LLM AI サイバーセキュリティとガバナンスのチェックリスト

～　失敗しない大規模言語モデル導入のために　～

OWASP 大規模言語モデル
アプリケーション リスク トップ10 チーム 監修

1.1 版
令和 6 年 4 月 10 日

OWASP PDF v3.8.0a 20240410-142125

本書で提供される情報は、法的助言を提供するものではなく、また提供することを意図するものでもありません。すべての情報は一般的な情報として提供することを目的としています。

本書には、他の第三者のウェブ・サイトへのリンクが含まれていますが、
OWASPは、それら第三者のサイトのコンテンツを推奨または保証するものではありません。

改定履歴

2023-11-01　英語版　0.1　initial draft
2023-12-06　英語版　0.5　public draft
2024-02-15　英語版　0.9　pre-release draft
2024-02-19　英語版　1.0　public release v 1.0
2024-04-10　英語版　1.1　public release v 1.1
令和6年4月10日　日本語版　1.1

目 次

図 表

第一章　概 要

すべてのインターネットユーザーと企業は、来るべき強力な生成的人工知能（生成AI）アプリケーションの波に備えなければなりません。生成AIは様々な業界に革新と効率化、商業的成功をもたらす大きな可能性を秘めている一方、発展の初期段階にある他の強力なテクノロジーと同様、予期せぬ新たな課題ももたらすからです。

人工知能は過去50年の間に著しい発展を遂げ、企業の様々なプロセスをサポートしてきましたが、多くは目立たないものでした。ChatGPTが広く知られるようになると、個人のレベル、並びに企業の間で、大規模言語モデル(Large Language Models, LLM)の開発と利用が推し進められることとなりました。当初、これらテクノロジーの用途は、学術的な研究や特定の企業の使用に限られており、一部の人にしか知られていませんでした。しかし、多くのデータが使えるようになり、コンピュータの能力が向上し、生成AIが進歩し、そしてLlama 2、ElevenLabs、Midjourneyのようなツールがリリースされるにおよび、AIはニッチなものから一般に広く受け入れられるものへと変貌をとげています。生成AI技術がより身近なものになっただけでなく、企業が業務にAIを統合し活用するためには、確たる戦略を策定する必要性が浮き彫りになってきたのです。

最初にいくつかの重要な用語を定義しておきましょう。

- □ 人工知能(AI)は、従来、人間の知能を必要としてきたタスクを、機械が達成できるようにするために必要な、コンピュータサイエンスの多くの分野を包含する用語です。機械学習と生成AIは、人工知能(AI)に含まれる2つの分野です。
- □ 機械学習はデータから学習できるアルゴリズムの作成に重点を置いています。機械学習アルゴリズムは、一連のデータを繰り返し入力することにより学習し、未知のデータの変動予測や分類を行うことができます。
- □ 生成AIは、機械学習アルゴリズムが新たなデータを作り出すことができるようにしたものです。
- □ 大規模言語モデル（LLM）は、人間と同じように文章を理解し、かつ、文章を作り出すことができる生成AIモデルの一種です。
- □ AIの用語で「モデル」とは、入力データに基づいて予測を行うように訓練されたシステムを指します。LLMは特に自然言語の大規模なデータ、すなわち大量の文章から学習されるため、大規模言語モデルと呼ばれています。

我々は今、生成AIソリューションの安全性の確保と管理という未知の領域に入ってきています。生成AIは、敵対者が攻撃戦略を高度化することにも使えるため、その急速な進歩は両刃の剣で、攻撃の脅威と防御の必要性が急速に拡大することも意味するのです。

この「LLM AI サイバーセキュリティとガバナンスのチェックリスト」は、コンテンツを自動生成するLLMを使ったアプリケーションに対する攻撃の脅威と防御に焦点をあて、解説していきます。

責任ある信頼できるAI

AIの利用をめぐる利点と課題が明らかになり、規制や法律が成立するにつれ、『責任ある信頼できるAI』の原則は、漠然とした懸念から、確立された基準へと進化しています。

OWASP AI Exchange ワーキンググループは、このような変化を監視しAIのあらゆる側面における、広範で複雑な要件の検討に取り組んでいます。そのなかで、信頼できるAIの条件を次のように考えています。

図 1.1　信頼できる AI の条件
Montreal Ethics Institute Example より抜粋

対象読者

本書は、経営幹部、技術者、サイバーセキュリティ、プライバシー、コンプライアンス、法務、DevSecOps、MLSecOps、サイバーセキュリティチーム、および保守担当者の各分野のリーダーを対象としています。

AIを企業の成功のために活用するだけでなく、安全でないAIの実装や、AIの実装を急ぐことによるリスクにも注目し、急速に変化するAIの世界で、常に一歩先を行くための努力をしていらっしゃるリーダーやチームの皆さん、チャンスをつかみ、課題と戦い、リスクを軽減するための戦術を構築しようとされている全ての方々に読んでいただけることを狙いとしています。また、技術およびビジネス・リーダー

の方々には、LLMを使用するリスクとメリットを理解し大規模言語モデル戦略を策定する際に、組織を防衛するために重要な領域とタスクの包括的なリストをいち早く作成する助けになることを目的としています。

OWASP 大規模言語モデル アプリケーション リスク トップ10 チームは、このリストが読者の皆様の組織で既にお持ちの防御手法の改善や、この新しいLLMテクノロジーを使用する際の脅威への対応策の開発に役立てることを願っています。

なぜチェックリストが必要なのか

チェックリストは、戦略を策定する際に、正確性を向上させ、目的を明確にし、統一性を保ち、集中した慎重な作業を促し、見逃しや細部の見落としを減らします。チェックリストは、戦略策定への確度を高めるだけでなく、継続的な改善のためのシンプルで効果的な手段を提供し、将来の組織の革新をより確かなものにします。

将来にわたって網羅的ではない

本書は、急速に変化する技術的、法的、及び様々な規制の中で、組織が初期のLLM戦略を策定する際の支援を目的していますが、全ての使用ケースや方策を網羅するものではありません。時には、このチェックリストのもとにして、その使用ケースや管轄区域の必要に応じて、評価及び実務を拡張する必要がでてくるでしょう。

大規模言語モデルの課題

大規模な言語モデルは、いくつかの深刻で固有の問題に直面しています。最も重要なことの1つは、LLMの開発は、大量の訓練データを用いてモデルを作り出すので、「制御の流れ」と「データ」を厳密に区別、または分離することができないということです。通常のコンピュータ・プログラムでは「制御の流れ」の記述しているプログラムを修正することで、ソフトウエアの不具合を修正することができますが、LLMの不具合が特定できたとしても、訓練データのどこを変えると不具合が修正できるかは、多くの場合、一筋縄ではいかないのです。
もう1つの重大な課題は、LLMは設計上、非決定論的であり、プロンプト、リクエストが同じようでも、時々に異なる結果をもたらすことです。すなわち、LLMは意味論的な検索を採用しているので、モデルのアルゴリズムがレスポンスに含まれる語に微妙に異なった優先順位をつけることです。従来のインターネット検索で使われているキーワード検索とは大きく異なり、結果の一貫性と信頼性に影響を与えます。LLMでは「幻覚」と呼ばれる現象があります。それは、モデルのトレーニングに使ったデータの欠落やデータの欠陥が引き起こす不確かさがある場合、生成プロセスが非決定論的な動作であるため、同じような状況でも結果が異なることにより生じます。

信頼性を向上させ、ジェイルブレイク(註1)、モデルトリック(註2)、幻覚に対する攻撃を防御する方法

はありますが、攻撃手法が巧妙になってくると、防御策を実装するコストとその実効性の間でトレードオフをする必要が出てきます。

【註】

1. ジェイルブレイク
   従来のコンピュータでは、制限されている機能にアクセスする方法を見つけ出し、実行することを「ジェイルブレイク」という。LLMアプリケーションでは、ユーザーがプロンプトを巧妙に仕組むことによりLLMアプリに有害な動作をするように仕向けること。LLMが進歩するにつれ、ジェイルブレイクにたいする防御も進歩しているが、攻撃者によるアタック手法も巧妙になっている。
2. モデルトリック
   モデルの訓練に使用されたデータの不完全性をついて、「全く同じように見えるデータだが、モデルは全く異なる結果をだす」場合や「全く異なるように見えるデータだが同じ結果をだす」という欠陥を見つけ出し、誤動作にいたらしめること。

LLMの使用が増加するにつれ、従来のソフトウエアには無かった領域が攻撃対象となり、リスクが増大します。LLMアプリへの攻撃には、そのような未知のものもありまずが、既知の攻撃手法を使った場合が多いのも事実です。たとえば、ソフトウェア部品表（SBoM）、サプライチェーン、データ損失保護（DLP）、認証されたアクセスなど、多くはおなじみの問題です。また、生成AIの進歩は、攻撃者が生成AIを利用すれば、攻撃者の効率性、能力、有効性を高めることになります。

【註】

3. ソフトウェア部品表（SBoM）
   自らのソフトウエア中で使用するオープンソース・ソフトウエアなど第三者が作成したもの
4. サプライチェーン
   自らのソフトウエア中で使用するベンダーが作成したもの
5. データ損失保護（DLP）
   重要データが使用中、転送中、保管中に失われたり、漏えいすることから守ること
6. 認証されたアクセス
   ソフトウエア・ランセンスを受けることにより受領者に与えられるアクセスならびに使用権

攻撃者は、LLMと生成AIを利用し、組織、個人、政府システムを攻撃する手法を、いち早く変えてきています。LLMは、新しいゼロデイ脆弱性（註7）を埋め込んだり、検知を回避するように設計された新しいマルウェア（註8）の作成を、容易にするために使うことができるのです。また、LLMを使って今までにない、より巧妙なフィッシング詐欺（註9）の手口を生み出すことも可能です。動画や音声を偽造するなどして、ソーシャルエンジニアリング（註10）の攻撃をさらに高度な、見分けにくいものにできます。さらに、AIツールは、侵入を実行し、巧妙なハッキング手法を開発することにも使えます。今後、AI技術を利用して、個々の攻撃対象に特化した犯罪、かつ複合的な犯罪が発生することも十分に考えられます。それに対し、適切な防御や攻撃されてしまった場合の回復に対する特別な解決策が求められるようになってきます。

【註】

7. ゼロデイ脆弱性
   開発チームに紛れ込んだハッカーによりソフトウエアのリリース前に埋め込まれた脆弱性
8. マルウェア
   コンピューターやその利用者に被害をもたらすことを目的とした、悪意のあるソフトウェアのこと。ウイルスは、プログラムの一部を書き換え、自己増殖していくマルウェアです。
9. フィッシング詐欺
   送信者を詐称した電子メールを送りつけたり、偽の電子メールから偽のホームページに接続させたりするなどの方法で、クレジットカード番号、アカウント情報（ユーザID、パスワードなど）といった重要な情報を盗み出す行為
   総務省　国民のためのサイバーセキュリティサイトより
10. ソーシャルエンジニアリング
    ネットワークに侵入するために必要となるパスワードなどの重要な情報を、情報通信技術を使用せずに盗み出す方法です。その多くは人間の心理的な隙や行動のミスにつけ込むものです。

組織はまた、LLMを活用しない場合に発生する脅威にも直面しています。例えば、競争上の不利、顧客やパートナーからの時代遅れという市場認識、パーソナライズされたコミュニケーションの拡張不能、イノベーションの停滞、運用上の不備、プロセスにおける人為的ミスのリスクの増大、人的資源の不適切な配分などです。

さまざまな種類の脅威を理解しビジネス戦略と統合すれば、大規模言語モデル（LLM）を使用する場合、使用しない場合のメリットとデメリットを比較検討して、そのデメリットばかりを恐れることによりビジネス目標の達成を妨げるのではなく、むしろ加速させるように、LLMの導入戦略を見極めることができるのです。

AIの脅威の分類

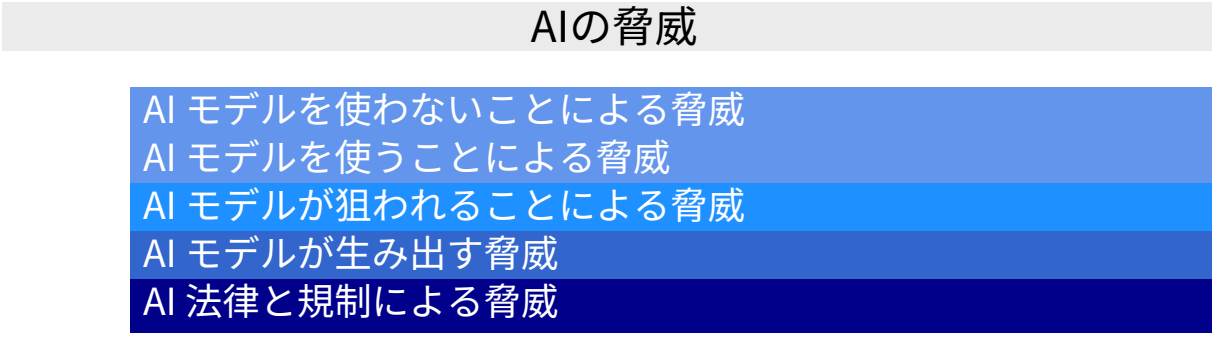

図 1.2　AIの脅威の分類
作成 sdunn

AIのセキュリティとプライバシーのトレーニング

AIと生成AIを理解すること、並びにLLMの構築・購入・利用により将来おこり得る事象を理解するためのトレーニングは、組織全体の従業員にとって有益です。LLMの、許される使い方と許されない使い方、セキュリティ意識に関するトレーニングは、すべての従業員を対象とするものだけでなく、人事部、法務部、開発者、データチーム、セキュリティチームなど、特定の職種に特化したものも必要です。

AIを公正に使用するための基礎と健全なAIとは何かを、早い段階から理解することは、将来のAIサイバーセキュリティ啓発の成功の礎となります。ユーザーはAIに対する基本的な対応の仕方だけでなく、AIが社会にもたらすの功罪、倫理的かどうかを判断する規範を身につけることができるのです。

LLMのセキュリティとガバナンスを既存の施策に組み込む

AIは、サイバーセキュリティ、プライバシー、法律や様々な規制への対応に新たな局面をもたらしますが、問題を特定し、脆弱性を発見し、脆弱性を修正し、潜在的なセキュリティ問題を軽減する最善の方法は以下のように目新しいものではありません。

□ AIシステムの管理方法が、既存の組織慣行と統合されていることを確認する。
□ AIシステムが、既存のプライバシー・ガバナンス・セキュリティ施策に従うことを確認する。
□ 必要に応じて、AIに固有のプライバシー・ガバナンス・セキュリティ施策を策定する。

セキュリティの基本原則

LLMの機能は、異なるタイプの攻撃と攻撃対象領域をもたらします。LLMは、プロンプト・インジェクション、安全でないプラグイン設計、リモート・コード実行などの複雑なビジネス・ロジックのバグに対して脆弱です。そのようなビジネス・ロジックのバグを解決するには、既存の施策が役立ちます。社内の製品セキュリティチームが、セキュアなソフトウェアのレビュー、アーキテクチャ、データガバナンス、サードパーティの評価を理解しているかどうか。サイバーセキュリティチームは、更に、音声クローニング（註11）、なりすまし（註12）、キャプチャ（註13）のバイパスなど、LLMによって影響が大きくなる問題を見つけるために、十分な対応をしているかチェックする必要があります。
機械学習、NLP（自然言語処理）、NLU（自然言語理解）、ディープラーニング、そしてLLM（大規模言語モデル）や生成AIの最近の進歩を考慮すると、サイバーセキュリティ・チームやソフトウエア開発管理チームと一緒にこれらの技術分野に精通した専門家を含めることが重要になってきます。彼らの専門知識は、組織がこれらの技術を採用するのに必要なだけでなく、新たな課題に対する革新的な分析や対応を開発する上でも役立ちます。

【註】
11. 音声クローニング
個人の録音された音声を分析し、その音声の高低、特徴などを使って、会話を合成すること

12. なりすまし（impersonation）
    音声だけでなく、行動や、容姿などをビデオ、画像を合成すること
13. キャプチャ
    オンライン・ログインの際に、パスワード入力の他に、人間であること確認するために作った仕組みで、画像の種類を判断させるなどの手法がある

リスク

ISO31000（註）の定義では、リスクとは「目的に対する不確実性の影響」です。本チェックリストで述べるLLMリスクは、敵対的リスク、安全リスク、 法律リスク、規制リスク、世評リスク、財務リスク、競争リスクに対応しています。

【註】

ISO31000

脆弱性と対応策

脆弱性を分類し脅威情報を共有するために開発された OVAL（註14）、STIX（註15）、CVE（註16）、CWE（註17）などのシステムは、大規模言語モデル（LLM）や予測モデルに固有の脆弱性や脅威を監視し、攻撃があった場合、警告する機能をまだ開発中です。AI/MLシステムとそのサプライチェーンに対する脆弱性や脅威が見つかった際には、脆弱性分析のためのCVEやサイバー脅威インテリジェンス（CTI）の情報交換のためのSTIXのような、確立し広く使われている標準は大きな助けになります。

【註】

14. OVAL
    Open Vulnerability and Assessment Language
15. STIX
    Structured Threat Information eXpression
16. CVE
    Common Vulnerabilities and Exposures
17. CWE
    Common Weakness Enumeration

第二章　LLM導入戦略の策定

大規模言語モデル（LLM）アプリケーションの急速な拡大により、業務で使用するあらゆるAI/MLシステム（生成AIと従来の予測AI/MLの両方）が注目され、評価される機運が高まっています。そのため、これまで見過ごされていたシステムを標的とする攻撃者や、法的、プライバシー、責任、保証の観点から軽視されていたガバナンスや法的課題などの潜在的なリスクが明らかになってきました。AI/MLシステムを業務に活用する組織にとって、包括的なポリシー、ガバナンス、セキュリティ・プロトコル、プライバシー対策、管理責任基準を評価・確立し、これらのテ技術が安全かつ倫理的にビジネス・プロセスと整合していることを確認することが重要です。

攻撃者（敵対者）は、企業、国民、政府機関にとって最も直接的で有害な脅威です。彼らの目的は、金銭的な利益からスパイ活動まで多岐にわたり、重要な情報を盗み、業務を妨害し、信用を傷つけることにあります。さらに、攻撃者は、AIや機械学習などの新技術を活用して手法を高度化し、攻撃のスピードを早めて防御が彼らの攻撃に先んじることを困難にしています。

多くの組織にとって差し迫ったLLMの脅威に、「シャドーAI」があります。「シャドーAI」とは、従業員が未承認のオンラインAIツールや安全でないブラウザのプラグインを使用したり、標準的なソフトウェアの承認プロセスを回避してアップデートやアップグレードすることによりLLM機能もつサードパーティのアプリケーションを使用している場合です。

LLM導入ステップ

ステップ 1: 強靭（きょうじん）であること

▶ 脅威モデリングに基づいて直近の脅威を見つける
▶ 脅威モデルシナリオにより内部・外部からの攻撃要因を見つけセキュリティ管理を確認する
▶ 取り巻く環境を精査し、有名ブランドになりすます「ごろつきアプリ」を見つける

ステップ 2: 既存ポリシーを改定する

▶ 契約、NDA、ガバナンス、セキュリティを見直し、LLM・生成AIの使用や脅威を加える

ステップ 3: トレーニングと教育

▶ セキュリティのトレーニングや社員教育、開発・法務部門向けのトレーニングを改定し、LLM・生成AIの使用や脅威を加える

ステップ 4: 組織のリーダーを取り込む

▶ 管理職、ビジネス・リーダーなど主だった役職に働きかけ、LLM・生成AIの解決戦略を定める
▶ リスク回避戦略を実行する

ステップ 5: 第三者リスク管理施策を改定する

▶ 第三者やベンダーが作成したAI機能に対して査察、精査が必要となる

ステップ 6: 導入戦略を策定する

図 2.1　LLM導入ステップ
作成 sdunn

導入戦略

LLMの適用範囲は、一般消費者向けのアプリケーションを活用するものから、個人データで独自のモデルをトレーニングするものまで多岐にわたります。利便性と統制の適切なバランスを決定するには、各々の場合、機密性の度合い、どの程度必要な機能か、それを決定するために利用可能なリソースのコストなどの要件を見極める必要があります。それら要件を見極めるための枠組みとして、5つの導入方法を以下に示します。

導入方法の種別

タイプ 1: 一般的な使用法を使う

▶ 大規模言語モデルを一般的な使用法でアクセスする
▶ 会社の方針を決め、従業員のトレーニングを行って、リスクを軽減する
▶ 利点：柔軟に、早く実験を進めることができる
▶ 例：Perplexity, ChatGPT, big-AGI

タイプ 2: モデルのAPIを経由してアクセスする

▶ ベンダー提供の大規模言語モデルのAPIを使いアクセスする
▶ 会社の方針を決め、従業員のトレーニングを行って、リスクを軽減する
▶ 利点：使用するAPIによりある程度の管理下で早く実験できる
▶ 例：Claude, ChatGPT, Gemini

タイプ 3: モデルをランセンスする

▶ エンタープライズ・クラスのサポート付きLLMをランセンスする
▶ 自社で管理しリスクを低減する。会社の方針を決め、従業員のトレーニングを行う。
▶ 利点：社内ツールやワーフローと統合し管理強化できる
▶ 例：Microsoft Enterprise CoPilot, Amazon Codewhisperer, SalesForce Einstein GPT

タイプ 4: 学習済みモデルを使う

▶ 自社データ、カスタムデータを使い基盤モデルを微調整する
▶ 透明性、機能性を増し、LLM幻覚を減らしてリスク低減する。会社の方針を決め、従業員のトレーニングを行う。
▶ 利点：機能強化、LLM幻覚を低減
▶ 例：QwenLM/Qwen 1.5, DBRX, Starling 7B

タイプ 5: 微調整済みモデルを使う

▶ 特別仕様のモデルを自社データを使い更に微調整する
▶ 透明性、機能性を増し、LLM幻覚を減らしてリスク低減する。会社の方針を決め、従業員のトレーニングを行う。
▶ 利点：一層のカスタマイズができる
▶ 例：Google MedPalm, Amazon Bedrock, Llama2, LegalAI

タイプ 6: モデルを自社開発する

▶ 自社仕様のAI/MLモデルを自社開発する
▶ 最高度の透明性と管理ができる。会社の方針を決め、開発者、従業員のトレーニングを行う。
▶ 利点：大きな投資が必要だが、高度な自社仕様

図 2.2　導入方法の種別
作成 sdunn

第三章　失敗しないLLM導入のためのチェックリスト

敵対的リスク

敵対的リスクは、相手として攻撃者だけでなく競合他社も含んでいます。

- □ 競合他社がどのようにAIに投資しているかを精査してください。AIの導入にはリスクがありますが、将来のマーケットシェアに影響を与えるビジネス上のメリットもあります。
- □ 生成AIの出現によりセキュリティ保護がもはや有効に機能しなくなった可能性のある現行の管理策を調査する。たとえば、パスワードのリセットに音声認識を利用している場合など、従来は困難だった攻撃が新たな生成AIの出現により、容易に行えるようになりました。
- □ インシデント対応計画と手引きを更新し、生成AIによる攻撃やAI・ML特有のインシデントへの対応を強化する必要があります。

脅威のモデリング

脅威を特定しプロセスとセキュリティ防御を検討するために、脅威モデリングを実施することを強く推奨します。脅威モデリングは、アプリケーション、ソフトウェア、システムのセキュリティに関する合理的な意志決定を行うための体系的で、反復可能な一連のプロセスです。生成AIを使った攻撃に対する脅威モデリングやLLMを導入する直前の脅威モデリングは、リスクを特定・軽減し、データを保護し、プライバシーを保護し、ビジネス内の安全でコンプライアンスに準拠した統合を確実にする、最も費用対効果の高い方法です。

- □ 攻撃者は、組織、従業員、経営陣、ユーザーに対する攻撃をどのようにして加速させるのでしょうか？生成AIを使用した大規模で高度な「なりすまし攻撃」を予測する必要があります。スピア・フィッシング攻撃（註1）にLLMを使うことにより、非常に巧妙で、狙う人・組織を絞りこむことができるようになります。
- □ 企業の顧客やクライアントへの攻撃に、スプーフィング（註2）や生成AIが生成したコンテンツを通じて、生成AIがどのように利用される可能性があるでしょうか？
- □ あなたのLLMソリューションへ有害または悪意のある入力やクエリがあった場合、それを検出し、未然に防ぐことができますか？
- □ LLMのすべての信頼境界で、内部のシステムとデータベースを、外部からの不正アクセスから保護できますか？
- □ 許可されたユーザーによる誤用を防ぐために、組織内部からの脅威の緩和策を講じていますか？
- □ 知的財産を保護するために、独自のモデルやデータへの不正アクセスを防止できますか？
- □ コンテンツ・フィルタリングを自動化することで、有害または不適切なコンテンツの生成を防ぐことができますか？

【註】

1. スピア・フィッシング攻撃
   特定の個人や組織を狙った偽メールによるフィッシング攻撃で、目的はログイン・パスワードなどの情報を盗む、あるいは、狙ったデバイスにマルウエアを感染させること
2. スプーフィング
   ハッカーが金融機関の口座などへのアクセス権限を得るために、他人のデバイスやユーザーになりすますこと

AI資産目録

AI資産目録は、社内で開発されたソリューションと外部またはサードパーティのソリューションの両方に適用する必要があります。

□ 既存のAIサービス、ツール、所有者を記録し、資産管理で固有のタグを指定すること。
□ ソフトウェア部品表（SBOM）は、アプリケーションに関連するすべてのソフトウェア部品、依存関係、およびメタデータの包括的なリストです。ソフトウェア部品表（SBOM）にAIを含むソフトウエア部品を含めること。
□ AIに使用するデータソースとデータの機密性（保護、機密、公開）を記録管理すること。
□ 現在の攻撃サーフェスのリスクを判断するために、設置されたAIソリューションのペンテスト（註3）またはレッドチーム（註4）が必要かどうかを確認する。
□ AIソリューションの導入時検査プロセスを策定する。
□ SBoMの要件に従い、熟練したIT管理スタッフを社内または社外から確保する。

【註】

3. ペンテスト（ペネトレーションテスト）
   ネットワーク、PC・サーバーやシステムの脆弱性を検証するテスト手法の1つ。実際にシステムに攻撃を仕掛け侵入を試みることから、「侵入テスト」とも呼ばれる。
4. レッドチーム
   セキュリティの脆弱性を検証するためなどの目的で設置された、その組織とは独立したチームのことで、対象組織に敵対したり、攻撃したりといった役割を担う。

AIセキュリティとプライバシー研修

□ 従業員と積極的に関わり、計画されているLLM導入についての従業員の懸念を理解し対処する。
□ オープンで透明性のあるコミュニケーション文化を確立し、組織のプロセス、システム、従業員の管理とサポート、顧客との関係における予測AI・生成AIの使用と管理、リスクへの対処法について意思疎通をはかる。
□ 倫理、責任、法的問題（保証、ライセンス、著作権）についての理解をすべてのユーザーに徹底

する。
□ セキュリティ意識向上トレーニングを更新し、生成AI関連の脅威を含める。ボイスクローニング、イメージクローニング、スピアフィッシング攻撃の増加を前提としたものとする。
□ 生成AIソリューションを採用する時には、DevOpsとサイバーセキュリティの両方のトレーニングを行い、AIのサイバーセキュリティとセキュリティを保証するためのデプロイメント・パイプラインを含める。

ビジネスケースの確立

確固としたビジネスケースは、AIソリューションのビジネス価値を判断し、リスクと価値のバランスを取り、投資対効果を評価・検証するために不可欠です。ビジネスケースの例を以下に示します。

□ 顧客体験の向上
□ 運用効率の向上
□ より良い知識管理
□ イノベーションの強化
□ 市場調査と競合分析
□ 文書作成、翻訳、要約、分析

ガバナンス

LLMにおけるコーポレート・ガバナンスは、組織に透明性と説明責任を提供するために必要です。デジタルプロセスへの防衛にスピーディに対応するために、テクノロジーとビジネスが目的とするユースケースに精通しているAIプラットフォーム、並びに、プロセス管理者を置くことは、十分に確立された企業の必要条件です。

□ 組織のAI RACIチャート（責任者、説明責任者、相談役、報告先）の確立
□ AIリスク、リスク評価、組織内のガバナンス責任を文書化し、担当者を割り当てます。
□ 技術的実装を含め、データの分類と使用制限に関するデータ管理ポリシーを確立します。モデルは、システムのあらゆるユーザの最小アクセス・レベルに分類されたデータのみを使用すべきです。例えば、データ保護ポリシーを更新して、保護されたデータや機密データをビジネス管理されていないツールに入力しないことを明記します。
□ 確立されたポリシー（例：善行基準、データ保護、ソフトウェア使用）に支えられたAIポリシーを作成します。
□ 様々な生成AIツールを従業員が使用する際、使用してよいかどうかを占めす一覧表を公開します。
□ 組織が LLM 生成モデルで使用するデータのソースと管理を文書化する。

法務

AIの法規定は決まっていない点が多く、非常に大きなコストがかかる可能性があります。IT、セキュリティ、法務の三部門の連携は、不確定部分を特定し、不明瞭な決定に対処するために不可欠です。

- □ 製品保証を製品開発の流れの中で明確にし、AIに関わる製品保証の責任者を割り当てます。
- □ 生成AIを考慮して、既存の利用規約を見直し、更新します。
- □ AI EULA契約書のレビューを行います。生成AIプラットフォームのエンドユーザーライセンス契約は、ユーザープロンプトの扱い方、アウトプットの権利と所有権、データプライバシー、コンプライアンス、責任の所在、プライバシー、アウトプットの使用制限など、多岐にわたり大きく異なります。
- □ AIが生成したコンテンツによる盗作、偏見の伝播、知的財産の侵害に関連する責任を組織が負うことを防ぐために、顧客向けのEULA、エンドユーザー契約を変更します。
- □ コード開発に使用されている既存のAI支援ツールの見直し。チャットボットが書いたコードを製品に使用する場合、その製品に関する企業の所有権を脅かす可能性があります。すなわち、生成されたコードの状態や保護、生成されたコードを使用する権利を誰が保持しているかが問題になる可能性があります。
- □ 知的財産に対するリスクの確認。チャットボットによって生成された知的財産は、生成プロセスが、著作権、商標、または特許保護の対象となるデータを使用する場合、危険にさらされる可能性があります。AI製品が侵害する素材を使用した場合、AIのアウトプットにリスクが生じ、知的財産の侵害につながる可能性があります。
- □ 免責条項のある契約を見直しましょう。免責条項は、責任につながる事象の責任を、その事象についてより過失があった人、あるいはその事象を阻止する可能性があったがそうしなかった人に押し付けようとするものです。責任につながる事象を引き起こしたのは、AIの提供者か、あるいはその利用者かを判断するための基準を設定する必要があります。
- □ AIシステムに起因する潜在的な傷害や物的損害に対する賠償責任を検討する。
- □ 従来のDirectors&Officers(D&O賠償責任保険や商業賠償責任保険は、AIの利用を完全に保護するには不十分である可能性が高いので、保険の見直しを行います。
- □ 著作権に関する問題の特定。著作権には人間が著作者であること必要です。LLMツールが悪用された場合、組織は剽窃、偏見の伝播、知的財産権侵害の責任を負う可能性もあります。
- □ ベンダーが開発または提供するサービスに関して、生成AIの適切な使用に関する契約条項を含めること。
- □ 権利行使が問題となる可能性がある場合、または知的財産権侵害が懸念される場合、従業員または請負業者に対する生成AIツールの使用を制限または禁止すること。
- □ 従業員管理や雇用の評価にAIを使用すると、差別待遇クレームや差別的影響クレームの原因となる可能性があります。
- □ AIソリューションが適切な同意や承認なしに機密情報を収集したり共有したりしないことを確認してください。

規制（米国、カナダ）

EUのAI法は最初の包括的なAI法になると予想されていますが、適用されるのは早くても2025年の見込みです。EUの一般データ保護規則（GDPR）はAIを特に取り上げていませんが、データ収集、データセキュリティ、公平性と透明性、正確性と信頼性、説明責任に関する規則を含んでおり、生成AIの利用に影響を与える可能性があります。米国では、AI規制はより広範な消費者プライバシー法の中に含まれています。米国では10の州で法律が成立しているか、2023年末までに施行される予定です。

カナダは、生成AIシステムの開発責任と管理に関する任意の行動綱領（Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems）を公開しています。
一方、Artificial Intelligence and Data Act (AIDA) には、それ以上強い規制が盛り込まれる見込みです。

米国雇用機会均等委員会（EEOC）、消費者金融保護局（CFPB）、連邦取引委員会（FTC）、米国司法省公民権局（DOJ）などの連邦機関は、雇用の公平性を厳しく監視しています。

- □ 国、州、またはその他の政府固有のAIコンプライアンス要件を決定する。
- □ 従業員の電子的監視および雇用関連の自動意思決定システムの制限に関するコンプライアンス要件の決定（バーモント州、カリフォルニア州、メリーランド州、ニューヨーク州、ニュージャージー州）
- □ 顔認識とAIビデオ分析に必要な同意に関するコンプライアンス要件の決定（イリノイ州、メリーランド州、ワシントン州、バーモント州）
- □ 従業員の採用や管理に使用されている、または検討されているAIツールを確認します。
- □ ベンダーが適用されるAIに関する法律やベストプラクティスを遵守していることを確認します。
- □ 人事採用プロセスでAIを使用した製品を使っている場合、文書化します。モデルがどのように訓練され、どのように監視され、差別や偏見を避けるために修正されたものを追跡しているかを確認します。
- □ どのような宿泊オプションが含まれているかを尋ね、記録しておきましょう。
- □ ベンダーが機密データを収集しているかどうかを尋ね、文書化してください。
- □ ベンダーやツールがどのようにデータを保存・削除し、入社前の顔認識やビデオ分析ツールの使用を規制しているかを尋ねてください。
- □ AIに関連して、コンプライアンス上の問題が発生する可能性のある他の組織固有の規制要件を確認してください。例えば、1974年従業員退職所得保障法（Employee Retirement Income Security Act of 1974）には、チャットボットでは対応できない可能性のある退職金制度のための教育義務要件があります。

大規模言語モデルソリューションの使用または実装

- □ 脅威モデル LLM のコンポーネントとアーキテクチャの信頼境界を見極める。

□ データ・セキュリティ、データが機密性に基づいてどのように分類され、保護されているかを検証してください。ユーザの権限はどのように管理され、どのような保護措置が取られていますか？
□ アクセス・コントロール、最小権限アクセス・コントロールの実施、徹底的な防衛策の実施
□ トレーニングパイプラインセキュリティは、トレーニングデータガバナンス、パイプライン、モデル、アルゴリズムに関する厳格な管理を必要とします。
□ 入力と出力のセキュリティでは、入力の検証方法と、出力がどのようにフィルタリングされ、サニタイズされ、承認されるかを評価します。
□ 監視と応答、ワークフロー、監視、応答をマッピングし、自動化、ロギング、監査を理解します。監査記録の安全性を確認します。
□ 製品リリースプロセスにおけるアプリケーションテスト、ソースコードレビュー、脆弱性評価、レッドチーミングの実施。
□ LLMモデルまたはサプライチェーンに脆弱性が存在するどうかチェックします。
□ プロンプト・インジェクション、機密情報の流出、プロセス操作など、LLMソリューションに対する脅威や攻撃の影響を調査します。
□ モデルポイズニング（註5）、不適切なデータの取り扱い、サプライチェーン攻撃、モデルの盗難など、LLMモデルに対する攻撃や脅威の影響を調査します。
□ サプライチェーンセキュリティ、第三者監査、侵入テスト、第三者プロバイダーのコードレビューを依頼します。(初期および継続的に）
□ インフラストラクチャ・セキュリティ、ベンダーがレジリエンス・テストを実施する頻度は？可用性、スケーラビリティ、パフォーマンスに関するSLAはどうなっていますか。
□ インシデント対応手順書を更新し、卓上演習にLLMインシデントを盛り込みます。
□ 生成AIのサイバーセキュリティを他のアプローチと比較するベンチマーク指標を策定し、または、拡張して、期待される生産性の向上を測定します。

【註】
5. モデルポイズニング
攻撃者がモデルの訓練データに不正なデータを混入し、モデルが訓練されること

試験、評価、検証、妥当性確認（TEVV）

NIST AIフレームワークでは、AIシステム運用者、ドメインエキスパート、AI設計者、ユーザー、製品開発者、評価者、監査人を含む、AIライフサイクル全体にわたる継続的なTEVVプロセスを推奨しています。TEVVには、システムの妥当性確認、統合、テスト、再較正、AIシステムのリスクや変更をナビゲートするための定期的な更新のための継続的なモニタリングなど、さまざまなタスクが含まれます。

□ AIモデルのライフサイクルを通じて、継続的なテスト、評価、検証、妥当性確認を確立します。
□ AIモデルの機能性、セキュリティ、信頼性、堅牢性に関する定期的な経営指標と最新情報の提供。

モデル・カードとリスク・カード

モデルカードとリスクカードは、大規模言語モデル（LLM）の透明性、説明責任、および倫理的な導入を向上させるための基本要素です。
モデル・カードは、AIシステムの設計、能力、制約に関する標準化されたドキュメントを提供することで、ユーザがAIシステムを理解し、信頼できるようにします。
リスクカードは、バイアス、プライバシーの問題、セキュリティの脆弱性など、潜在的な悪影響をオープンに扱うことでこれを補い、危害防止への積極的なアプローチを促します。これらの文書は、AIの社会的影響を慎重に扱い、強調して対応する土壌を確立するため、開発者、ユーザー、規制当局、倫理学者にとって重要です。これらのカードは、モデルを作成した組織によって開発され、維持されており、AI技術が倫理基準と法的要件を満たすことを保証し、AIエコシステムにおける責任ある研究と展開を可能にする上で重要な役割を果たしています。

モデルカードは、MLモデルに関連する以下のような属性を含んでいます。

- モデルの詳細：モデルに関する基本情報、すなわち、名前、バージョン、タイプ（ニューラルネットワーク、決定木など）、および意図された使用方法。
- モデルのアーキテクチャ：層の数とタイプ、活性化関数、その他の主要なアーキテクチャの選択など、モデルの構造の説明を含みます。
- トレーニングデータと手法：データセットのサイズ、データソース、使用した前処理やデータ補強技術など、モデルの学習に使用したデータに関する情報。また、使用されたオプティマイザ、損失関数、チューニングされたハイパーパラメータなど、トレーニング手法の詳細も含まれます。
- パフォーマンス測定基準：精度、正確度、再現率、F1スコアなど、さまざまな測定基準におけるモデルのパフォーマンスに関する情報。また、データの異なるサブセットでのモデルのパフォーマンスに関する情報も含まれることもあります。
- 潜在的なバイアスと限界：不均衡なトレーニングデータ、オーバーフィッティング、モデルの予測におけるバイアスなど、モデルの潜在的なバイアスや制限をリストアップします。また、新しいデータへの適応能力や特定の使用例への適合性など、モデルの限界に関する情報も含まれます。
- 責任あるAIへの配慮：プライバシーに関する懸念、公平性、透明性、またはモデルの使用による潜在的な社会的影響など、モデルに関連する倫理的または責任あるAIの考慮事項。また、モデルのさらなるテスト、検証、モニタリングに関する推奨事項が含まれる場合もあります。

モデルカードに含まれる正確な機能は、モデルのコンテキストと使用目的によって異なる可能性がありますが、その目的は、機械学習モデルの作成と展開に公開性と説明責任を与えることです。

□ モデルカードの確認

□ リスクカードがある場合はそれを確認
□ サードパーティを通じて使用されるモデルを含む、あらゆる導入モデルのモデルカードを追跡し、維持するプロセスを確立します。

RAG：大規模言語モデルの最適化

ファインチューニングは、事前に訓練されたモデルを最適化するための手法で、既存のモデルを新しいドメイン固有のデータで再学習させ、タスクやアプリケーションのパフォーマンスに合わせて行います。ファインチューニングにはコストがかかりますが、パフォーマンス向上には不可欠です。

検索補強型生成（Retrieval-Augmented Generation, RAG）は、最新の利用可能な知識ソースから適切なデータを検索することにより、大規模な言語モデルの能力を最適化し補強する、より効果的な方法として発展してきました。RAGは特定のドメイン用にカスタマイズすることができ、ドメイン固有の情報の検索を最適化し、特殊な分野のニュアンスに合わせて生成プロセスを調整します。RAGは、LLM最適化のための、より効率的で透明性の高い手法と考えられており、特にラベル付きデータの収集が限られていたり、高価であったりする場合に適しています。また、RAGの利点の一つは、検索段階で新しい情報を継続的に更新することができるため、継続的な学習をサポートすることです。

RAGの実装はいくつかのステップがあります。まず、埋め込み情報モデルの導入から始まり、知識ライブラリの索引付け、クエリ処理のための最も関連性の高い文書の検索まで、いくつかの重要なステップを含みます。関連するコンテキストの効率的な検索は、文書埋め込み情報の保存とクエリに使用されるベクトルデータベースを使って行います。

RAG 参照・リンク
□ 検索拡張世代（RAG）とLLM：例
□ 12 RAGの問題点と解決案

AIレッドチーム

AIレッドチームとは、攻撃者が悪用できる脆弱性が存在しないことを確認するために、AIシステムを敵対的に攻撃するテストシミュレーションです。これは、バイデン政権を含む多くの規制機関やAI管理機関によって推奨されています。しかし、レッドチーミングだけでは、AIシステムに関連するすべての実害を検証する包括的な解決策とはなりません。アルゴリズムによる影響評価や外部監査など、他の形式のテスト、評価、検証、妥当性確認と組み合わせる必要があります。

□ AIモデルとアプリケーションの標準的なプラクティスとして、レッドチームテストを導入します。

第四章　LLM導入に役立つリソース

OWASP 大規模言語モデル・アプリケーション リスク トップ10

LLM01: プロンプト・インジェクション
巧妙な入力によって大規模な言語モデル（LLM）を操作し、LLMが意図しない動作を引き起こします。システムのプロンプトを直接、上書きする手法、外部ソースからの入力を操作し、間接的に行う手法があります。

LLM02: 安全が確認されていない出力ハンドリング
LLMの出力を細かくチェックせずにバックエンドシステムに送った場合、バックエンドの脆弱性をつかれ、意図しない結果を引き起こすことです。悪用されると、XSS、CSRF、SSRF、特権の昇格、リモート・コードの実行といった深刻な結果につながる可能性があります。

LLM03: 訓練データの汚染
LLMの訓練データが改ざんされ、セキュリティ、有効性、倫理的行動を損なう脆弱性やバイアスなどが入り込むことです。訓練データの情報源として、Common Crawl、WebText、OpenWebText、書籍などが使われます。

LLM04: モデルのDoS
LLMが計算リソースを大量に消費するようにしむけ、LLMを使ったサービスの品質低下や高コストを狙ったものです。

LLM05: サプライチェーンの脆弱性
LLMアプリケーションが使用するコンポーネントやサービスの脆弱性によって引き起こされる攻撃です。サードパーティのデータセット、事前に訓練されたモデル、およびプラグインを使用することで脆弱性が増す可能性があります。

LLM06: 機微情報の漏えい
LLMは、その応答の中に意図せず機密データを含めてしまう可能性があり、不正なデータアクセス、プライバシー侵害、セキュリティ侵害につながります。これを軽減するためには、データの浄化と厳格なユーザー・ポリシーを導入することが極めて重要です。

LLM07: 安全が確認されていないプラグイン設計
LLMプラグインにおいて、入力の安全性が確認されておらず、あるいはアクセスコントロールが不十分な場合、悪意のあるリモート・コード実行のような結果をもたらす可能性があります。

LLM08: 過剰な代理行為
この問題は、LLMベースのシステムに与えられた過剰な機能、権限、または自律性に起因し、意図しない結果を招くことがあります。

LLM09: 過度の信頼
十分監督されていないLLMに過度に依存したシステムやユーザーは、LLMが生成したコンテンツが不正確または不適切なものであることに気づかず、誤った情報、誤ったコミュニケーション、法的問題、セキュリティの脆弱性に直面する可能性があります。

LLM10: モデルの盗難
独自のLLMモデルへの不正アクセス、モデルのコピー、または流出が含まれます。その影響は、経済的損失、競争上の優位性の低下、機密情報へのアクセスの可能性などです。

図 4.1　OWASP 大規模言語モデル･アプリケーション リスク トップ10

OWASP 大規模言語モデル・アプリケーション リスクの所在箇所

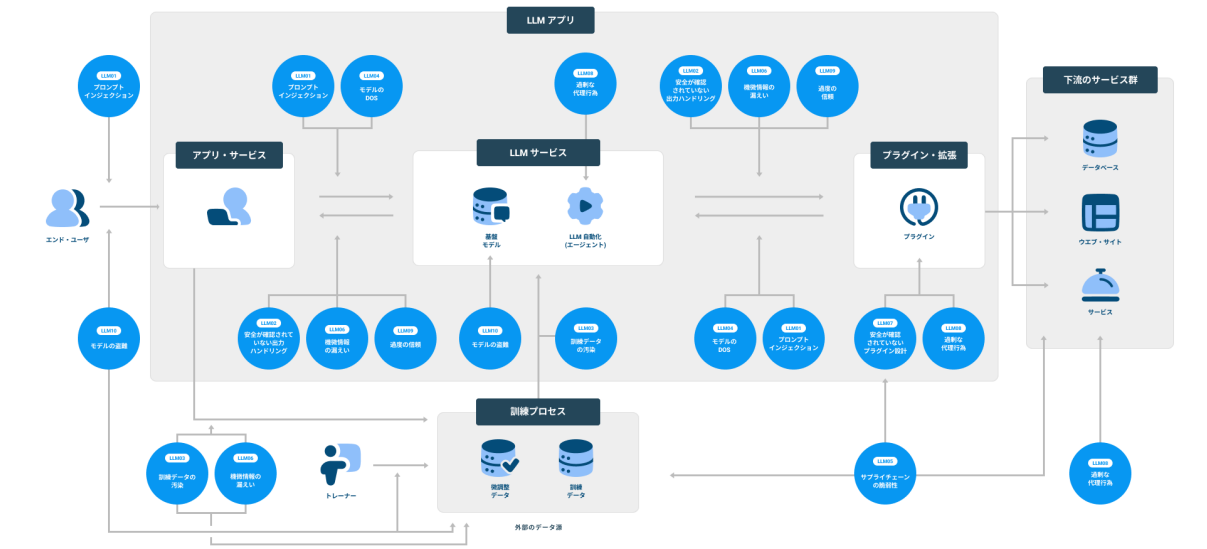

図 4.2　OWASP 大規模言語モデル・アプリケーション リスクの所在箇所

OWASPのリソース

LLMソリューションを導入すると、それまでになかった領域が攻撃対象となり、新たな課題が発生するため、特別な戦術や防御が必要になります。同時に、よく知られている問題と類似した問題も発生し、すでに確立されたサイバーセキュリティの手順と緩和策を適用することができる場合もあります。LLMサイバーセキュリティを、それらすでに確立されたサイバーセキュリティ管理、プロセス、手順と統合することで、脅威に対する脆弱性を減らすことができます。

これらがどのように統合されるかは、OWASP Integration Standards（OWASP統合基準）をご覧ください。

OWASP リソース・リンク

OWASP SAMM

説明

ソフトウェア保証成熟度モデル

お勧めする理由と使用方法

効果的で組織の安全な開発ライフサイクルを分析し、改善するための測定可能な方法です。SAMMは、ソフトウェアライフサイクル全体をサポートします。SAMMはリスクを少しずつ絞り込む手法を取り、ソフトウエア開発におけるギャップを特定し、優先順位を付け、要点を絞って、セキュアな開発を可能にします。

OWASP リソース・リンク

OWASP AI Exchange

OWASP 機械学習セキュリティ トップ10

説明

OWASPのプロジェクトは、全世界でAIセキュリティを高め、標準化と相互協力を支援していま

す。

OWASP AI ExchangeはOWASP AIセキュリティ・プライバシーガイドのための窓口です。

OWASP 機械学習セキュリティトップ10は、MLシステムのセキュリティの問題を対象にしています。

お勧めする理由と使用方法

このプロジェクトは、ML Top 10 を含み、安全でプライバシーを保護するAIシステムの設計、作成、テスト、調達に関する明確で実行可能な施策を提供するため、常時、更新されている文書です。これは、AIのグローバルな規制とプライバシー情報のための最高のOWASPリソースです。

OWASP リソース・リンク

Open Common Requirement Enumeration OpenCRE

説明

OpenCRE (共通要件列挙)は、セキュリティ標準とガイドラインの両方を外観するためのプラットフォームで、少しずつ絞り込んでいくことができます。

お勧めする理由と使用方法

標準、規格を検索するため。規格名またはコントロールタイプで検索できます。

OWASP リソース・リンク

OWASP脅威モデリング

説明

アプリケーションの脅威モデリングを行うための構造化された正式なプロセスです。

お勧めする理由と使用方法

脅威モデリングについて広範な知識を得ることができます。アプリケーションのセキュリティに関する情報を体系的にまとめています。

OWASP リソース・リンク

OWASP CycloneDX

説明

OWASP CycloneDXは、サイバーリスク低減のための高度なサプライチェーン機能を提供するフルスタックの部品表（BOM）規格です。

お勧めする理由と使用方法

ソフトウェアはサードパーティやオープンソースのコンポーネントを使用することが多くなっています。これらのコンポーネントは、複雑でそれぞれの用途に応じて異なった方法で使われ、目的とする機能を実現します。SBOM（Software Bill of Materials）は、リスクを特定し、透明性を高め、迅速な影響分析を可能にする、すべてのコンポーネントの正確な部品表です。

EO 14028 は、米国連邦政府システムのSBOMに関する最低要件を規定しています。

OWASP リソース・リンク

OWASP ソフトウェアコンポーネント検証規格（SCVS）

説明

アクティビティ、コントロール、ベストプラクティスを作成するためのフレームワークを確立するコ

ミュニティ主導の取り組みです。

お勧めする理由と使用方法

ソフトウェアサプライチェーンのリスクを減らし、さらに進んだソフトウェア・サプライチェーン管理を実現するための施策、管理、最善の対処法を策定するのに役立ちます。

OWASP リソース・リンク

OWASP API セキュリティ プロジェクト

説明

アプリケーション・プログラミング・インターフェース（API）のセキュリティに関するユニークな脆弱性とセキュリティリスクを理解し、軽減するための戦略とソリューションについて説明しています。

お勧めする理由と使用方法

アプリケーションはAPIを経由して接続されています。ユーザと組織を保護するためには、APIの誤設定や脆弱性を減らすことが必須です。ビルド環境と製品環境のセキュリティテストとレッドチームに使用することができます。

OWASP リソース・リンク

OWASP Top 10 CI/CD セキュリティ リスク

説明

継続的インテグレーションと高い頻度の自動デプロイのセキュリティのための要点を示しています。

お勧めする理由と使用方法

開発者の手から製品環境まで、ソフトウエアをデプロイするプロセスのセキュリティを高めるために役立ちます。ビルド環境と製品環境のセキュリティテストとレッドチームに使用することができます。

OWASP リソース・リンク

OWASP アプリケーション セキュリティ 検証基準 ASVS

説明

アプリケーション・セキュリティ検証標準（ASVS）は、ウェブアプリケーションの技術的なセキュリティ管理をテストするための基礎を提供し、また、開発者に安全な開発のための要求事項のリストを提供します。

お勧めする理由と使用方法

ウェブアプリケーション用セキュリティ要件、セキュリティテスト、一覧表を作成するのに使用します。様々なアプリケーションの使い道、ユーザーの流れに沿ったリリース・テストを確立するために役立ちます。

OWASP リソース・リンク

OWASP 脅威およびセーフガードマトリックス (TaSM)

説明

ビジネスを守るために、すぐに適用可能な方策を示す一覧を提供します。

お勧めする理由と使用方法

この一覧により、主要な脅威をNISTサイバーセキュリティフレームワークの機能（特定、保護、検知、対応、回復）に重ね合わせ、強固なセキュリティ計画を構築することができます。組織全体のセキュリティを追跡し、報告するダッシュボードとして使用できます。

OWASP リソース・リンク

欠陥 道場

説明

オープンソースの脆弱性管理ツールで、テンプレート化、レポート作成、数値化により、テストプロセスを合理化します。使用例を示すツールも用意されています。

お勧めする理由と使用方法

「欠陥 道場」を使えば、テンプレートを使い、一般的な脆弱性を検出し、レポート生成、数値化し、脆弱性のログ作成時間を短縮できます。

MITREのリソース

LLM の脅威が頻発するようになったことで、攻撃対象領域を防御するために「強靭さ第一」の価値が再認識されてきています。従来のTTPS（註1）は、LLMにおける新たな攻撃対象領域と機能を取り込んでいます。MITREは、実際に起こった例に基づいて攻撃者の戦術と手順を理解するために、確立され広く受け入れられている枠組みを使っています。

【註】

1. TTPS：Tactics, Techniques, and Procedures

攻撃者がサイバー攻撃を仕掛ける際にとる行動、戦術、手順を記述したもの

組織のLLMセキュリティ戦略に、MITRE ATT&CK と MITRE ATLAS を適用することで、LLMセキュリティが現在のAPIセキュリティ標準などのプロセスでカバーされている部分、カバーされていない部分を特定することができます。

MITRE ATT&CK（Adversarial Tactics, Techniques, and Common Knowledge）は 、MITRE Corporationによって作成されたフレームワーク、データマトリックス集、および評価ツールで、世界中で使用されている知識リポジトリです。MITRE ATT&CKマトリックスは、攻撃者が特定の目標を達成するために使用する戦略集を含んでおり、これらの目的は戦術ごとに分類されています。さらに、攻撃順に示されており、偵察（reconnaissance）から始まり、最終的な目標である殲滅（exfiltration）または破壊（impact）へと進んでいきます。

MITRE ATLASは「Adversarial Threat Landscape for Artificial Intelligence Systems」の略で、悪質な行為者による機械学習（ML）システムへの攻撃の実例に基づいた知識ベースです。ATLASはMITREのATT&CKアーキテクチャに基づいており、その戦術と手順はATT&CKを補完するものです。

MITRE リソース・リンク

MITRE ATT&CK

説明

実際の観察に基づく攻撃者の戦術とテクニックの知識ベース

お勧めする理由と使用方法

具体的な脅威モデルと方法論の開発の基礎として使用します。組織内の既存の管理策を攻撃者の戦術や技法に照らし合わせ、抜けている部分やテストすべき領域を特定します。

MITRE リソース・リンク

MITRE AT&CK 作業台

説明

ローカルな知識ベースを使って、ATT&CKの作成または拡張する際に使うツールです。

お勧めする理由と使用方法

ATT&CK 知識ベースをコピーし、新しい、または更新されたテクニック、戦術、緩和策グループ、および組織に特化したソフトウェアで拡張することができます。

MITRE リソース・リンク

MITRE ATLAS

説明

MITRE ATLAS (Adversarial Threat Landscape for Artificial-Intelligence Systems)は、攻撃戦術や攻撃手法をMLに特化した知識ベースです。実例やレッドチーム・セキュリティチームによるデモ、研究者の挙げた可能な手法が網羅されています。

お勧めする理由と使用方法

すでに知られているML特有の脆弱性を知ることができます。

MITRE リソース・リンク

MITRE ATT&CK Powered Suit

説明

ATT&CK Powered Suit は、MITRE ATT&CK知識ベースをあなたの手元に置くブラウザ拡張機能です。

お勧めする理由と使用方法

ワークフローを中断することなく戦術、テクニックなどを検索することができます。

MITRE リソース・リンク

The Threat Report ATT&CK Mapper (TRAM)

説明

CTI(Cyber Threat Intelligence)レポートのTTP(tactics,techniques,procedures)検索を自動化します。

お勧めする理由と使用方法

CTIレポートに見られるTTPをMITRE ATT&CKに対応させる手順は、エラーが発生しやすく、時間がかかります。TRAMはLLMを使用し、最も一般的な50のテクニックについてこのプロセスを自動化します。Juypterノートブックも用意されています。

MITRE リソース・リンク

Attack Flow v2.1.0

説明

アタックフローはサイバー攻撃者が、様々な攻撃技術をどのように組み合わせ、目的を達成するのかを記述するための言語です。

お勧めする理由と使用方法

攻撃者がどのようなテクニックを使うかを理解することで、防御者は敵の動きを理解し、自らの防御態勢を向上させることができます。

MITRE リソース・リンク

MITRE カルデラ
Caldera 用のプラグイン

説明

攻撃者の動きをシミュレートし、対するレッドチームを支援し、インシデント応答を自動化するように設計された、サイバーセキュリティ・プラットフォームです。

お勧めする理由と使用方法（プラグインへのリンク付）

プラットフォームのコア機能を拡張し、エージェント、レポート、TTP のコレクションなどの追加機能を提供するプラグインも用意されていま。

MITRE リソース・リンク

CALDERA ラグイン: アーセナル

説明

AI対応システムへの攻撃者をシミュレートするプラグインです。

お勧めする理由と使用方法

このプラグインは、CALDERA とインターフェースするため、MITRE ATLAS に定義された TTP を提供します。

MITRE リソース・リンク

アトミック・レッド・チーム

説明

MITRE ATT&CKフレームワークで使う、攻撃者をシミュレートするテストのライブラリです。

お勧めする理由と使用方法

コントロールの検証とテストのために使用することができます。セキュリティチームは、Atomic Red Teamを使用することで、自分たちの環境を素早く、ポータブルに、再現性よくテストすることができます。アトミックテストはコマンドラインから直接実行できます。

MITRE リソース・リンク

MITRE CTI ブループリント

説明

サイバー脅威情報のレポートを自動化します。

お勧めする理由と使用方法

CTIブループリントは、サイバー脅威インテリジェンス（CTI）アナリストが、高品質で実用的なレ

ポートを効率的な作成を支援します。

AI脆弱性リポジトリ

名称

AI インシデント・データベース

説明

過去に失敗したAIアプリケーションの事例を蓄積しています。大学の研究グループによって維持され、クラウドソーシングされています。

名称

OECD AI インシデント・モニター (AIM)

説明

AIに関連する課題を理解するためのわかりやすい出発点を提供します。

AIモデルの脆弱性を追跡する企業

1. Huntrバグバウンティ：プロテクトAI
   AI/ML向けバグ報奨金プラットフォーム
2. AI脆弱性データベース(AVID)：ガラク
   モデルの脆弱性データベース
3. AIリスクデータベース：ロバスト・インテリジェンス
   モデルの脆弱性データベース
4. LVE Repository
   オープン LLM Vulnerability and Exposures レポジトリ

AI調達ガイダンス

名称

世界経済フォーラム：AI責任の導入：民間セクターによるAIソリューション調達のためのガイドライン：インサイトレポート 2023年6月

説明

AIシステムの調達のための、標準ベンチマークと評価基準は、まだ開発初期段階にあります。この調達ガイドラインは、エンド・ツー・エンドの調達プロセスにおける考慮事項の基準を提供しています。このガイダンスを使うと、既存の「サードパーティ・リスク・サプライヤーとベンダー調達プロセス」を強化することができます。

チーム

OWASP 大規模言語モデル アプリケーション リスクトップ10 サイバーセキュリティとガバナンス チェックリストの作成に貢献された皆さんに感謝いたします。

チェックリストの作成

Sandy Dunn
Heather Linn
John Sotiropoulos
Steve Wilson
Fabrizio Cilli
Aubrey King
Bob Simonoff
David Rowe
Ken Huang
Emmanual Guilherme Junior
Andrea Succi
Jason Ross
Talesh Seeparsan
Anthony Glynn
Julie Tao
Cédric Lallier
Tetsuo Seto
Ads Dawson

日本語版の作成

Tetsuo Seto
Riotaro Okada