



背景

生成AIを悪用したフィッシング攻撃の拡大により、ひと月あたりの被害件数が過去最多を更新するなど、被害は増大する一方です。この状況に対し、生成AIを用いてフィッシングサイトを検出することで被害削減に取り組んでいます。

成果の概要

NTTでは人間の心理的な弱みを狙うWeb上のソーシャルエンジニアリング攻撃の収集・検出技術を創出してきました。今回、Webサイトの自動巡回技術と連携することで、生成AIを用いた高精度なフィッシングサイト検出を実現しました。

開発課題



① Webサイトの情報はそのまま生成AIに入力できない

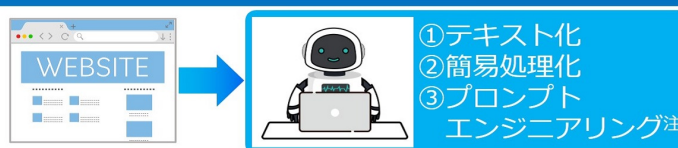


これは
フィッシングサイト？

わかりません

② 単純な命令では判定ミスが多発

技術概要



- ① Webサイトのスクリーンショット画像をテキスト化
- ② HTMLやテキストを簡略化
- ③ プロンプトエンジニアリング^注により推論のプロセスを制御
注) LLMに対しての指示や命令を最適化する手法



これは
フィッシング
サイトです

98%以上の精度^注で
フィッシングサイトを検出

注) GPT-4の利用において

技術のポイント

- 生成AIによる高度な文脈理解を活用して、フィッシングサイトの心理的な働き掛けを特定する世界初の検出技術
- フィッシングサイトに偽装されたブランドを特定し、正規のWebサイトとの不一致を発見することで高精度に悪性判定する能力
- Webサイトの自動巡回により大量のフィッシングサイトを短時間で収集・分析することで、人手による監視や検出作業のコストを削減

この研究がもたらす未来

生成AIを悪用したフィッシング攻撃の大量発生 of 将来を見据え、防御側として生成AIを用いた高精度なフィッシングサイト検出技術で被害を未然に防ぎます。

出展企業

日本電信電話株式会社、NTTセキュリティホールディングス株式会社

問い合わせ先

rdforum-exhibition@ml.ntt.com