Nomura Research Institute Group

NEWS RELEASE

2023 年 12 月 18 日

NRI セキュアテクノロジーズ株式会社

NRI セキュア、生成 AI を活用したシステム向けの セキュリティ診断サービス「AI Red Team」を提供開始

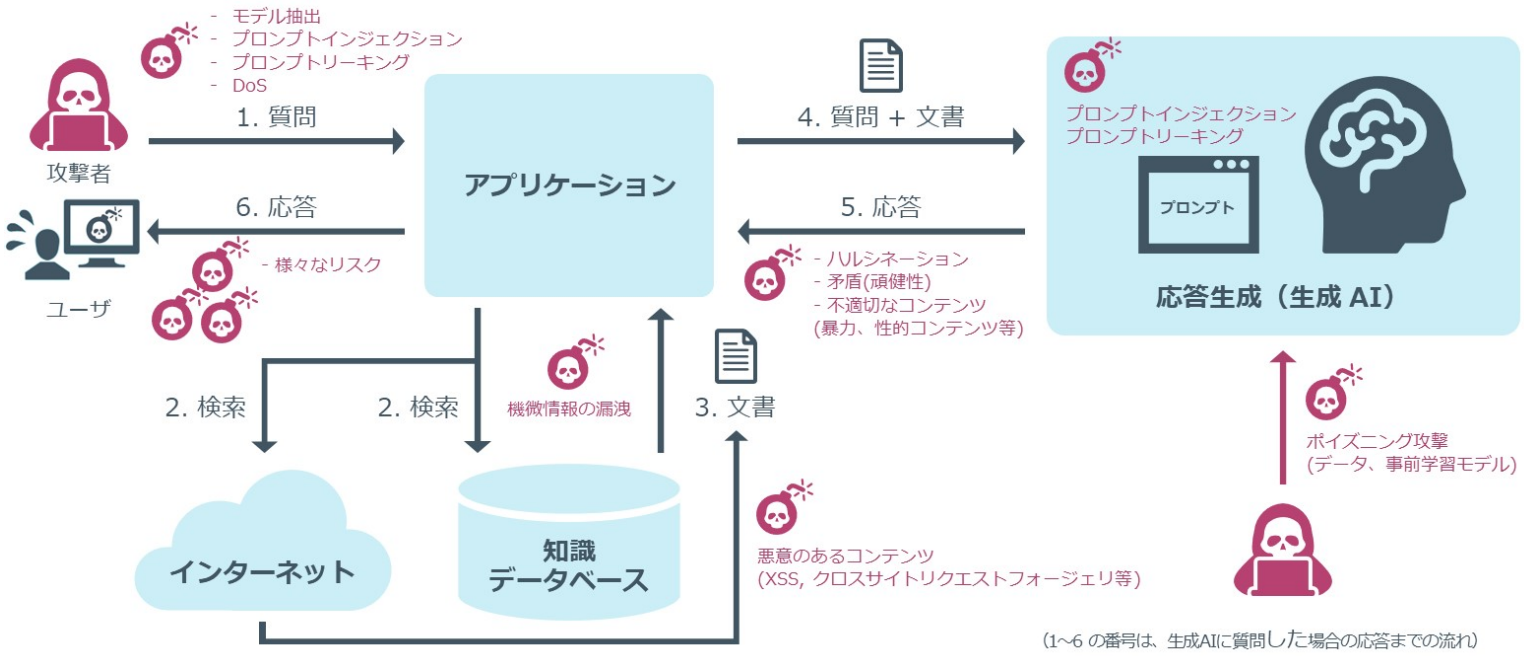
～ リスクベースアプローチで大規模言語モデル（LLM）とシステム全体を 2 段階で診断 ～

NRI セキュアテクノロジーズ株式会社（本社：東京都千代田区、代表取締役社長：建脇 俊一、以下「NRI セキュア」）は、「AI セキュリティ統制支援」サービスのラインナップの一つとして、新たに生成 AI を利用するシステムやサービスを対象にした、セキュリティ診断サービス「AI Red Team（以下、本サービス）」の提供を本日開始します。

■ AI が抱えるセキュリティ上の問題点

近年、多くの分野で生成 AI、とりわけ大規模言語モデル（LLM）[1] の利用が増加の一途をたどっています。LLM への期待が高まる一方で、LLM には、「プロンプトインジェクション」[2]、「プロンプトリーキング」[3] と呼ばれる脆弱性や、「ハルシネーション」[4]、「機微情報の漏洩」、「不適切なコンテンツの生成」、「バイアスリスク」[5] といったリスクが存在することが浮き彫りになってきました（図を参照）。LLM を利活用する企業は、これらの問題を把握し、適切な対策を施す必要があります。昨今、生成 AI を用いたシステムやサービスに特化したセキュリティ診断の重要性が叫ばれており、諸外国では独立した外部の専門家による診断の必要性について言及され始めています。

図：LLM を活用したシステムのイメージとリスクの例



■ 本サービスの概要と特長

本サービスでは、NRI セキュアの専門家が実際のシステムに擬似攻撃を行うことで、LLM を活用したサービスにおける AI 固有の脆弱性と、その AI と連携する周辺機能を含めたシステム全体の問題点を、セキュリティ上の観点から評価します。

AI は、それ単体でサービスとして機能することはなく、周辺機能と連携することで具体的な提供サービスを構成します。そのため、LLM 単体のリスクを特定するだけではなく、サービス全体を俯瞰して、「リスクが顕在化した場合に、システムやエンドユーザに悪影響を与えるかどうか」というリスクベースアプローチでも評価する必要があります。

本サービスでは、LLM 単体でのセキュリティリスクを洗い出し、LLM を含むシステム全体を評価するという、2 段階に分けた診断を実施します。診断の結果、見つかった問題点と緩和策をまとめた報告書を提供します。

本サービスの主な特長は、以下の 2 点です。

1. 独自開発した自動テストと専門家による調査で、効率的・網羅的・高品質な診断を実施

NRI セキュアは、LLM 向け DAST [6]を採用し、自動でテストが可能な診断用アプリケーションを独自に開発しました。このアプリケーションを用いることで、効率的かつ網羅的に脆弱性を検出することができます。さらに、LLM のセキュリティに精通したエンジニアが診断にあたり、自動テストではカバーできない各システム固有の問題点も洗い出し、検出された脆弱性を調査で深く掘り下げます。

2. システムやサービス全体における実際のリスクを評価し、対策コストを削減

生成 AI は、確率的に出力を決定する性質があります。また、内部の動作を完全に理解することは困難であるという特性上、システムの部分的評価で脆弱性を洗い出すには限界があります。NRI セキュアでは、長年培ってきたセキュリティ診断のノウハウを組み合わせることにより、システムやそれが提供するサービス全体を包括的に診断したうえで、AI に起因する脆弱性が顕在化するかどうかを診断します。本サービスは、AI 固有の問題を評価するだけでは対処が難しい「OWASP Top10 for LLM」[7] にも対応が可能です。

また、仮に AI そのものに脆弱性があった場合、システム全体から見た実際のリスクの程度を評価することによって、実施が難しい AI そのものの脆弱性対応をせずに済むよう、代替の対策案を提示することができます。その結果、対策コストを抑えることが期待できます。

本サービスの詳細については、次の Web サイトをご参照ください。

https://www.nri-secure.co.jp/service/assessment/ai-red-team

NRI セキュアでは、生成 AI を利用したシステム等のセキュリティ対策を継続的に支援していくために、本サービスと対をなすサービスとして、AI アプリケーションの定期的なモニタリングを実施する「AI Blue Team」サービスを開発しています。「AI Blue Team」サービスは 2024 年 4 月の提供開始を予定しており、現在、PoC（Proof of Concept：概念検証）に参加していただける企業を募集しています。

NRI セキュアは今後も、企業・組織の情報セキュリティ対策を支援するさまざまな製品・サービスを提供し、安全・安心な情報システム環境と社会の実現に貢献していきます。

---

[1] 大規模言語モデル（LLM）：LLM は、Large Language Model の略で、大量のテキストデータを利用してトレーニングされた自然言語処理モデルのこと。

[2] プロンプトインジェクション（Prompt injection）：主に、攻撃者が入力プロンプトを操作して、モデルから予期しない、または不適切な情報を取得する試みを指す。

[3] プロンプトリーキング（Prompt leaking）：攻撃者が入力プロンプトを操作して、もともと LLM に設定されていた指令や機密情報を盗み出そうとする試みを指す。

[4] ハルシネーション： AI が事実に基づかない情報を生成する現象のこと。

[5] バイアスリスク：トレーニングデータの偏りやアルゴリズム設計により、偏った判断や予測を引き起こす現象のこと。

[6] DAST：Dynamic Application Security Testing の略で、実行中のアプリケーションをテストし、潜在的なセキュリティ脆弱性を動的に評価する手法。

[7] OWASP Top10 for LLM：世界的なコミュニティである「Open Web Application Security Project（OWASP）」によって作成された、大規模言語モデル（LLM）固有の 10 大セキュリティリスクのこと。

【ニュースリリースに関するお問い合わせ先】

NRI セキュアテクノロジーズ株式会社　広報担当

TEL：03-6706-0622　E-mail：info@nri-secure.co.jp

【ご参考】

NRI セキュアが提供する AI セキュリティ関連サービスについては、以下の Web サイトおよび一覧表をご参照ください。

https://www.nri-secure.co.jp/service/ai-security

「AI セキュリティ統制支援」サービスの一覧

AIサービスラインナップ	サービス概要	支援対象フェーズ	提供状況
AIリスクガバナンス構築支援サービス	AI利活用推進におけるガバナンス対応全般（AI利用ポリシーやガイドライン整備、データガバナンス策定等）に関する、コンサルタントによる人的な態勢整備を支援するサービス	企画構想～開発～導入・運用	提供中
AIセキュリティ診断サービス (AI Red Team)	攻撃者の視点で、AIシステムを攻撃し、潜在リスクを洗い出して改善に寄与するサービス	導入～運用	本日開始
セキュアAI基盤構築支援サービス	企業がAIシステムを安全に利用するにあたって、各種セキュリティ対策を施した専用環境を提供、構築を支援するサービス	開発	2023年度中開始予定
AI品質・適合性検証サービス	欧州AI規制やISO標準規格等、リスクベース規格への適合性検証を行うことでAIシステムの品質を評価するサービス	開発～導入・運用	2023年度中開始予定
データセキュリティサービス	AIシステム内のデータの管理態勢・データ流通を可視化するソリューションの提供によりデータ管理に関するビジネスリスクを軽減するサービス	開発～運用	2023年度中開始予定
AIセキュリティ監視サービス (AI Blue Team)	AIシステムを独自のメソッドでモニタリングし攻撃を検出・防御することで安定運用に寄与するサービス	運用	2024年4月開始予定