LIBRUS

AIセキュリティ最新動向・起こりそうなトラブル解説

自己紹介

翁駿暁（オウシュンギョウ）

浙江大学情報コンピューターサイエンス学部、早稲田大学大学院 国際情報通信研究科卒業。

シンプレクス株式会社にてカリフォルニア大学ロサンゼルス校金融工学短期研修了したのち、証券会社における取引システムの開発、導入等のメンバー、リーダー等を担当。

アビームコンサルティング株式会社、PwCコンサルティング合同会社にて金融機関を中心としたITコンサルティング業務に従事。

現在、Librus株式会社取締役COOを担当。
社内のIT技術領域全般を管掌し、システム開発およびサイバーセキュリティを中心とした技術戦略の推進を担当

趣味：ゲーム（PS5、Switch）、アニメ、筋トレ、格闘技、ゴルフ

会社概要

会社概要

私たちは「ストラテジー」と「テクノロジー」を通じてあらゆるビジネスをより豊かにすることを目指します。

社名	Librus株式会社
設立年月	2017年11月
本社所在地	〒105-0004　東京都港区新橋6丁目13-12 VORT新橋Ⅱ 4F
資本金	2,000万円
代表取締役 CEO	鎌田光一郎：青山学院大学 法学部卒業。SMBC日興証券株式会社にて証券営業、経営管理業務に従事したのちPwCコンサルティング合同会社に転籍。金融機関に対するコンサルティング業務に従事。その後、Librus株式会社を設立、代表取締役に就任。
取締役 COO	翁駿暁：浙江大学 理工学院 情報コンピューターサイエンス学部、早稲田大学大学院 国際情報通信研究科卒業。情報通信学修士。シンプレクス株式会社にてカリフォルニア大学ロサンゼルス校金融工学短期研修修了したのち、証券会社における取引システムの開発、導入等のメンバー、リーダー等を担当。その後、アビームコンサルティング株式会社、PwCコンサルティング合同会社にて金融機関を中心としたコンサルティング業務に従事。現在はLibrus株式会社取締役COOに就任。
従業員数	110名
認証	ISO 9001、ISO/IEC 27001 情報セキュリティサービス基準適合サービスリスト 人材派遣業/人材紹介業
主な所属	公益社団法人経済同友会 一般社団法人ソフトウェア協会など

事業領域

当社はサイバーインテグレーターとして、「Cyber Security」「SI&DX」「Strategy」の3領域を主軸に事業活動を展開しています。

CS	サイバーセキュリティ	・組織のリスク管理の体系化/高度化、**セキュリティ診断/SOC**などを通じた、急増する**サイバーセキュリティリスク**への対応支援 ・**社内の個人情報の検出/管理**や**WAF**ソリューション導入支援
SI & DX	システム設計/開発/保守運用	・高品質かつ迅速な**ウェブ/スマートフォンアプリケーション**その他の開発 ・高品質のドキュメントをベースとした**低リスク/高品質でプロジェクト**の推進 ・**コンサルタント、プロジェクトマネージャー、エンジニア**の調達支援
	AI/データマネジメント	・DXに伴う**データ分析**や活用、**基盤構築**支援 ・AIを活用した新サービスの立ち上げや既存サービスの高度化支援 ・その他データを活用して企業収益の最適化推進
Str	事業戦略/新規事業/DX	・事業戦略の高度化/その他コスト改善等 ・デジタルや金融ビジネスの知見を活用した新規事業の立ち上げ/運用 ・DXによる事業の最大化やコストの最適化
	マーケティング	・コーポレートサイトやIRサイトのリニューアルを通じた**ブランディング強化** ・デジタルを通じた自社の**ブランドや認知度を向上** ・**サイト制作**や**広告運用**、**マーケティング分析**を通じた、収益性の向上
	ファイナンス&リスク	・企業価値向上などを目的としたM&A仲介/アドバイザリーサービス ・IR/その他財務コンサルティングサービス ・コーポレートガバナンス、リスクマネジメント等に関するコンサルティングサービス

はじめに｜数字で見る、AIセキュリティの現在地

AIセキュリティのリスクは、もう統計にはっきり表れている

80%超

従業員の8割超が、会社非公認のAIツールを業務利用（シャドーAI）

会社が把握しているのは氷山の一角

約27%

AIツールへの入力プロンプトの約3割に、機密・社外秘情報が含まれる

悪意ではなく「便利さ」が漏洩の入口に

+67万ドル（約1億円）

シャドーAIが関与した情報漏洩は、侵害コストを平均67万ドル押し上げ

調査対象の侵害の20%にシャドーAIが関与

今日話すこと

5つの「起こりそうな」シナリオ ― すべて実在の事件・脆弱性ベース

① シャドーAIと機密漏洩　② 間接プロンプトインジェクション　③ エージェント・MCP
④ AIコーディングとサプライチェーン　⑤ ディープフェイク詐欺

対策と最新動向

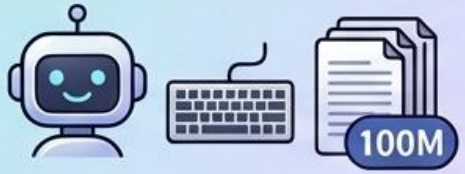
攻撃者のAI利用、フロンティアAIの脅威、
規制・標準の動き、対策フレームワーク

出所：IBM Cost of a Data Breach Report 2025、各種Shadow AI利用実態調査（2025-2026）

AIの進化状況｜4年で「道具」から「実行者」へ

生成AIは4年で「文章を書く道具」から「自律的に作業する実行者」になった

2022-23｜生成AIの登場



ChatGPTの衝撃

- 公開2ヶ月で1億ユーザー
- 文章・コード・画像の生成が一般業務に一気に浸透

2024｜マルチモーダルx推論

「目と耳」を持つAI

- 画像・音声・長文脈に対応
- 複雑な推論が実用レベルに
- 社内RAG活用が本格化

2025｜エージェント元年

AIが「作業」を実行する

- ツール操作・自律実行が標準化（MCP）
- 開発現場の主力に
- 初のAI主導型攻撃も発生

2026｜フロンティア時代

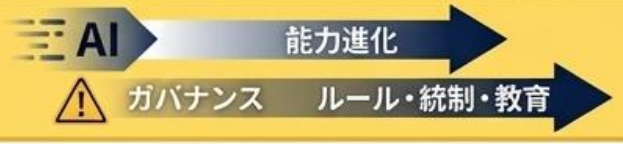
専門家超えの領域が拡大

- ゼロデイ自律発見など「超人的」能力が出現
- 能力閾値・安全枠組みの運用が前提に

デュアルユースこそがAIセキュリティの本質

AIの能力向上は「攻撃にも防御にも」等しく効く

→ 能力の進化スピードに、利用ルール・統制・教育が追いついていないことが、今日のすべてのシナリオの背景にある



出所：各社公表資料・公開研究より作成

守る対象が「システム」から「入力・データ・自律性」へ広がった

守る対象が「システム」から「入力・データ・自律性」へ広がった

従来のセキュリティ

システムを守る

- ネットワーク／OS
- アプリケーションの脆弱性
- マルウェア対策

WAF・EDR・アンチウイルス等の
既存対策でカバー

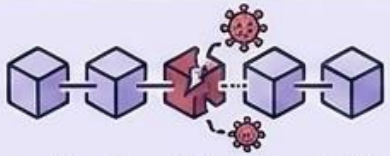
1 LLM01 プロンプトインジェクション

自然言語が「攻撃コード」になる。
世界標準OWASPでも第1位のリスク。

2 LLM02 機密情報の漏洩

プロンプト・学習データ・応答経由で
機密が流出。

3 LLM03 サプライチェーン



モデル・パッケージ・ツール定義の
汚染が波及。

4 LLM06 過剰な自律性

権限を持ちすぎたAIエージェントが
想定外の操作を実行。

国際標準（OWASP LLM Top 10 2025）でも「入力・データ・自律性」が最重要リスク。

シナリオ①　シャドーAIと機密情報漏洩

「この顧客リスト、傾向を分析して」— 貼り付けた瞬間、統制の外へ

STEP 1｜場面

社外AIに機密を貼り付け

従業員の8割超が非公認AIを利用。プロンプトの約3割に機密・社外秘が混入している

STEP 2｜進行

事業者側にデータが残る

ログ・学習データとして残存。規約次第で二次利用も。自社の統制・監査が及ばない

STEP 3｜被害

漏洩コストの増大

個情法の報告義務・信用失墜。シャドーAI関与の侵害はコストが平均**67万ドル増加**

今日から
できる対策

- ✓ AI利用をインベントリ化し、データ格付け（公開/社内/機密）と「入力してよい情報」を明文化
- ✓ 法人契約（学習オプトアウト・SSO・監査ログ）へ誘導し、DLPでプロンプト内の機密を自動検知

出所：IBM Cost of a Data Breach Report 2025

実例詳解｜シャドーAI起因の情報漏洩

2つの事件が示すもの ―「入力した瞬間」と「事業者側」の両方で漏れる

Case 1｜Samsung 半導体部門（2023年4月）

経緯：業務効率化のためChatGPT利用を解禁した直後、わずか約20日間で3件の機密入力が発覚

・設備の不具合解析のため、半導体関連の機密ソースコードを貼り付けて修正を依頼

・歩留まり・不良設備検出プログラムのコードを入力し、最適化を依頼

・社内会議の録音データから議事録を作るため、会議内容をそのまま入力

影響と対応：入力データは外部サーバーへ送信され回収不能。同社は全社で生成AI利用を一時禁止し、再開後も入力容量制限・社内専用AIの整備等の統制を導入

Case 2｜DeepSeek DB露出（2025年1月）

経緯：急成長中のAIサービスの内部データベース（ClickHouse）が、認証なしでインターネットから閲覧可能な状態でWiz Researchにより発見される

・利用者のチャット履歴・APIキー・内部運用情報など**100万件超のログ**が露出

・第三者がDBを操作し権限昇格まで可能な状態だった

影響と対応：通報後に速やかに修正。ただし「入力した機密がどこに保存され、誰がどう守るか」を利用者側は選べない構造的リスクが顕在化。各国で政府端末からの利用禁止が拡大

「便利さ」が先行するサービスほどデータの行き先は見えない ― 入力前の統制がすべて

出所：Samsung関連の各社報道（2023年5月）、Wiz Research公表（2025年1月）

シナリオ②　間接プロンプトインジェクション

メール1通・クリック0回で情報流出 ―「EchoLeak」が示した現実（CVE-2025-32711）

STEP 1｜仕込み

隠し命令入りメールを送付

白文字・HTMLコメントで人間には見えない命令を埋め込み。注入検知フィルタも回避

STEP 2｜実行

AIがメールを「文脈」として取得

M365 CopilotのRAGが悪性メールを参照し、命令を実行。ユーザー操作はゼロ

STEP 3｜被害

社内文書・メールを外部送出

画像の自動取得を悪用してデータを持ち出し。
CVSS 9.3の重大脆弱性

今日から できる対策

- ✓「データと命令をAIは区別できない」= パッチでは根絶できない構造的リスクとして扱う
- ✓ AIが参照するデータ範囲を最小化し、出力の監視と外部送信のフィルタリングを実装する

出所：Aim Security / Microsoft（2025年6月公表、サーバー側で修正済み）

シナリオ③　AIエージェント・MCPの悪用

「ツールを持ったAI」は、攻撃者にとって最高の踏み台になる

STEP 1｜接続

エージェントに社内ツールを接続

MCP等の標準化でメール・DB・決済までAIが操作可能に。導入が急拡大中

STEP 2｜汚染

ツール定義・参照データに命令混入

ツールポイズニング。多くのクライアントはツール定義を検証せずに受け入れる

STEP 3｜被害

AIが自律的に不正操作

データ持ち出し・コマンド実行。2026年は**CVSS 9超のMCP関連脆弱性**が続出

今日から できる対策

- ✓ 接続ツール・MCPサーバーを棚卸しし、出所を確認。エージェント権限は最小化・読み取り専用から
- ✓ 送信・削除・決済など重要操作は人間の承認（Human in the Loop）を必須にし、操作ログを監視

出所：MCP関連脆弱性の公開研究・アドバイザリ（2025-2026）

シナリオ④　AIコーディングとサプライチェーン

生成コードの約2割が「存在しないパッケージ」を参照している

STEP 1｜生成

AIが幻覚パッケージを提案

研究では生成コードの**19.7%が実在しないパッケージ**を参照。約20.5万種の幻覚名を確認

STEP 2｜先回り

攻撃者が同名の悪性パッケージを公開

幻覚名の58%は繰り返し出現＝予測可能。「スロップスクワッティング」として実害化

STEP 3｜被害

サプライチェーン汚染

悪性コードが依存関係経由で本番・全社へ。混入後の検知は極めて困難

今日から できる対策

- ✓ ハッシュ検証・依存スキャナで、パッケージの実在と完全性を必ず確認
- ✓ AI生成コードも人間レビュー＋SAST/SCAを必須ゲートにする（「動いたのでOK」を禁止）

出所：大学共同研究（USENIX Security 2025 採択論文ほか）

シナリオ⑤　ディープフェイク・AIフィッシング

ビデオ会議の「全員」が偽物だった ― Arup社、約2,500万ドルの被害（2024）

STEP 1｜準備

公開動画から役員を複製

CFOと同僚の姿・声を、決算説明会など公開映像から生成。数秒のサンプルで足りる

STEP 2｜接触

全員偽物の会議で送金指示

「機密案件」としてビデオ会議で指示。本人確認のつもりの通話が、逆に信用させた

STEP 3｜被害

15回・5口座に分割送金

約**2,540万ドル**が流出。発覚は本社への事後確認。資金は回収されていない

今日からできる対策

- ✓ 送金・権限変更は映像・音声で本人確認しない。別経路での確認が必須
- ✓ 「至急・機密・分割送金」という典型パターンを全社で共有し、訓練で体に覚えさせる

出所：香港警察発表・CNN報道（2024年5月）

視点の転換｜攻撃者もAIを使い始めた

2025年9月、初の「AIが実行する」スパイ攻撃が確認された（GTG-1002）

出所：Anthropic「Disrupting the first reported AI-orchestrated cyber espionage campaign」（2025年11月）

最先端AIはゼロデイ、大量脆弱性を自律発見する段階へ ― 攻防の時間感覚が変わる

ALL

全主要OS･ブラウザで発見

Windows、macOS、Linux等の主要OSと主要Webブラウザのすべてにおいて、未知の脆弱性（**ゼロデイ**）を特定しました。

99%超

驚異的な未パッチ率

AIによって大量に発見された深刻な脆弱性のうち、99%以上が発見時点においてパッチ未適用の状態であることが判明しました。

27年間

長期潜伏バグの特定

これまで世界中のセキュリティ専門家が誰一人として気づかなかった、27年間もシステムに潜伏し続けていた欠陥を発見しました。

未経験者でも可能

実動エクスプロイト生成

セキュリティの専門訓練を受けていないエンジニアであっても、AIの支援により実用的な攻撃コードの生成が可能になっています。

出所：Anthropic Red / BBC / CETaS

（参考）規制・ガイドラインの最新動向

「何をすべきか」は出揃いつつある ― 問われるのは実装

日本

AI推進法（2025年9月全面施行）

理念法。企業への罰則はないが、活用事業者にも努力義務

AI事業者ガイドライン 第1.2版

（2026年3月改訂）総務省・経産省。日本のAIガバナンスの統一指針

海外

EU AI Act ― 段階施行中

汎用AIモデルの義務（2025年8月～）

高リスクAIの義務（2026年8月～順次）

日本企業もEU域内提供なら対象になりうる

標準・フレームワーク

ISO/IEC 42001（AIマネジメント）

認証取得が取引条件になる動きも

NIST AI RMF ／ OWASP LLM Top 10

リスク評価・技術対策の実務標準として定着

セキュリティ対策は「コンプライアンス要件」へ ― ガイドライン準拠が取引・入札の前提になり始めている

出所：内閣府・総務省・経産省公表資料、EU AI Act

対策フレームワーク｜3つの層で守る

キーワードは「禁止」ではなく「安全に使える仕組み」

まとめ

LIBRUS

AIセキュリティは「未来の話」ではなく「明日の業務」の話

すべて現実に起きている

EchoLeak・Arup事件・AI主導型攻撃 — 今日の話はすべて実在の事件・脆弱性

構造的リスクと捉える

プロンプトインジェクションはパッチで消えない。仕組みと権限設計で抑え込む

始める材料は揃っている

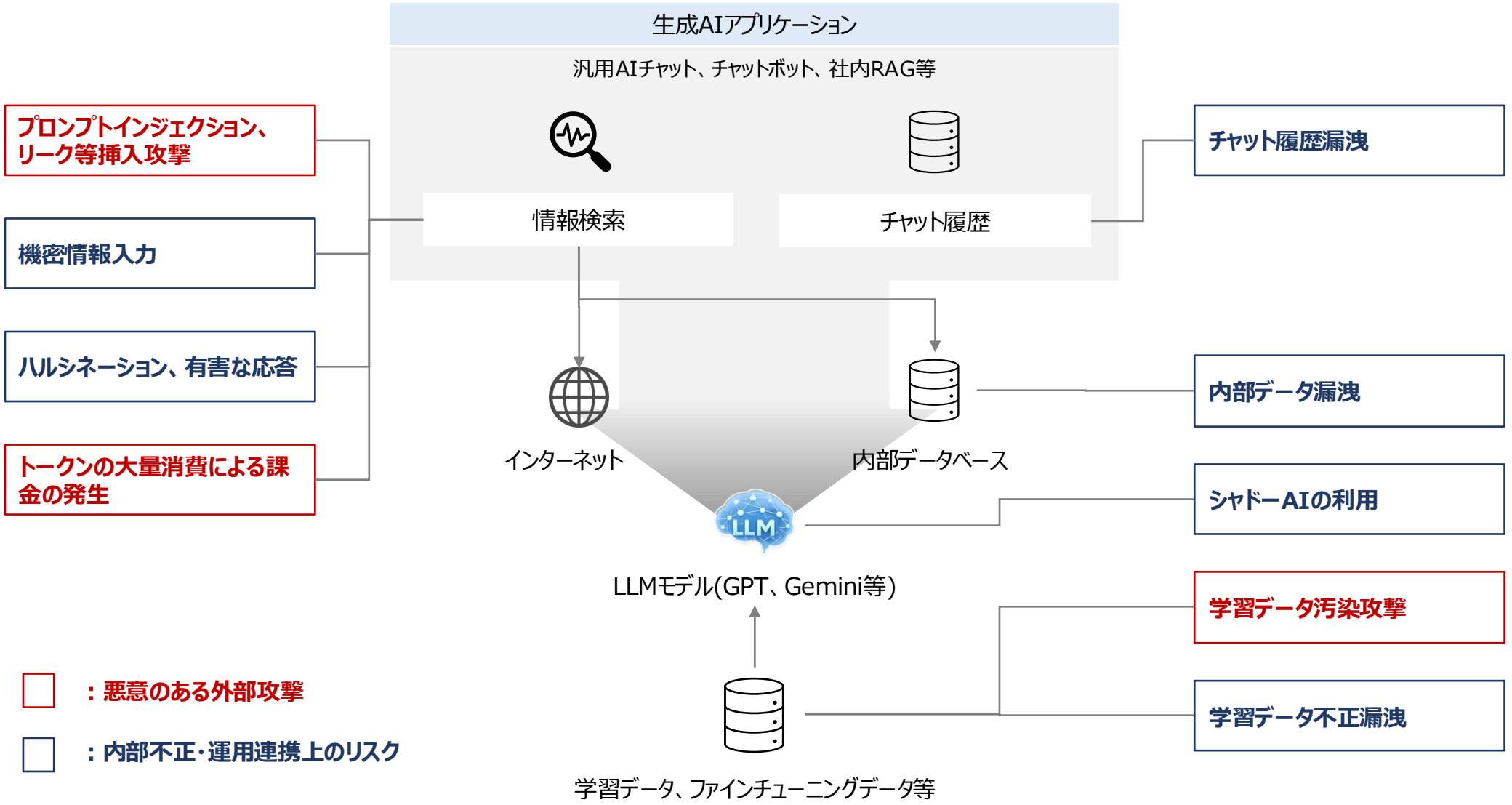
法令・ガイドライン・標準は出揃った。ポリシー1枚と承認ルールから着手する

あなたの会社のAI — 誰が・何に・どんなデータで使っているか、把握できていますか？

Librus AIセキュリティ関連取り組み

AIセキュリティリスク対応総合支援：概要

生成 AI の導入における主なセキュリティリスクは以下のとおりと想定しており、当社では、これらのリスクへの対応について総合的な支援が可能です。



AIセキュリティリスク対応総合支援：リスクアセスメント・ポリシー策定

ガイドラインベースおよびリスクシナリオベースのアプローチにより、AIに関連するリスクを体系的に評価し、適切なリスク対策を整理したうえで、AIポリシーの策定から継続的な運用までを一貫してご支援します。

サービスフロー

事前準備
<0.5~1ヶ月>

- **●対象AIスコープ定義**
  - 評価対象となるAIの種類、AIに関連する資産を定義
  - 評価対象関連資料の入手
- **●計画策定**
  - 実施スケジュール、評価アプローチ・基準ガイドラインの決定

評価実施
<1~3ヶ月>

- **●脅威・リスク特定**
  - 基準ガイドラインに沿って脅威・リスクを特定
  - 特定されたリスクの評価を実施
- **●対策・ロードマップ策定**
  - カードレール、入力制御等のリスク対応策と実施ロードマップを策定

報告
<0.5~1ヶ月>

- **●報告を実施**
  - 報告書の作成
  - 報告会の開催

ポリシー策定・運用支援

- **●AIポリシー策定・継続運用支援**
  - AI運用ポリシー策定支援
  - AI対策導入・推進支援
  - 継続的に組織・技術・人的対策支援

評価アプローチ

ガイドラインベース

ガイドライン	概要
NIST AI Risk Management Framework（AI RMF）	AIに関連するリスクを効果的に管理するためにNISTによって開発されたものであり、AIシステムに関連するリスクを企業が自律的に特定・評価・管理・対応するための、信頼性と安全性を高めるためのガイドラインとなる。「ガバナンス」「マップ」「測定」「マネジメント」に基づき、評価を実施する。
ISO/IEC 42001	2023年12月に発行され、AIに特化したマネジメントシステムの国際規格。AIを開発・提供・業務で利用する全ての組織に向けた規格であり、AIを安全かつ効果的に使うためのマネジメントシステムを作る際のルールが定められている。
経産省 AI事業者ガイドライン	経済産業省と総務省が共同で策定した、AI開発・提供・利用に関わる事業者向けの「日本におけるAIガバナンスの統一的な指針」

リスクシナリオベース

次のフレームワークを利用し、対象環境に対してリスクシナリオを作成し、影響評価を実施

フレームワーク	概要
MITRE ATLA（Adversarial Threat Landscape for Artificial-Intelligence Systems）	AIシステムに対する攻撃者の戦術と攻撃手法に関する、世界的にアクセス可能な最新のナレッジベースであり、現実に起こった攻撃の観察と、AIレッドチーミングやセキュリティグループ活動によるリアルなデモンストレーションに基づいたもの
OWASP Top 10 for LLM	大規模言語モデル（LLM）を利用するアプリケーションにおける、最も重要で一般的なセキュリティリスクを10項目にまとめたガイドライン。

AIセキュリティリスク対応総合支援：AIセキュリティ技術支援

AIシステムにおけるセキュアな構築、セキュリティ診断、セキュリティ対策等の技術的な支援を提供可能です。

セキュアなAIシステム構築支援 ▶ **AIシステムへのセキュリティ診断** ▶ **AIシステムにおけいセキュリティ対策構築支援**

セキュリティアーキテクチャ設計
- 企業データを隔離・保護するため、**閉域ネットワークや生成AI専用のネットワークセグメンテーション**を用いたAIシステム構成
- RAG 等で利用する各種データベースに対し、最小権限の原則に基づく**アクセス権限設計**を行う。

セキュアなAIモデル導入・設定
- 無料の生成AIサービスではなく、**エンタープライズ向けAI**モデルを採用
- モデルへのDoS攻撃を防ぐために、**利用レート制限**を適用

入力制御実装
- **プロンプトインジェクション**を防止するために、入力制御の実装

出力制御実装
- **機密情報検知、マスキング**等の制御禁止、禁止された情報や不適切な内容が出力されないようルールを適用

ログ監視仕組み構築
- 生成AI不正利用や情報漏洩のリスクを低減するために、**ログ取得**の仕組みを構築

弊社独自なツールを利用し、生成AIシステムへの**セキュリティ診断**を実施し、下記のようなAIセキュリティリスクを評価

※詳細は別紙参照

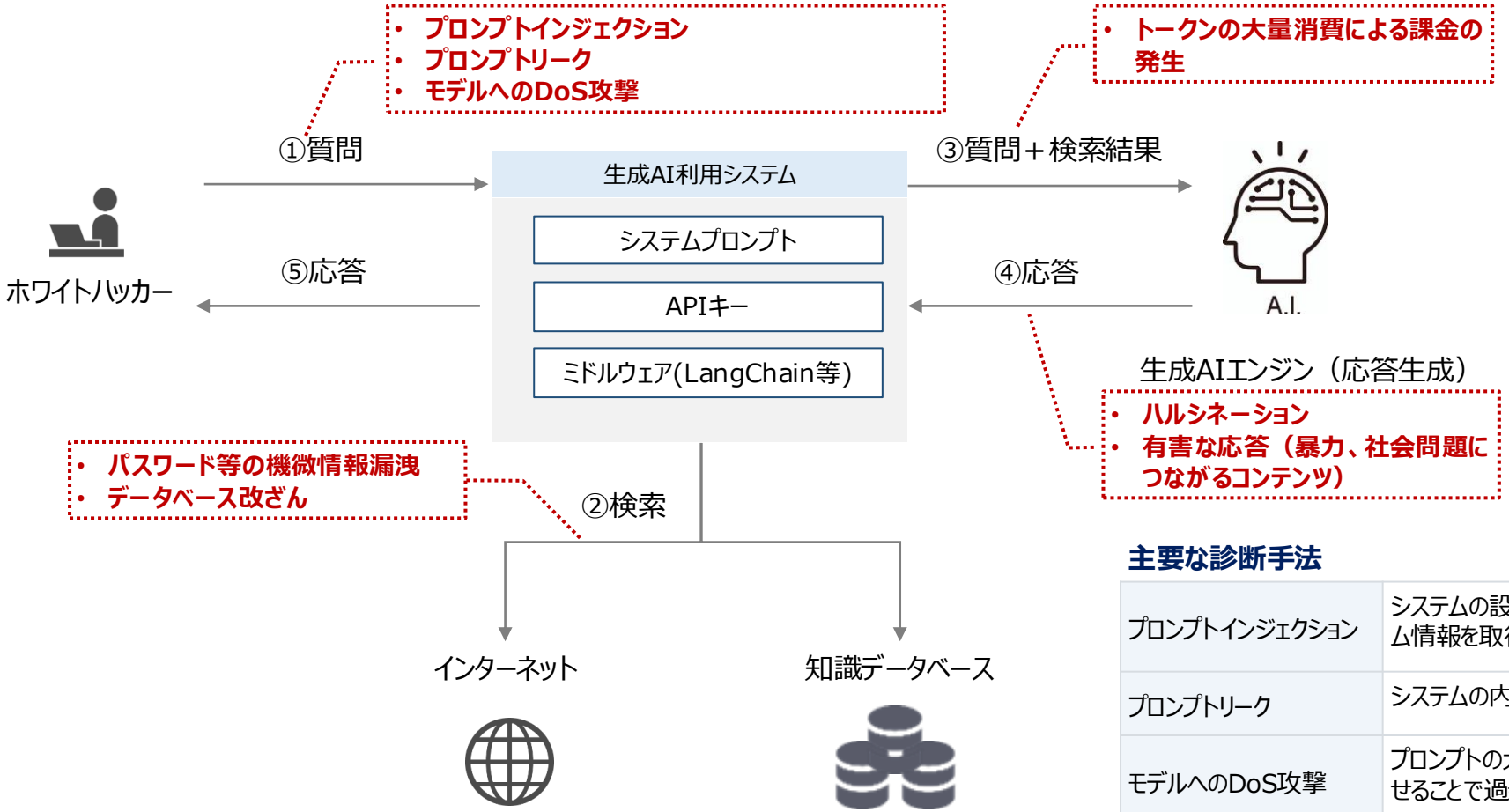
プロンプトインジェクション	システムの設定を回避して開発者が意図しない機密情報やシステム情報を取得する手法。
プロンプトリーク	システムの内部で設定されているプロンプトを出力する手法。
モデルへのDoS攻撃	プロンプトの大量入力によるサービス拒否や大量のトークンを消費させることで過剰な料金請求を発生させる手法

入力制御	**LLM Firewall・入力検証ソリューションの選定・導入支援** 代表ソリューション：Prompt Security、Protect AI、Guardrails AI
出力制御	**LLM出力制御ソリューションの選定・導入支援** 代表ソリューション：Prompt Security、Guardrails AI
データセキュリティ	**データセキュリティソリューションの選定・導入支援** 代表ソリューション：Azure Data Lake、Amazon Bedrock Knowledge Bases、Snowflake
ログ監視・異常検知	**ログ監視・異常検知ソリューションの選定・導入支援** 代表ソリューション：Splunk, Microsoft Sentinel,

生成AI　セキュリティ診断

生成AIを利用したシステムには様々なリスクが存在します。当社は攻撃者の視点から、AIによる情報漏洩や情報改ざんなどの特有のセキュリティ問題を特定するための診断サービスを提供しております。

生成AIを利用したシステムにおけるリスク例



主要な診断手法

プロンプトインジェクション	システムの設定を回避して開発者が意図しない機密情報やシステム情報を取得する手法。
プロンプトリーク	システムの内部で設定されているプロンプトを出力する手法。
モデルへのDoS攻撃	プロンプトの大量入力によるサービス拒否や大量のトークンを消費させることで過剰な料金請求を発生させる手法

生成AI　セキュリティ診断　観点・サンプル

当社の生成AIセキュリティ診断は、国際標準である「OWASP Top 10 for LLM」や、最新の攻撃手法・対策を取り扱った論文等をもとに、独自のサービスとして構築しております。診断観点および報告書サンプルは以下の通りです。

■ 診断の主要観点（Librus　Top10）

「OWASP Top 10 for LLM」や、最新の攻撃手法・対策を取り扱った論文等を踏まえ、弊社独自の診断観点として整理しております。

No.	リスク観点	概要
1	情報漏洩リスク	社内機密・顧客データ・個人情報・システム構成情報の漏洩
2	アクセス/認証リスク	API鍵盗難、権限管理不備、外部者の不正利用
3	プロンプト注入/外部操作リスク	意図しない指示強制、ガードレール突破、内部システム操作
4	システム設計上の安全性リスク	監査不足、ログ形骸化(ログの不正利用)、権限境界の欠陥
5	モデル汚染リスク	攻撃データ混入による性能劣化・バイアス誘導
6	法的・倫理的リスク	個人情報保護法、著作権、AI規制への抵触
7	サプライチェーン影響リスク	生成物が他システムへ連鎖し、被害・誤学習を誘発
8	製品品質/信頼性リスク	製品価値に関連するハルシネーション(事実の捏造)、不適切出力、誤誘導、ブランド毀損
9	ユーザー被害/誤誘導リスク	ユーザーが誤出力を信じることで被害が拡大
10	可用性・財務リスク	攻撃トラフィックによる利用量急増に起因する従量課金コストの異常増大および、サービス継続性を脅かす財務リスク

■ 報告書サンプル