G-31

# 生成 AI によるインシデントレスポンスの有用性について

奥野 拓真†　　川橋 裕‡
Takuma Okuno　Yutaka Kawahashi

## 1. はじめに

サイバー攻撃の脅威は年々増大しており，企業や組織におけるセキュリティ対策の重要性が高まっている．特に，迅速なインシデントレスポンスが求められる状況において，適切な対応策を即座に判断する能力が不可欠である．しかし，インシデント発生時の対応には高度な専門知識が必要であり，現場のエンジニアが常に的確な判断を下すことは容易ではない．加えて，サイバーセキュリティ分野では専門家の人材不足が指摘されており，十分な知識を持つ担当者が確保できない場合もある．実際に，「企業における情報セキュリティ実態調査[1]」によれば，企業規模を問わずセキュリティ人材が不足していると回答した企業は 90%を超えており，この事実は日本企業における共通の課題である．さらに，サイバー攻撃の高度化や多様化が進む中，インシデントの数は増加傾向にあり，限られた人材だけではすべてのインシデントに迅速かつ適切に対応することが難しい状況にある．

一方でセキュリティインシデント対応のプロセスは，主に専門的な知識と経験を持つセキュリティエンジニアによって運用されることを前提として設計されている．そのため，非専門家が対応を試みた場合，インシデントの診断や適切な対応の判断が難しく，対応の遅れや誤った処置のリスクが高まる．ここで非専門家とは，セキュリティに関する専門教育や訓練を受けておらず，基本的なセキュリティ概念を十分に理解していない一般的なエンドユーザーや IT 担当者を指す．このような課題に対処するため，専門家の知識を補助し，適切な対応を支援する技術の導入が求められている．

近年，生成 AI（Generative AI）がさまざまな分野で利用されており，インシデントレスポンスの分野でも実際に活用されるケースが見られる．生成 AI は，膨大なデータを学習し，ユーザの入力に応じて適切な情報を生成する機能を持つ．適切な質問を行えば，インシデントレスポンスの方針を導き出せる可能性がある．さらに，複雑な技術情報を平易な言葉に翻訳することで，非専門家でも直感的に理解しやすい形で情報を提供し，インシデント対応における意思決定を支援する役割を果たす可能性がある．

一方で，生成 AI が出力する情報は必ずしも正確とは限らず，誤った指示が対応の遅れや問題の悪化を招く恐れがある．攻撃の種類によっては適切な判断が下せない場合も考えられるため，生成 AI がどの程度インシデントレスポンスにおいて実用性を持ち得るのかを検証することが重要である．

以上の背景を踏まえ，本研究では，セキュリティ知識をほぼ持たない非専門家でも，大規模言語モデルを活用した既存の対話型サービスを通じて，サイバーセキュリティのインシデントレスポンスを効率的かつ効果的に実施できる可能性を検証し，有用性を評価した．具体的には，対話型生成 AI である ChatGPT[2]を用い，実際のインシデントを想定した複数のシナリオにおいて生成 AI の出力を分析し，適用可能性の評価を行った．

和歌山大学大学院 システム工学研究科

## 2. 関連研究・既存技術

### 2.1　関連研究

Jie Zhang らによる「When LLMs Meet Cybersecurity[3]」では，大規模言語モデル（LLM）のサイバーセキュリティ分野への応用について，300 以上の研究を対象とした体系的文献レビューが行われている．同研究では，脆弱性検出やコード修復，脅威インテリジェンスの分析など多様な応用シナリオが示され，LLM の技術的可能性と課題が広く論じられているが，非専門家の利用を想定した具体的な運用例や評価は行われていない．

Obaloluwa Ogundairo らによる「Natural Language Processing for Cybersecurity Incident Analysis[4]」では，自然言語処理（NLP）の技術を用いて，非構造化データからの情報抽出や報告支援，脅威インテリジェンスの分析を行う可能性について議論されている．ただし，こちらもセキュリティ専門家を主な対象としており，非専門家による実践的なインシデント対応支援には言及されていない．本研究は，これらの課題を補完する形で，非専門家が LLM を活用して実際にインシデントレスポンスを行う際の有用性と課題について検証を行った点に特徴がある．

### 2.2　既存技術

既存技術としては，Microsoft が提供する Security Copilot[5]と，データ分析基盤である Splunk[6]が挙げられる．Security Copilot は，生成 AI を活用してセキュリティ専門家のインシデント対応や脅威ハンティングを支援するツールであり，高度な自然言語応答によって迅速な判断を可能とする．しかし，Microsoft 製品との統合を前提としており，他社ツールやオンプレミス環境での柔軟な導入には制限がある．一方，Splunk はログデータやイベントデータを収集・解析し，自動化されたインシデント対応を可能にするプラットフォームであり，異常検知やアラートの優先順位付けなどを通じて SOC 業務の効率化を支援する．ただし，操作には高度な専門知識が求められ，非専門家に対する具体的な支援は想定されていない．本研究は，これらの技術が持つ制約を補完し，非専門家が大規模言語モデルを活用してインシデントレスポンスを実行可能とするアプローチの有用性を検証するものである．

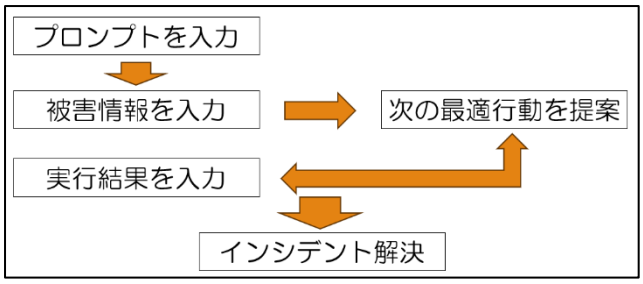
## 3. 提案手法

本研究では，LLM を活用した対話型サービスを用いて，セキュリティ知識を持たない非専門家でもインシデント対応を実施できる可能性を検証する．想定される対応の流れとしては，まずユーザが LLM に対して「特定の Web ページが不正にリダイレクトされている」といった問題の概要を入力し，LLM はそれに基づき原因を推定し，ログ確認などの追加情報を促す．ユーザが得られたログやエラー情報を LLM に入力すると，LLM はそれらを解析し，問題の特定と具体的な修復手順を提示する．たとえば index.html の修正や.htaccess の設定変更などが該当し，必要に応じてコマンドライン例も提示される．また，対応完了後にはパスワード変更や多要素認証といった再発防止策も提案される．このような手法により，非専門家であっても段階的に対応手順を把握しやすくなることが期待される．さらに，LLM を活用した対話型サービスには，リアルタイムでのログ解析と情報抽出による迅速な意思決定支援，複雑な技術内容の可視化による理解促進，そして対話形式による指示の受け取りやすさといった利点がある．本手法では，これらの特性を活かし，非専門家の実践的なインシデントレスポンスを支援することを目的としている．

## 4. 実験

実験では，ホスティングサービスを運営する企業を想定し，ネットワークセキュリティの知識が乏しい非専門家が，ChatGPT の指示に従ってインシデントに対応するというシナリオを設定した．実験環境としては，VMware 上に構築した FreeBSD 14.1 および Apache 2.4 を用いた仮想サーバを使用した．発生させたインシデントは 2 種類であり，1 つ目はユーザの Web ページが不正ログインにより改ざんされ，詐欺サイトへリダイレクトされるというものであり，2 つ目は SlowHTTPDoS 攻撃によりサーバのリソースが枯渇し，Web ページにアクセスできなくなるというものである．

ChatGPT には，非専門家としての立場や組織的背景，指示形式，制約条件を含む詳細なプロンプトを入力し，その指示に従って対応を進める形式とした．実験手順としては，通報内容を ChatGPT に入力し，指示に従って調査や修復を実施し，結果を再入力するというやりとりを繰り返しながらインシデント解決を目指した．




基本的に AI の指示に従ってのみ行動し，自発的な判断は行わないという条件の下，進捗が止まった場合にはヒントを段階的に与えることで AI の対応能力を検証した．評価は，チェック項目の達成状況や指示の正確性，誤情報の有無，行き詰まりの要因を記録し，各シナリオにつき 5 回ずつ繰り返して傾向を分析した．

## 5. 実験結果・考察

### 5.1 実験結果

インシデント 1 に関する実験では，5 回中 4 回において ChatGPT は適切な手順を提示し，ヒントなしでインシデントの解決に至った．この 4 回では細かなコマンドの違いはあったものの，index.html の修正，.htaccess の削除，ログ解析による攻撃者の特定，パスワード変更の提案といった一連の対応が適切に行われた．一方，残る 1 回では，ChatGPT が.htaccess の存在を誤って否定したため，関係のない修正が繰り返され，解決に至らなかったが，「.htaccess が他のディレクトリに存在する可能性がある」とのヒントを与えることで，その存在を特定し最終的に解決された．インシデント 2 に関する実験では，5 回すべてにおいて基本的に同様の手順で進行し，いずれのケースにおいてもヒント無しでは解決まで至ることはなかった。こちらから SlowHTTPDoS 攻撃であることを明示した後は，適切な対策が提示されたことから，解決できなかった主因は,攻撃手法を特定できていなかった点にあると考えられる．

### 5.2 考察

本研究の結果から，ChatGPT は一定の精度で問題解決が可能であるが，誤った仮説に基づいた場合，適切な解決に至るまでに試行錯誤を要する可能性が示された．特に，ファイルの特定や調査に関する指示が不完全な場合，ChatGPT がインシデントに関係のない修正を繰り返してしまうことがあり，その際には適切なヒントを提供することで軌道修正が可能となることが分かった．また，攻撃手法を明確に特定できないインシデントにおいては，複数の試行を経てようやく適切な対応策が提示される傾向が見られ，このようなケースでもヒントの有無が解決の効率に大きく影響することが示唆された．インシデント 1 で与えた「.htaccess が public_html 以外のディレクトリに存在する可能性がある」というヒントは，比較的気づきやすく，非専門家でも ChatGPT の指示を通じて意識する機会があると考えられるのに対し，インシデント 2 における「攻撃手法が SlowHTTPDoS 攻撃である」というヒントは，セキュリティの専門知識がほとんどない非専門家では思いつくことが難しい．特に，パケットキャプチャの結果からリクエストの送信間隔やセッションの持続時間を解析するという知識がなければ，その挙動を単なる「大量のリクエスト」としてしか認識できない可能性が高い．したがって，ChatGPT のインシデントレスポンス支援能力を向上させるためには，攻撃手法をより正確に識別し，状況の変化に適応した対応が取れるようにする質問やガイドラインの提示が重要であることが明らかとなった．

## 6. 今後の課題

今後は，より多様な攻撃パターンに対する検証や，プロンプトの構造に関する設計手法の確立，さらには AI 活用における能動的な情報提示の訓練手法などを含め，非専門家が実践的に活用できる支援環境の構築が求められる．また，インシデント対応の現場で生成 AI を実用的に運用するには，機密情報を外部に送信しない環境が求められる場面も多く，オープンソースの LLM を用いたローカル環境での実行可

能性や，社内データと連携した LLM 活用の方法も今後の重要な検討課題である．

## 参考文献


[1] 企業における情報セキュリティ実態調査～NRI Secure Insight 2023～
https://www.nri-secure.co.jp/hubfs/NRIS/download/ebook/NRISecure_Insight2023_h8ut.pdf
(参照 2025-02-10)

[2] ChatGPT | OpenAI
https://openai.com/ja-JP/chatgpt/overview/
(参照 2025-02-10)

[3] Zhang, J., Bu, H., Wen, H., Chen, Y., Li, L., Zhu, H.
”When LLMs Meet Cybersecurity: A Systematic Literature Review”
Cybersecurity volume 8, Article number: 55 (2025)

[4] Ogundairo, Obaloluwa. Broklyn, Peter.
”Natural Language Processing for Cybersecurity Incident Analysis”
Journal of Cyber Security (2024)

[5] Microsoft Security Copilot | Microsoft Security
https://www.microsoft.com/ja-jp/security/business/ai-machine-learning/microsoft-security-copilot
(参照 2024-11-19)

[6] Splunk (スプランク) | 企業のレジリエンス強化の鍵となる
https://www.splunk.com/
(参照 2025-02-08)