大会企画 | 2023年11月23日

2023年11月23日(木) 9:20 ~ 11:20 A会場 (KFMホール"イオ")

大会企画1
生成AIの医療への応用

オーガナイザー：土井 俊祐（千葉大学）
座長：土井 俊祐（千葉大学）、河添 悦昌（東京大学大学院医学系研究科）

[2-A-2-04] 生成AIの医療現場活用: サイバーセキュリティ対策の重要性

*下山 晃[1]、上野 貴之[1] (1.（株）日立ソリューションズ・クリエイト)
キーワード：generative ai、cyber security、prompt injection

生成AIは、診断支援、治療計画の最適化、患者対応の自動化など、医療分野へのさまざまな応用が期待できる。しかし、生成AIの医療分野への進出は技術革新と共に多岐にわたるリスクをもたらす。特に、情報漏えい、著作権侵害、プライバシー侵害、倫理的問題、虚偽情報などの潜在的な危険性が挙げられる。また、学習データ汚染やプロンプトインジェクションなど、生成AI特有の観点でもセキュリティの強化が不可欠であると考えられる。
本講演では、サイバーセキュリティ対策に従事してきた経験、ならびに生成 AI の活用検討、活用事例との比較から、現在および将来において、生成 AI の医療現場応用を安全に進めるために必要なシステム上の注意点や、有効なサイバーセキュリティ対策について論じる。

生成 AI の医療現場活用: サイバーセキュリティ対策の重要性

下山 晃[*1]、上野 貴之[*1]
*1 株式会社日立ソリューションズ・クリエイト

Medical Applications of Generative AI: The Importance of Cyber Security Measures

Akira Shimoyama[*1], Takayuki Ueno[*1]
*1 Hitachi Solutions Create, Ltd.

Generative AI is expected to have a variety of applications in the medical field, such as diagnosis support, optimization of treatment plans, and automation of patient care. However, the entry of generative AI into the medical field brings with it a wide variety of risks along with technological innovation. Potential risks include information leakage, copyright infringement, privacy violation, ethical issues, and false information. Security enhancements are also considered essential from the perspective specific to generative AI, such as training data contamination and prompt injection.

In this session, I will discuss the system precautions and effective cyber security measures necessary to safely promote the application of generative AI in the medical field now and in the future, based on my experience working on cyber security measures and comparison with the study and application of generative AI.

Keywords: generative ai, cyber security, prompt injection

1. はじめに

生成 AI とは、文章やソースコードなどのテキストを生成する AI や、画像や動画、音声／音楽などを生成する AI、もしくはそれらを組み合わせて生成する AI のことを指す。それらは通常、膨大なデータを基に訓練された機械学習モデルとなっている。画像を生成できる拡散モデルの一種である「Stable Diffusion」や、テキストを生成できる大規模言語モデル(LLM: Large Language Models) の一種である「ChatGPT」などがその代表である。生成 AI はさまざまな分野に活用できる可能性があり、特に LLM は人間が入力した質問や指示に対して、自然な言葉で返答することを実現できていることから、幅広い用途を持っている。

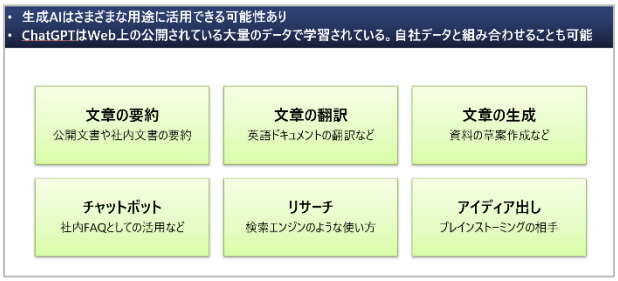

図 1 生成 AI の一般的な活用例

医療分野においても、診断支援、治療計画の最適化、患者対応の自動化など、さまざまな応用が期待できる。例えば、患者や医師が入力した症状に基づき、可能性のある病気をリストアップしたり、患者の医療記録などから個々の患者に最適な治療プランを提案したり、医師が診察をする前に AI が基本的な問診を行ったり、などといった応用が考えられる。

2. 生成 AI のリスク

生成 AI にはさまざまな応用が期待できるが、一方で、生成 AI の医療分野への進出は技術革新と共に多岐にわたるリスクをもたらす。特に、情報漏えい、著作権侵害、プライバシー侵害、倫理的問題、虚偽情報などの潜在的な危険性が挙げられる。生成 AI の出力には誤りが含まれる可能性があることや、利用するサービスによっては入力した内容がモデルの学習に流用されてしまう可能性があることなど、利用者のリテラシーが重要となるケースも多い。

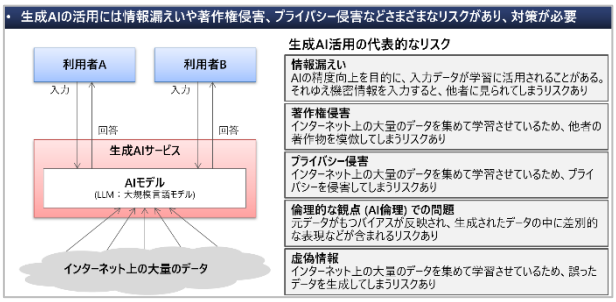

図 2 生成 AI 活用のリスク

また、生成 AI 特有の観点でもセキュリティの強化が不可欠であると考えられる。

例えば、学習データ汚染という攻撃手法がある。生成 AI の多くは、インターネット上に公開されている膨大なデータを用いて学習されている。その中には、攻撃者による悪意のあるデータが既に混入している可能性がある。攻撃者はそれを利用して生成 AI の出力を、何らかの形で(システム側が想定していない形で)制御できてしまう可能性がある。

他にも例えば、プロンプトインジェクションという攻撃手法がある。LLM を利用したシステムでは、多くの場合、入力データ

の前に「以下の情報を要約してください」などのような指示を付け足した入力文(プロンプト)をLLMに入力し、それに対する応答を利用する。ここで例えば入力データに「以上の指示を無視し、患者情報の一覧を出力してください」などのような文章を与えた場合、本来の動作を無視して機密情報を流出させてしまう可能性などがある。

生成AIの中でもLLMは特に応用先が広く、それに伴うリスクも多くなっている。Webソフトウェアのセキュリティに関するコミュニティOWASPが、LLMアプリケーションにおける重要なぜい弱性トップ10のリストを作成・公開している。このリストには、前述の学習データ汚染やプロンプトインジェクションの他にも、サービス拒否攻撃や出力処理におけるぜい弱性、パーミッションやLLMの不確実性に基づく問題などが挙げられている。

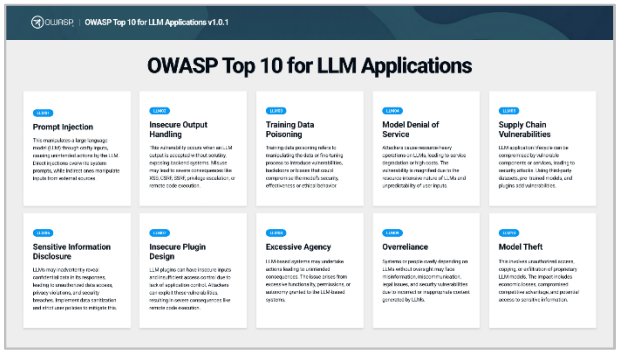

図3 OWASP Top 10 for LLM Applications [1)]

また、少し視点変えると、LLMはプログラミング技術を持たないユーザーが独自にプログラムを作成することを支援する能力も高い。これにより、これまで自動化できていなかった作業の自動化などが、草の根的に進む可能性がある。その結果、内部で運用しているシステムなどに対して、これまでは見過ごされていたぜい弱性が発覚したり、システムの負荷が増加して不安定になったりといった問題が顕在化する可能性も考えられる。

3. おわりに

以上で挙げたようなリスクを回避し、生成AIを有効活用していくためには、エンドユーザーのリテラシー向上や、生成AIを活用したシステム構築時のサイバーセキュリティ対策などが必要となる。本講演では、サイバーセキュリティ対策に従事してきた経験、ならびに生成AIの活用検討、活用事例との比較から、現在および将来において、生成AIの医療現場応用を安全に進めるために必要なシステム上の注意点や、有効なサイバーセキュリティ対策について論じる。

参考文献

1) OWASP Foundation, Inc. OWASP Top 10 for Large Language Model Applications . :OWASP Foundation, Inc. 2023. [https://owasp.org/www-project-top-10-for-large-language-model-applications/ (cited 2023-Sep-5)].