2023 年 6 月 9 日

報道関係各位

GMO サイバーセキュリティ by イエラエ株式会社

AI・大規模言語モデルを用いたアプリケーションの セキュリティリスクを評価 「GMO AI セキュリティ診断 for GPT」を提供開始

GMO インターネットグループでサイバー攻撃対策事業を展開する GMO サイバーセキュリティ by イエラエ株式会社（代表取締役 CEO：牧田 誠　以下、GMO サイバーセキュリティ by イエラエ）は、2023 年 6 月 9 日（金）より、大規模言語モデル（LLM）を用いたアプリケーションのセキュリティリスクを可視化する新サービス「GMO AI セキュリティ診断 for GPT」の提供を開始しました。

「GMO AI セキュリティ診断 for GPT」は、GPT 等の大規模言語モデル（LLM）を組み込んだアプリケーションに対し、敵対的プロンプト（Adversarial Prompting）を用いた擬似攻撃を行い、アプリケーションに内在するセキュリティリスクを調査するサービスです。

GMO サイバーセキュリティ by イエラエの、膨大な脆弱性診断の実績・経験に基づく高い技術力と、GPT 等の大規模言語モデル（LLM）への最新の攻撃手法の知見を組み合わせることで、セキュリティリスクを可視化し、お客様のシステムのセキュリティ対策を支援します。

【提供開始の背景】

生成 AI の研究や活用が進む中、OpenAI の「GPT」、Google の「PaLM」、Meta の「LLaMA」といった、様々な大規模言語モデル（LLM）の登場により、言語処理タスクの高精度な実行が可能となりました。また 2023 年 3 月 2 日（木）に ChatGPT の API が一般公開されたことで、製品やサービスに GPT を利用する対話型のアプリケーションが急速に普及しています。GPT ベースのサービスが急速に普及する一方で、対話を通じてサービスの設定情報や秘密情報を抜き出そうとする攻撃（プロンプトインジェクション）が行われるなど、新たなセキュリティリスクが発生しています。

新たに提供を開始する「GMO AI セキュリティ診断 for GPT」は、GPT 等の LLM を組み込んだサービスに対するセキュリティリスクに対応しています。機密情報や金融資産を扱う Web アプリケーションに対す

る脆弱性診断の実績と経験に基づく高い技術力に加え、GPT 等の大規模言語モデル（LLM）に関する最新の攻撃手法の知見を組み合わせ、みなさまのセキュリティ対策を支援します。

【「GMO AI セキュリティ診断 for GPT」の特徴】
（URL：https://gmo-cybersecurity.com/service/ai-security-for-gpt）

■LLM を利用したアプリケーションに対するセキュリティリスクについて

GPT 等の大規模言語モデル（LLM）を利用して構築したアプリケーションは、プロンプトを攻撃の入口として、敵対的プロンプト（Adversarial Prompting）を利用した攻撃が行われる可能性があります。

敵対的プロンプト（Adversarial Prompting）とは大規模言語モデル（LLM）の出力を意図しないものへ誘導するよう設計された一連の入力のことです。大規模言語モデル（LLM）を利用してアプリケーションを構築する際には、一般的な Web アプリケーションのリスクに加え、こうした大規模言語モデル（LLM）を利用して構築したアプリケーション特有のリスクを認識し、セキュリティ対策を講じることが重要です。

敵対的プロンプトに関するセキュリティリスクの例として、以下のようなものが存在します。

（1）プロンプトインジェクションによる出力の乗っ取り
（2）プロンプトリークによる秘密情報の漏洩
（3）コンテンツポリシーの回避による非倫理的なコンテンツの出力（ジェイルブレイク）

「GMO AI セキュリティ診断 for GPT」は、上記のような大規模言語モデル（LLM）を利用して構築したアプリケーション特有のセキュリティリスクに対応可能な診断パッケージです。また予算や求めるセキュリティレベルに応じて 3 つの診断プランから選択が可能です。

■プラン説明

・ライトプラン：安価に敵対的プロンプトのセキュリティリスクを可視化したいお客様向けのプラン
・ベーシックプラン：はじめて GPT 等の大規模言語モデル（LLM）を用いたアプリケーションを構築するお客様向けのプラン
・アドバンスドプラン:関連機能のセキュリティリスクを総合的に評価したいお客様向けのプラン

	ライトプラン	ベーシックプラン	アドバンスドプラン
敵対的プロンプトによる機密情報漏洩の診断	○	○	○
敵対的プロンプトによる出力の汚染	○	○	○
ホワイトボックスによる設定診断	×	○	○
Indirect Prompt Injection 等の最新の攻撃事例を踏まえた手動診断	×	×	○
調査対象アプリケーション毎にカスタマイズした診断	×	×	○

■価格について

プランの詳細、価格については、サービスページよりお問い合わせください。

（サービスページ：https://gmo-cybersecurity.com/service/ai-security-for-gpt）

【GMO サイバーセキュリティ by イエラエについて】
（URL：https://gmo-cybersecurity.com/）

「誰もが犠牲にならない社会」をミッションに掲げ、国内最大規模のホワイトハッカーを組織するサイバーセキュリティのプロフェッショナルカンパニーです。Web アプリケーションやスマホアプリ、企業の基幹システムなどに対するサイバー攻撃に対する高度なセキュリティ対策を提供し、持続可能な事業継続をサポートしています。

以上

【報道関係お問い合わせ先】

●GMO サイバーセキュリティ by イエラエ株式会社
営業部　帰山
TEL：03-6276-6045
E-mail：info@gmo-cybersecurity.com

●GMO インターネットグループ株式会社
グループコミュニケーション部広報担当 青柳
TEL：03-5456-2695　E-mail：pr@gmo.jp

【サービスに関するお問い合わせ先】

●GMO サイバーセキュリティ by イエラエ株式会社
営業部
TEL：03-6276-6045
E-mail：info@gmo-cybersecurity.com

【GMOサイバーセキュリティ byイエラエ株式会社】（URL:https://gmo-cybersecurity.com/）

会社名	GMOサイバーセキュリティ byイエラエ株式会社
所在地	東京都渋谷区桜丘町 26 番 1 号　セルリアンタワー
代表者	代表取締役CEO　牧田　誠
事業内容	■Web アプリ及びスマホアプリ脆弱性診断 ■ペネトレーションテスト ■不正利用(チート)診断 ■IoT 脆弱性診断 ■自動車脆弱性診断 ■フォレンジック調査 ■CSIRT 支援 ■クラウドセキュリティ診断 ■クラウドセキュリティ・アドバイザリー
資本金	1 億円

【GMOインターネットグループ株式会社】（URL：https://www.gmo.jp/）

会社名	GMOインターネットグループ株式会社（東証プライム市場　証券コード：9449）
所在地	東京都渋谷区桜丘町 26 番 1 号　セルリアンタワー
代表者	代表取締役グループ代表　熊谷 正寿
事業内容	■インターネットインフラ事業 ■インターネット広告・メディア事業 ■インターネット金融事業 ■暗号資産事業
資本金	50 億円