URL　https://journal.ntt.co.jp/article/30159
DOI　https://doi.org/10.60249/24115107

主 役 登 場

生成AIが切り拓く フィッシング攻撃対策の未来

小出 駿　Takashi Koide

NTTセキュリティ・ジャパン
研究主任

大規模言語モデル（LLM：Large Language Models）に代表される生成AI（人工知能）の登場は，ビジネス，教育，医療，芸術や娯楽など多くの分野に革新をもたらしています．しかし，その一方で生成AIがサイバー攻撃に悪用されるという新たな脅威も生まれています．有名人や政治家の精巧な偽動画を用いた印象操作や詐欺への誘導，災害や事件に関する偽画像の拡散など，サイバー攻撃の手口は日々巧妙化しています．日本国内でも，生成AIを用いてマルウェアを作成した事例で逮捕者が出るなど，これらの脅威が現実のものとなっています．

このような状況下で，私たちは生成AIをサイバーセキュリティ対策に活用する研究開発に取り組んでいます．その中でも特に注力しているのが，フィッシング攻撃対策です．フィッシング攻撃は，メールやショートメッセージサービス（SMS），偽のWeb広告など，多様な経路を通じてユーザを偽サイトに誘導します．そこでは巧妙な文言や偽の警告により，クレジットカード番号や個人情報，ログイン情報などを入力させようとします．情報を盗まれると，アカウント乗っ取りや金銭的被害など，深刻な事態に発展する可能性があります．

従来のフィッシング攻撃対策では，機械学習技術を用いた検出モデルの頻繁な更新，人手による確認や情報収集が必要でした．これには多大なコストと時間がかかり，日々進化し変化し続ける攻撃に追いつくには大きな困難が伴います．そこで私たちは，LLMを活用した新しいフィッシングサイト検出技術の開発に着手しました．

この技術の特徴は，疑わしいWebサイトに自動でアクセスし，そこから得られる情報をLLMが解釈しやすいかたちに変換する点です．具体的には，プロンプトエンジニアリングという手法を用いて，LLMにフィッシングサイトを判別するための具体的な指示を与えます．さらにWebサイトのスクリーンショット画像やHTMLソースコード，URLを入力することで，これらの情報を基にWebサイトの正当性を判断するのです．検証の結果，驚くべきことに99%以上という高い検出精度でフィッシングサイトを判別できることが分かりました．なぜこれほどの精度が出せたのでしょうか．

LLMはインターネット上の膨大な情報を学習しています．そのため，フィッシングサイトが狙う企業やサービスのブランドに関する多種多様な知識を持っています．この知識を基に，正規サイトとのURLの違いや，不自然な文章表現などを根拠として，ブランドの偽装を見抜くことができます．また，偽のウイルス感染警告，架空の宅配通知，アカウント異常など，人を騙すための典型的な手口を文脈から正確に特定し，それらが偽情報であることを識別できるのです．

私たちの研究は，世界でも先駆的なものでした．LLMが登場した当初，これを悪性サイトやマルウェアの検出に使えるのではないかというアイデアは議論されていました．しかし，実際にその有効性を検証した研究は存在しませんでした．私たちはOpenAI社のGPT-4が発表されてからわずか3カ月という短期間で，システムの構築と検証を行い，論文を公開しました．これにより，LLMを用いたフィッシングサイト検証の有効性を世界で初めて明らかにしたのです．その後も研究開発を重ね，マルチモーダルモデルを活用した画像入力による検出精度の向上など，さらなる改善に取り組んでいます．

しかし，サイバーセキュリティの世界では，攻撃と防御の技術が日々進化を続けています．生成AIの発展により，より巧妙な詐欺手法が生み出されつつあります．そのため，私たちは常に攻撃者の一歩先を行く必要があります．今後の目標として，より迅速にフィッシングサイトの情報を配信し，ユーザによる不用意なアクセスを未然に防ぐシステムの構築をめざしています．また，リアルタイムで悪性サイトを判定し，ユーザに危険度を即座に通知できる仕組みの研究にも取り組んでいます．技術の進歩は，私たちの生活を豊かにする一方で，新たな脅威も生み出します．しかし，同じ技術を応用することで，これらの脅威に対する新たな防御策を構築できるのです．社会の安心と安全を守るため，これからも最先端の研究開発に挑戦し続けます．