365/microsoft-365-copilot-ai-security

Microsoft 365 Copilot における攻撃対策

データ流出防止

Microsoft 365 Copilot の階層化されたセキュリティ モデルは、次のようなデータ流出の可能性があるシナリオなど、従来および新たな脅威に対処します

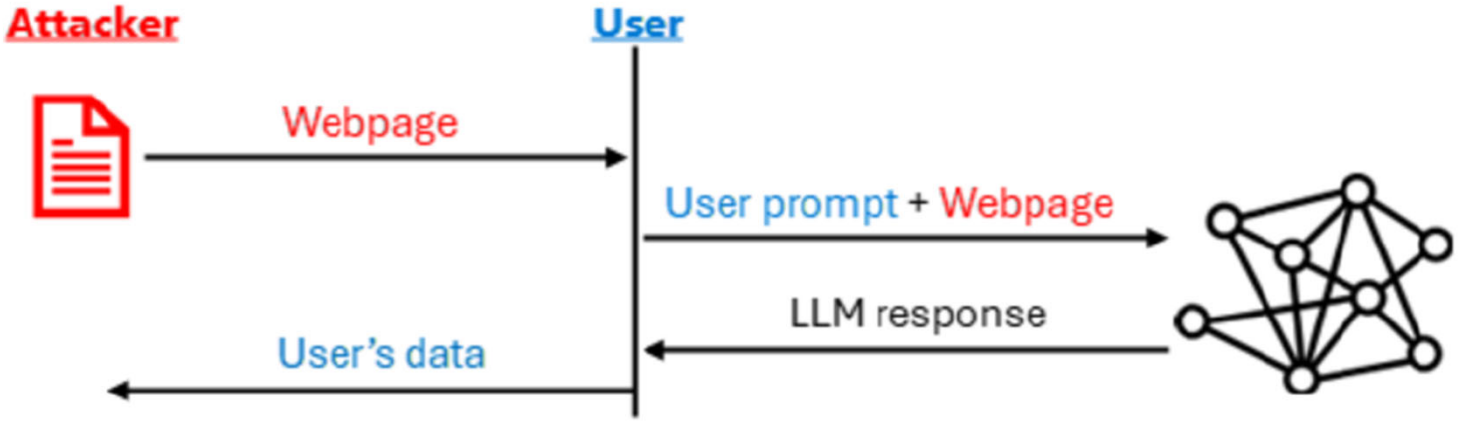
- 認証されていないイメージ URL。ユーザーが機密データを含むイメージを生成し、ブラウザーツールを使用して URL を抽出した後、画像を外部で共有します。 認証なしでイメージにアクセスできる場合は、 条件付きアクセス や 秘密度ラベルなどのエンタープライズコントロールをバイパスできます
- あるテナントのユーザーが悪意のあるイメージを生成し、別のテナントのユーザーとQR コードなどの悪意のあるイメージによって、匿名 URL を共有する。URL が認証によって保護されていない場合、アクセス制御が適用されない可能性があります

https://learn.microsoft.com/ja-jp/copilot/microsoft-365/microsoft-365-copilot-ai-security

Microsoft 365 Copilot における攻撃対策

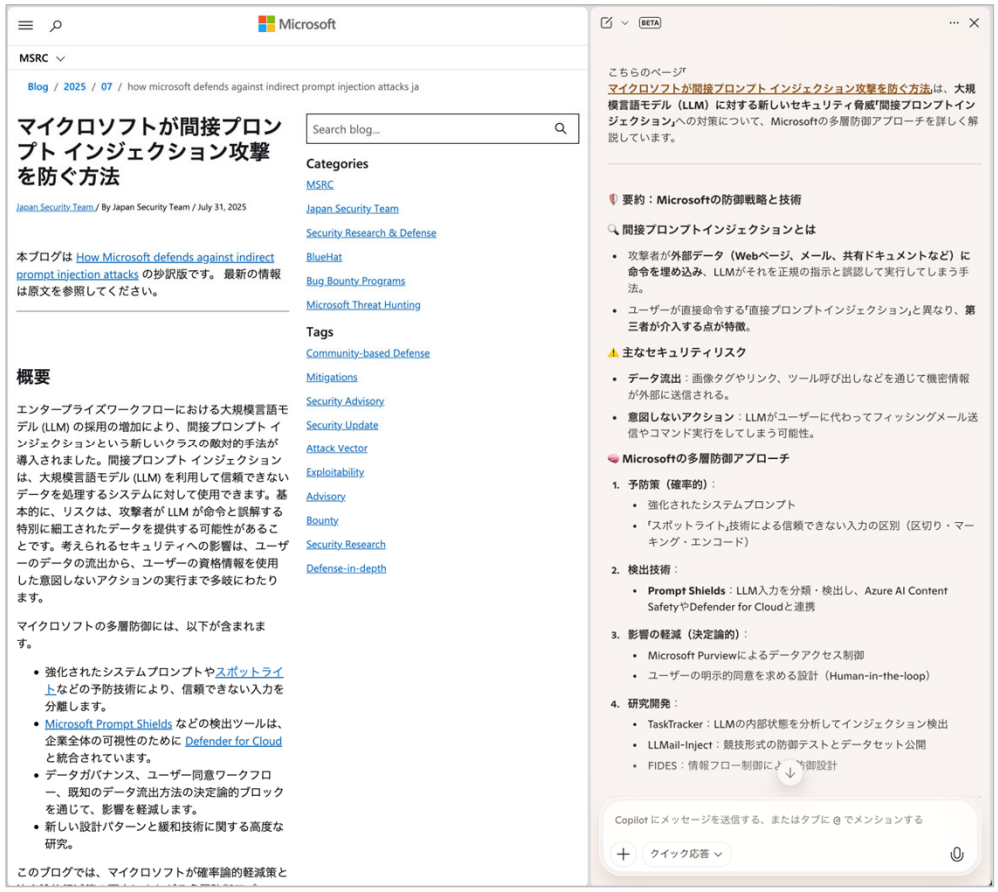
間接プロンプトインジェクション防御

エンタープライズワークフローにおける大規模言語モデル (LLM) の採用の増加により、間接プロンプト インジェクションという新しいクラスの敵対的手法が導入されました。間接プロンプト インジェクションは、大規模言語モデル (LLM) を利用して信頼できないデータを処理するシステムに対して使用できます。基本的に、リスクは、攻撃者が LLM の命令だと誤解する特別に細工されたデータを提供する可能性があることです。考えられるセキュリティへの影響は、ユーザーのデータの流出から、ユーザーの資格情報を使用した意図しないアクションの実行まで多岐にわたります。



https://www.microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks-ja

Microsoft 365 Copilot における攻撃対策



例えば、左の図のようにWebページの内容を参照しながら「このページを要約してください」というプロンプトを実行した場合、このページになんらかの仕掛けがされていた場合に、プロンプトにそれを含んでしまう可能性があるという課題があります
このような攻撃に対して、以下の対応を提案しています

- 入力時点だけではなく、処理前のプロンプトのチェック（Prompt Shields）
- データアクセス制御（Purview）
- ユーザーの明示的同意を求める設計（Human-in-the-loop）

様々な段階でのプロンプトチェックとアクセス制御でデータ流出や権限奪取を防御します

https://www.microsoft.com/en-us/msrc/blog/2025/07/how-microsoft-defends-against-indirect-prompt-injection-attacks-ja

プライバシー・バイ・デザインとコンプライアンス

Microsoft 365 Copilot は、「 Microsoft 365 Copilot のデータ、プライバシー、セキュリティ」で説明されているプライバシーとコンプライアンスの基準に準拠しています。セキュリティ制御によって適用される保護には、次のものがあります。

- データ アクセスの適用
- 暗号化と分離
- コンプライアンス ツール
- AI ライフサイクル全体でデータを保護する
- EU データ境界
- AI ワークロードのクロスクラウド ガバナンス
- ポリシーの統合と適用

https://learn.microsoft.com/ja-jp/copilot/microsoft-365/microsoft-365-copilot-ai-security

AI開発、運用基盤におけるマルチモデル活用

Microsoft AI Foundryでは、「責任あるAI（Responsible AI）」の原則に基づき、モデルの品質・安全性・セキュリティ・ガバナンスを多層的に評価し、信頼できるモデルやサービスを選定しています

1. **安全性とセキュリティ**
   - ユーザーや組織に害を及ぼさないよう、プロンプトインジェクション対策や情報漏洩防止を実装
   - Azure AI Content Safetyなどと連携し、有害コンテンツの検出と遮断を行う
2. **説明可能性（Explainability）**
   - AIの判断根拠を明示できるよう、透明性のある設計を採用
   - TaskTrackerなどの技術で、LLMの内部状態を可視化・分析し、信頼性を担保
3. **責任あるAI（Responsible AI）**
   - MicrosoftのAI原則（公平性、安全性、プライバシー、包括性、説明責任）に準拠
   - Human-in-the-loop（人間の介在）を重視し、自律性と制御のバランスを取る
4. **情報フロー制御と権限管理**
   - FIDESなどの技術により、AIがアクセス・処理できる情報の範囲を制限
   - Microsoft Purviewと連携し、データガバナンスとコンプライアンスを確保
5. **脅威テストによる評価**
   - LLMail-Injectなどの攻撃シナリオを用いた防御テストを実施
   - 実際の脅威に対する耐性を評価し、信頼性の高いエージェントを選定

AIセキュリティは攻撃ベースではなくガバナンスで対策する

攻撃ベースの対策は後追いになるばかりではなく、対策の隙間を産んでしまい、それを攻撃者に狙われてしまいます

データフローやアクセスコントロールを管理するのはもちろん、システムやサービス全体のガバナンス（全てを把握し、課題を発見したら即時修正する）を徹底することで、将来の攻撃対策を実施できます

最近ではこれをAttack Surface Managementなどといって、一部の対応にとどまることがありますが、本来の攻撃可能性、つまり脆弱性の露出について包括的に考えて対応することが重要です