生成AIを活用するシステムへの効果的なセキュリティ監視の導入

生成AI技術がもたらす新たなセキュリティリスクは従来のITセキュリティとはまったく異質なものになる。非決定的な手法も採用した継続的な監視、多層的な防御戦略、そして最新の脅威に対応するためのインテリジェンス更新が生成AIを安全かつ効果的に活用する条件となる。

生成AIシステム固有のセキュリティリスク

ChatGPTに代表される生成AI技術の急速な進化と普及により、多くの企業がこの革新的な技術を活用し始めている。しかし、その利便性と引き換えに、従来のITシステムとは異なる新たなセキュリティリスクが浮上してきている。

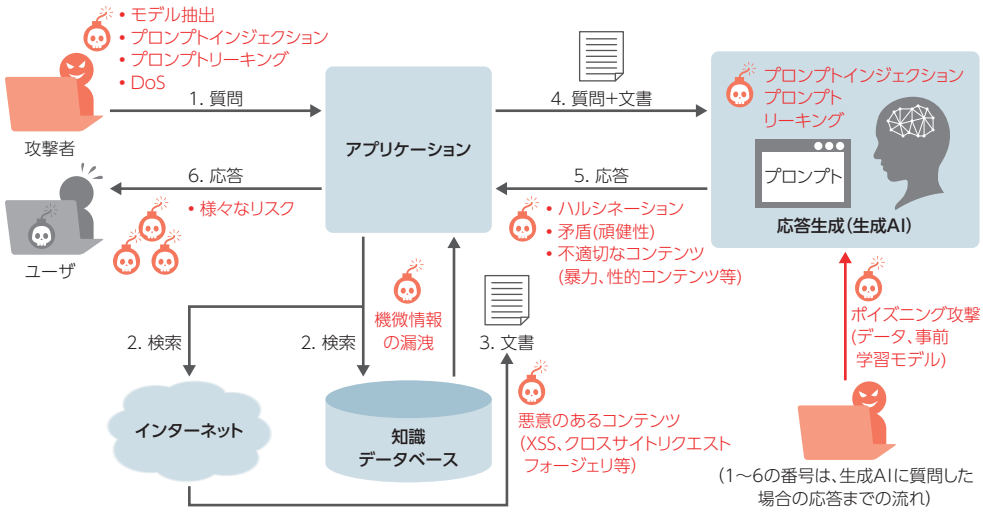
生成AIを活用したシステム（以下、生成AIシステム）、特に大規模言語モデル（LLM）を利用したシステムは、多様なセキュリティリスクにさらされている。プロンプトインジェクション攻撃[1]によるAIの操作や、プロンプトリーキング攻撃[2]等の内部設定や機密情報の不正取得といった攻撃。また、AIが不適切なコンテンツを生成したり、学習データに含まれる機密情報を意図せず出力したりするリスクも存在する。さらに、AIが事実と異なる情報を、自信を持って提示する「ハルシネーション（幻覚）」や、学習データのバイアスを反映して特定のグループに不利な結果を生成するといった公平性の問題も懸念されるところだ。

これらのリスクは、従来のITシステムにおけるセキュリティ対策では十分に対処できない新たな脅威といえよう。例えば、昨今企業での導入が進むRetrieval-Augmented Generation（RAG）をみてもユーザー入力から最終出力までの各段階で上述したリスクが複雑に絡み合っている（図表）。

従来のセキュリティ対策との違い

また、生成AIシステムのセキュリティ対策が従来のITセキュリティに比べて難易度が高いのは、入力が自然言語であるという点に起因している。自然言語による入力は極めて柔軟で予測困難なため、従来のような決定的なルールベースのセキュリティ対策では、攻撃の検知や防御が困難となる。例えば、SQLインジェクション攻撃に対しては特定のパターンや構文を検出することで防御できるが、プロンプトインジェクションの場合、攻撃の形態が多様で正常な入力との区別が難しいため、同様のアプローチでは十分な防御ができ

図表　LLMを利用したシステムにおける脅威イメージ



（出所）野村総合研究所

NOTE

1) LLMを操作しようとする技術で、通常の指示の中に隠れた命令を挿入することで、意図しない動作をさせることを目的とする。
2) LLMにプロンプトの一部または全体を開示させようとする攻撃手法。通常、システムプロンプトやその他の機密情報を含む可能性のある指示を抽出することを目的とする。
3) 既知の攻撃パターンや不正な入力のデータベース（シグネチャ）と、入力されたテキストの類似度を計算する手法。完全一致でなくても、類似した攻撃を検出できる利点がある。
4) テキストの意味や意図を理解するための分析手法。単純な単語マッチングではなく、文脈や言語の構造を考慮して、テキストの真の意味を解釈しようとする。

ない。また、生成AIシステムは入力に対して非決定的に応答を生成するため、出力の予測や制御も困難である。これは、従来のシステムでは想定されていない新たな課題となっている。

効果的な監視アプローチの重要性

このような課題に対処するためには、従来の決定的手法と新たな非決定的手法を効果的に組み合わせたアプローチが有効だと考える。

まず、従来のような決定的アプローチ（正規表現によるテキストマッチ等）が有効な場合は、それらを積極的に活用する。例えば、既知の攻撃パターンや特定のキーワードの検出には、このアプローチが効率的かつ効果的である。しかし、生成AIの柔軟性と予測困難性を考慮すると、決定的アプローチだけでは十分とはいえない。

そこで、決定的アプローチでは防ぎきれないリスクに対しては、非決定的アプローチを多段階で適用する必要がある。これには、シグネチャとの類似度比較[3]、セマンティック分析[4]などが含まれる。また、入力と出力のどちらも監視することが重要だ。入力監視では主にプロンプトインジェクションや不適切な要求を、出力監視では機密情報の漏洩や有害コンテンツの生成、ハルシネーション、後続の処理で危険な攻撃につながる可能性のある出力を検出する。さらに、入出力の相関分析により、より高度な異常パターンの検出も可能となる。

またこうしたアプローチを継続するためにもインテリジェンスの更新も不可欠である。生成AIの技術進化や新たな攻撃手法に迅速に対応するため、脅威情報の収集・分析、攻撃手法のシミュレーション、検知ルールの更新を行う。また、インシデント対応から得られた知見を活用し、システムの防御能力を継続的に強化することも重要である。

包括的なセキュリティ戦略が不可欠

効果的な生成AIシステムのセキュリティには、スポットでの診断と継続的監視を組み合わせた包括的アプローチが不可欠だ。システムの開発や更新時には、AIモデル、プロンプト、学習データ、アクセス制御、入出力フィルタリングの有効性といった観点でのセキュリティ診断が重要である。しかし、診断だけでは変化する脅威に対応することはできない。そこで、リアルタイムの入出力分析、異常検知と自動対応などの継続的な監視体制が必要となる。この監視から得られた知見を活用し、インシデント分析、脅威パターンの学習、セキュリティポリシーの更新を通じて、防御能力を段階的に向上させるのである。

さらに、組織全体でのセキュリティ意識とAIリテラシーの向上も重要である。開発者、運用者、エンドユーザーが生成AIのリスクと適切な使用法を理解することで、より安全なAI活用が実現する。

このように、技術的対策と組織的取り組みを統合した包括的なセキュリティ戦略が、生成AIシステムの安全な運用には不可欠といえよう。

Writer's Profile

田篭 照博 Teruhiro Tagomori
NRIセキュアテクノロジーズ
チーフエキスパートセキュリティエンジニア
専門はAIセキュリティ
focus@nri.co.jp