報道関係者各位

ウィズセキュア、ChatGPT のサイバー攻撃への悪用の可能性をリサーチ

～ 受け取るメッセージ／コンテンツへのより慎重な対応が必要に ～

2022 年 1 月 19 日
ウィズセキュア株式会社

人間の発音に近いテキストを数秒で提供する言語モデルの利用は、人類の歴史における転換点をもたらすとされています。先進的サイバーセキュリティテクノロジーのプロバイダーである WithSecure（本社: フィンランド・ヘルシンキ、CEO: Juhani Hintikka、日本法人: 東京都港区、以下、ウィズセキュア）は同社が EU（欧州連合）の『Horizon 2020』研究・イノベーションプログラムによるプロジェクトの支援を受けて実施した、機械学習によってテキストを生成する高精度の言語 AI である GPT-3（Generative Pre-trained Transformer 3）を用いた新しいリサーチにおいて明らかになった事項について発表し、ChatGPT のサイバー攻撃への悪用の可能性について警鐘を鳴らしています。

GPT-3 や GPT-3.5 のような自己回帰言語モデルを採用した使いやすいツールが広くリリースされ、インターネットに接続できる環境であれば、誰でも人間のような音声を数秒で生成することができるようになりました。わずかなインプットから多用途の自然言語テキストを生成することは、サイバー犯罪者の興味を引くことは必至でしょう。同様に、ウェブを使って詐欺やフェイクニュース、誤報を流す人たちも、信頼が置けるあるいは説得力のあるテキストを瞬時に作成するツールに関心を持つかもしれません。

サイバーセキュリティの観点からは、大規模な言語モデル、言語モデルが生成できるコンテンツ、そしてそのコンテンツを生成するために必要なプロンプトのリサーチは非常に重要であると言えます。まず、こうしたリサーチにより、現在のツールで何が可能または不可能なのかを知ることができ、私たちは、そして社会全体がそうしたテクノロジーの潜在的な悪用に注意を払う必要があります。次に、モデルのアウトプットを使用して、悪意のあるコンテンツ（有害なスピーチやオンラインハラスメントなど）のデータセットを生成し、その後、そのようなコンテンツを検知する方法を作成し、その検知メカニズムが有効であるかどうかを判断するために使用することができます。また、本リサーチで得られた知見は、今後、より安全な大規模言語モデルの作成に役立てることができます。

ウィズセキュアが主導した本リサーチで研究されたユースケースは、以下のカテゴリーに分類されます。

●フィッシングコンテンツ: ユーザーを騙して悪意のある添付ファイルを開かせたり、悪意のあるリンクにアクセスさせるように設計されたメールやメッセージ
●ソーシャルノミネーション: 個人への嫌がらせやブランド毀損を目的としたソーシャルメディア上のメッセージ
●社会的検証: 広告や販売を目的としたソーシャルメディア上のメッセージ、または詐欺を正当化するために作成されたメッセージ
●スタイルトランスファー: モデルを騙して、特定のライティングスタイルを使用させるように設計された手法
●オピニオントランスファー: モデルを騙して、あるテーマについて意図的に意見を述べるように仕向ける手法
●プロンプト作成: コンテンツに基づいてプロンプトを作成するようにモデルに依頼する方法
●フェイクニュース: GPT-3 がフェイクニュースの記事を生成できるかを実験

本リサーチで行った実験により、大規模言語モデルを用いれば、たとえ学習データに関連情報が含まれていなくても、スピアフィッシング攻撃に適したメールスレッドの作成、特定の人物の文体の模倣である「テキストディープフェイク」、文章内容への意見の適用、特定の文体での執筆、説得力のあるフェイク記事の作成が可能であることが証明されました。こうしたことから、私たちはこのようなモデルがサイバー犯罪や攻撃に用いられるテクノロジーの向上サポートとなる可能性があると結論づけました。

リサーチ結果について、ウィズセキュアのインテリジェンスリサーチャーである Andy Patel (アンディ・パテル) は、次のように述べています。
「インターネットに接続できる人なら誰でも強力な大規模言語モデルにアクセスできるようになったという事実は、非常に深刻な問題をもたらします。つまり、私たちが受け取るメッセージやコンテンツは、AI によって書かれたものかもしれないと、まず疑念を抱くことが必要だということです。今後、AI を使用して有害なコンテンツと有用なコンテンツの両方を生成するには、書かれたコンテンツの意味と目的を理解することができる検出の戦略が必要になるのです。ChatGPT が GPT-3 テクノロジーを誰でも使えるようにする以前に、当社ではこのリサーチを始めていました。この開発によって、私たちのリサーチ結果はより大きな意味を持つようになったと考えています。なぜなら、私たちは映画『ブレードランナー』のデッカードのような捜査官であり、私たちが 相手にしている知性 が『本物』なのか『人工のもの』なのかを見極めようとしているからです。」

(ウィズセキュアのインテリジェンスリサーチャーである Andy Patel)

本リサーチの詳細 (英語) およびレポートのダウンロードについては以下のページをご覧ください。
https://labs.withsecure.com/publications/creatively-malicious-prompt-engineering

WithSecure Web サイト:
https://www.withsecure.com/jp-ja/
WithSecure プレスページ:
https://www.withsecure.com/jp-ja/whats-new/pressroom

WithSecure について

WithSecure™は、IT サービスプロバイダー、MSSP、ユーザー企業、大手金融機関、メーカー、通信テクノロジープロバイダー数千社から、業務を保護し成果を出すサイバーセキュリティパートナーとして大きな信頼を勝ち取っています。私たちは AI を活用した保護機能によりエンドポイントやクラウドコラボレーションを保護し、インテリジェントな検知と対応によりプロアクティブに脅威を探し出し、当社のセキュリティエキスパートが現実世界のサイバー攻撃に立ち向かっています。当社のコンサルタントは、テクノロジーに挑戦する企業とパートナーシップを結び、経験と実績に基づくセキュリティアドバイスを通じてレジリエンスを構築します。当社は 30 年以上に渡ってビジネス目標を達成するためのテクノロジーを構築してきた経験を活かし、柔軟な商業モデルを通じてパートナーとともに成長するポートフォリオを構築しています。
1988 年に設立された WithSecure は本社をフィンランド・ヘルシンキに、日本法人であるウィズセキュア株式会社を東京都港区に置いています。また、NASDAQ ヘルシンキに上場しています。詳細は www.withsecure.com をご覧ください。また、Twitter @WithSecure_JP でも情報の配信をおこなっています。