生成系 AI のセキュリティに関する 4 つの主要な質問への回答

セキュリティ、プライバシー、コンプライアンスの確保を支援しながら、
生成系 AI を迅速に導入する

この日本語ガイドは、自社の組織に生成系 AI を安全に導入する方法を計画または検討しているビジネスリーダー、特に IT 意思決定者やセキュリティチームリーダーを対象としています。

目次

はじめに

準備、設定、生成：生成系 AI を迅速かつ安全に導入

生成系 AI の競争は始まっています。企業は生産性とエクスペリエンスを大きく向上できる可能性から、カスタマーエクスペリエンスとアプリケーションの改革に突き進んでいます。

生成系 AI の時代はまだ始まったばかりですが、組織は既にあらゆる事業部門で目に見えた成果を実現しています。しかし、セキュリティの専門家からは注意が喚起されています。セキュリティの専門家は、生成系 AI の導入に慎重に取り組むべき理由として、データのプライバシー、モデルのバイアス、有害なコンテンツの生成 (ディープフェイクなど)、モデルへの悪意のある入力によるリスクを挙げています。

組織は自社のデータ、ユーザー、信用をどう保護するのか明確な戦略を立てて生成系 AI に取り組むことが不可欠です。同時に、迅速な導入とカスタマーエクスペリエンスの向上を実現する必要があります。

これには多くの課題がありますが、AI、機械学習 (ML)、データ保護、クラウドワークロードセキュリティに関する標準的なベストプラクティスが適用されることには変わりはないということを認識しておく必要があります。実際、お客様の組織は生成系 AI の安全を確保する準備が予想以上に整っているかもしれません。

今、生成系 AI のワークロードを適切に保護する方法を確立すれば、組織全体のイノベーション促進につながるでしょう。そしてチームは大きな構想を追求する自信を持ち、ビジネスの成長に集中できるようになります。

この日本語ガイドでは、より安全な生成系 AI ワークロードの実現に向けて踏み出す際に、問うべき重要な 4 つの質問について解説します。

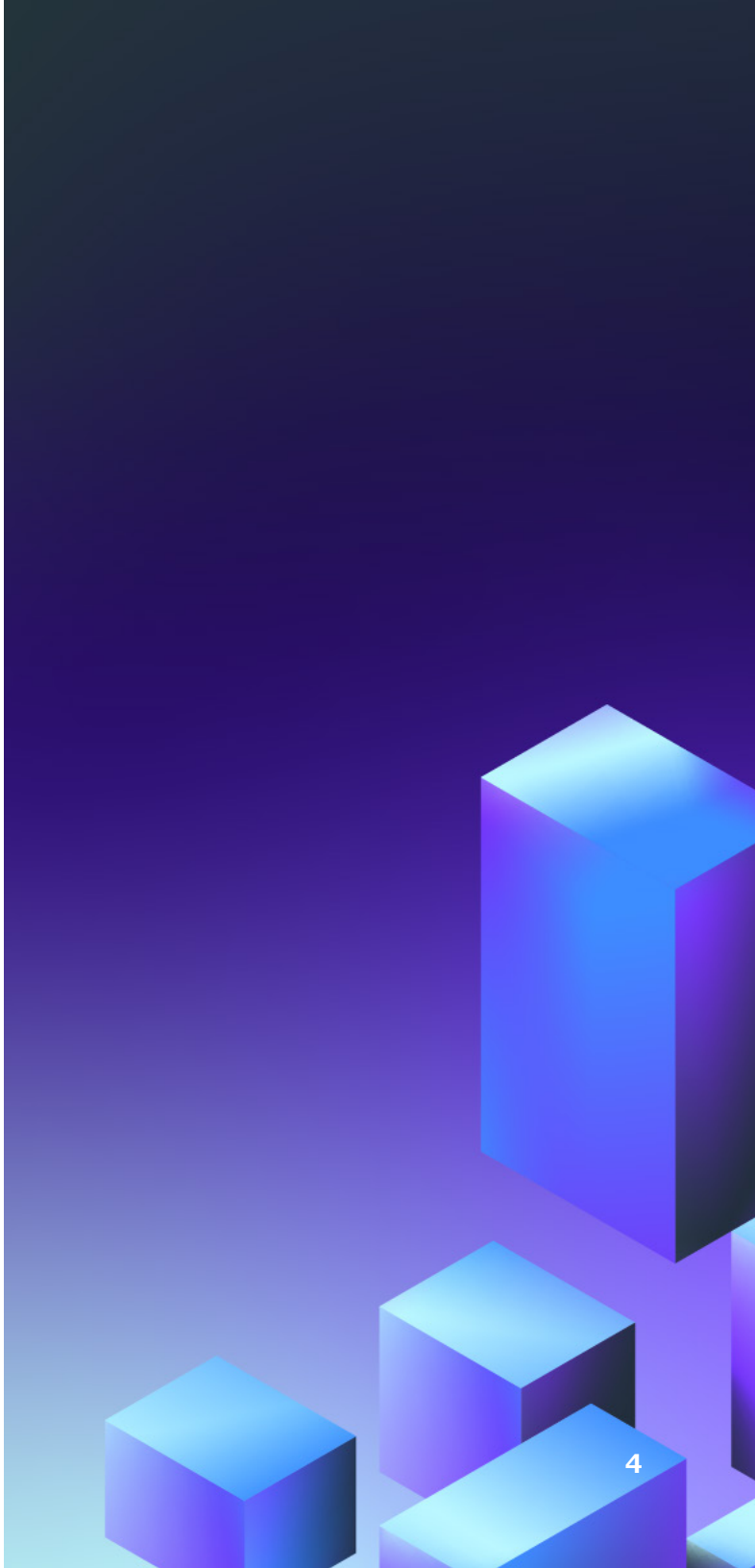
1. **何を保護するべきか**
2. **コンプライアンスに関する懸念にどう対処するのか**
3. **モデルが意図したとおりに動作しているかをどう確認するのか**
4. **何から始めるべきか**

質問 1:

何を保護するべきか

生成系 AI アプリケーションを安全に開発してデプロイするには、何を保護する必要があるのかを正確に理解しておくことが重要です。この作業は以下の 3 つのカテゴリに分類するとよいでしょう。

- **クラウドワークロードの保護**
- **データの保護**
- **生成系 AI アプリケーションの保護**



クラウドワークロードの保護

生成系 AI をセキュリティとプライバシーの目標を満たしつつ使用するには、クラウドのインフラストラクチャ、サービス、設定の全体を保護することから始めます。それにはまず、セキュリティの責任の中で自社が担う部分とクラウドプロバイダーが担う部分とを区別する必要があります。

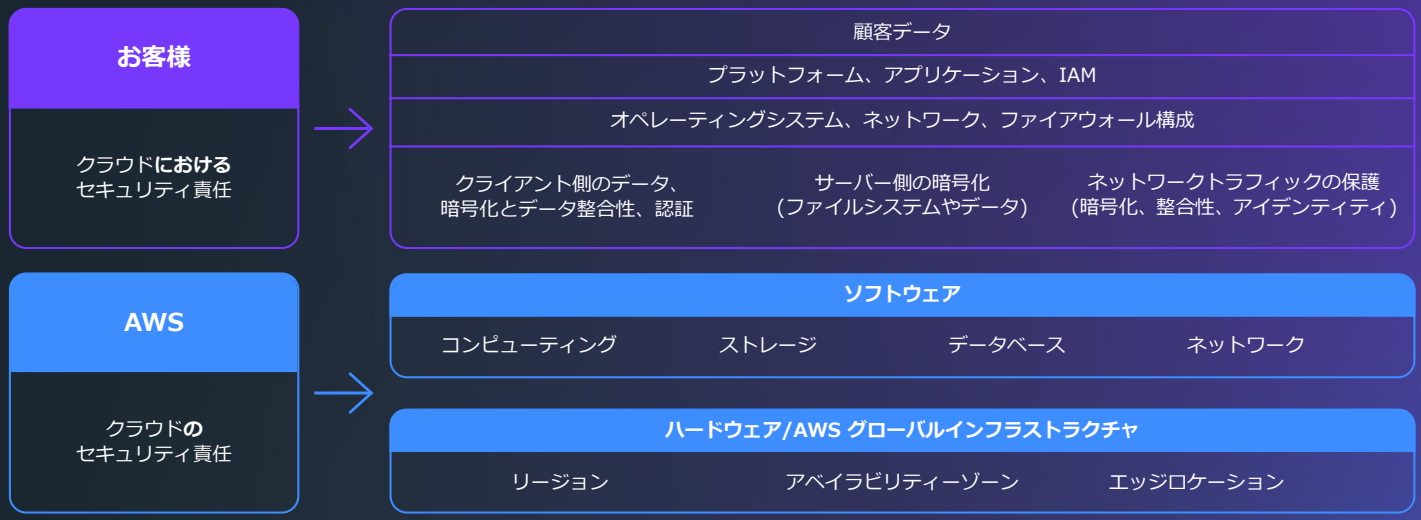
Amazon Web Services (AWS) をご利用のお客様は、この分野のガイダンスを**責任共有モデル**でご覧いただけます。大まかに説明すると、AWS クラウドで提供されるすべてのサービスを実行するインフラストラクチャの運用、管理、制御は、AWS が責任を持つというものです。これは、**「クラウドのセキュリティ」**と呼ばれています。

一方 AWS のお客様は、ゲストオペレーティングシステムの管理 (更新やセキュリティパッチなど)、その他の関連アプリケーションソフトウェアの管理、AWS より提供されるセキュリティグループファイアウォールの構成に責任を持ちます。

お客様の責任範囲と具体的な義務は、お客様が使用する AWS サービスによって異なります。これは**「クラウドにおけるセキュリティ」**と呼ばれます。

生成系 AI は新たに注目されている分野ですが、従来のセキュリティベストプラクティスがこれまで同様出発点として役立ちます。これには、次のようなセキュリティハイジーンの取り組みが含まれます。

- Identity and Access Management (IAM)
- 侵入検知とレスポンス
- インフラストラクチャ保護
- データ保護
- アプリケーションのセキュリティ



データの保護

次に、生成系 AI アプリケーションで使用されるデータのセキュリティとプライバシーを確保する必要があります。これには、機密情報、貴重な知的財産 (IP)、および個人を特定できる情報 (PII) が含まれる場合があります。

生成系 AI アプリケーションは、膨大な量のデータに基づいてトレーニングされた基盤モデル (FM) によって稼働しています。FM はこのデータを分析してパターンを特定し、新しく類似コンテンツを生成する方法を学習します。特定のビジネス要件を満たす生成系 AI アプリケーションを構築するには、通常、既存の FM をお客様のデータでトレーニングしカスタマイズしなくてはなりません。

このデータを保護するには、データプライバシー管理と IAM ポリシーのベストプラクティスを検討するとよいでしょう。

FM をカスタマイズするときは、そのモデルのバージョンが安全に保存されており、なおかつ FM 自体を改善する際に使用されているものではないことをチーム内でよく確認してください。**Amazon Bedrock** でシングルテナントの専用容量を設定すると、推論用インスタンスが **Amazon Virtual Private Cloud** (Amazon VPC) に割り当てられ、**Amazon Simple Storage Service** (Amazon S3) への読み取りと書き込みが行えます。

IAM を効果的に設定することで、適切なリソースに適切なユーザーとマシンが適切な条件下でアクセスしているかを検証できます。**AWS Well-Architected フレームワーク**には、設計原則とアーキテクチャのベストプラクティスが記載されており、アイデンティティの管理に役立ちます。このリソースは IAM ポリシーの開発に役立つツールであるだけでなく、脅威検出やネットワークセキュリティなど、セキュリティ上の懸念事項への対処にも有用です。

生成系 AI アプリケーションの保護

アプリケーションレベルで生成系 AI を保護するには、リスクを継続的に特定、分類、修正、軽減する必要があります。最初のステップは、環境とデータを安全に保つための既存のベストプラクティスを実装することです。

そこから、セキュリティを移行する方法を開発プロセスの初期段階で検討していくのです。こうすることで取り組みを効率化でき、開発チームはセキュリティをボトルネックにすることなく、より迅速かつ自由にイノベーションを進めることができます。

次に、どのような AI アプリケーションにも重要な 3 つの要素 (入力、出力、モデル) を保護する方法を検討します。

入力の保護

まず、AI システムに入力するデータを確認します。改ざん、なりすまし、プロンプトインジェクションなどのインテグリティ攻撃のリスクを減らすため、ユーザーが FM に直接アクセスする際は、必ず入力フィルタリングを行わなければなりません。ここで挙げた攻撃手法は、統制をすり抜けたり、モデルを悪用したりします。これ以外にも、入力を保護するために検討すべき戦略として、データ品質の自動化、継続的な監視、脅威モデリングがあります。

出力の保護

生成系 AI アプリケーションの出力でのリスクには、情報漏えい、知的財産インシデント、さらに、組織の信用を損なうおそれのあるモデルの誤用や悪用などがあります。脅威モデルを開発する際には、情報のフットプリントと使用されるコンテキストを考慮し、複合的な行動検出と監視も組み込みましょう。

モデルの保護

最後に、攻撃者がどうやってモデルそのものやモデルに関連するコンポーネントからデータを削除しようとするのかを考えてみましょう。リスクには、現実世界の歪曲やモデルに入力されたデータの改ざん、モデルの完全性や可用性の侵害があります。お客様のビジネス目標を狙う脅威をモデル化し、こうした脅威シナリオに対する監視を実装しましょう。

コンプライアンス要件

質問 2:

コンプライアンスに関する懸念にどう対処するのか

生成系 AI アプリケーションの設計と開発に伴うリスクを軽減すれば、パートナーや顧客との信頼関係を構築し、ブランドの信用を維持して継続的にコンプライアンス要件に対応できます。

生成系 AI アプリケーションの法的規制はまだ初期段階にあり、ベストプラクティスについてのコンセンサスは確立されていません。そのため、さまざまな地域で基準や管理が相反するという混乱した状況を乗り越えることが、複雑で継続的な課題となっています。

法務アドバイザーやプライバシーの専門家に相談し、生成系 AI アプリケーションを構築する際の要件と影響を評価してください。これには、特定のデータやモデルを使用する法的権利の検証、プライバシー、生体認証、差別禁止、その他ユースケース固有の規制に関する法律の適用可否の判断が含まれる場合があります。

法的要件は州、県、国によって異なり、さらに新しい AI 規制が世界中で提案され続けているということを留意してください。これらの考慮事項は将来のデプロイや運用段階で再度見直す必要があります。

仲間や AI のエキスパート、政府機関と連携することで、お客様の組織が AI の法的および倫理的基準に真摯に対応していることを顧客に示すことができ、同時にコンプライアンスの維持にもなります。最近、Amazon はホワイトハウスと大手 AI 企業 6 社とともに、**責任ある安全な AI 開発への自発的な取り組み**を開始しました。Amazon はこの取り組みの価値を実証しながら、さらに将来のコラボレーションの基礎を築いています。

AI に内在するリスク

機械学習を使用するあらゆるソリューションと同様、生成系 AI アプリケーションには従来のソフトウェアを超えるリスクがあります。生成系 AI を使用してアプリケーションを安全に構築してデプロイするには、次のようなリスクを軽減するための戦略を練り、策定する必要があります。

- バイアス、虚偽、誤解、有害性、攻撃性のある出力
- 規模による複雑性とコスト
- 増大しすぎ、もしくは古くなったデータセットや、意図するコンテキストと関係しないデータセット
- 拡大する不透明性と再現性に対する懸念
- 未整備のテスト基準とテスト手順

次のセクションでは、このようなリスクの一部を軽減するための幅広い戦略を見ていきます。生成系 AI アプリケーションが与える専門的、組織的、社会的な影響を定義するためのベストプラクティスについて説明します。

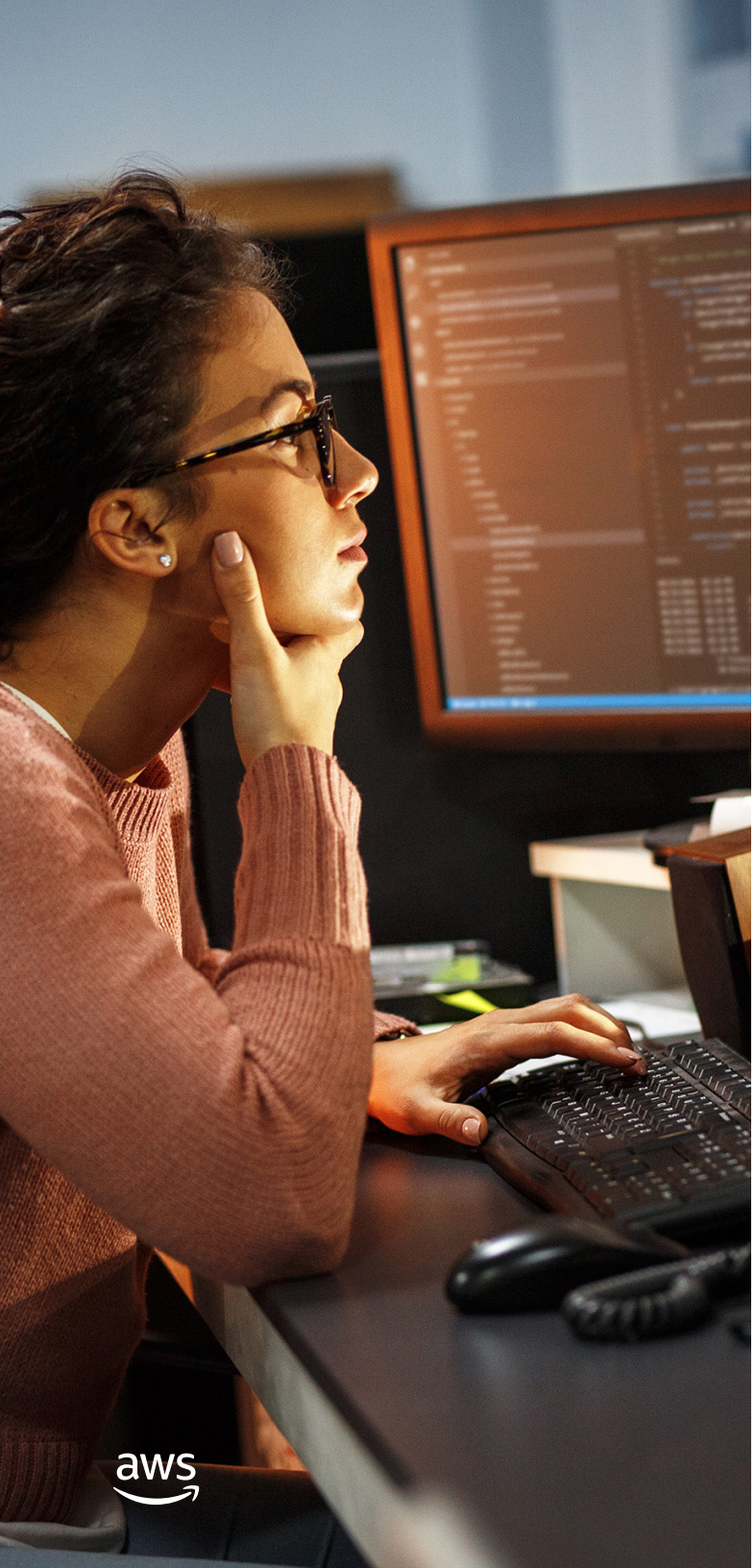
質問 3:

モデルが意図したとおりに動作しているかをどう確認するのか

責任ある生成系 AI の使用を徹底することは、基本的なビジネスタスクとしてだけでなく、継続的なイノベーションを実現するための重要な手段としても重視されるようになってきました。

FM は膨大なデータセットでトレーニングされ、類似するコンテンツの生成方法を理解するために複雑な分析を行います。多くの FM が目覚ましい成果をもたらしていますが、「ゴミを入れたら、ゴミが出てくる」(GIGO) という古くからの格言が今でも当てはまります。FM に不正確、不完全、またはバイアスのあるデータを入力した場合、FM が出す結果にも同様の欠陥が生じるおそれがあります。

欠陥のあるデータは、誤用、悪意のある行為、その他のリスクを生む可能性があります。生成系 AI アプリケーションのユーザー、適用範囲、機能が拡大するほど、これらの問題がもたらす潜在的な影響も大きくなります。



責任ある AI を促進する

責任ある AI の戦略を取ることで、こうしたリスクに対処できます。責任ある AI には、説明可能性、公平性、ガバナンス、プライバシー、セキュリティ、堅牢性、透明性などの側面があります。また、多種多様な文化や人口統計がアプリケーションによってどのように捉えられ、取り扱われ、影響を与えるのかを理解することも含まれます。

生成系 AI 導入の初期段階から責任ある AI について考え、さらにアプリケーションのライフサイクル全体を通じて、ビジョンの重要な部分として取り組み続けることが最善です。比較的小さな、簡単なアクションから始めましょう。次に、責任ある AI が長期的に設計、開発、運用にどのような影響を与えるかを評価します。

責任ある AI とガバナンスポリシーを策定する際には、お客様の生成系 AI アプリケーションがユーザー、顧客、従業員、社会にどのような影響を与えるのかを考慮してください。アルゴリズムの公平性、多様で包括的な表現、バイアスの検出にも必ず対応しましょう。

有害性に取り組む

大規模言語モデル (LLM) における有害性とは、不快、侮辱的、不合理なテキストの生成を指します。生成系 AI アプリケーションの有害性を減らし、公平性を確保するための戦略は数多くあります。例えば、トレーニングデータから攻撃的な言葉やバイアスのある表現を見つけ出し、削除することもできます。また、アプリケーションの特定のユースケースや、対象ユーザー、最も使われるプロンプトやクエリに焦点を当て、より絞り込んだ公平性テストを実施することもできます。

さらに、異なる種類や程度の有害性を識別する注釈付きのデータセットで、ガードレールモデルをトレーニングすることもできます。これによって FM は、トレーニングデータ、入力プロンプト、生成した出力から好ましくないコンテンツを自動的に検出してフィルタリングする方法を学習できます。

プライバシーを保護する

生成系 AI アプリケーションを扱う際に、機密情報、企業秘密、知的財産が意図せず公開されるのを防ぐために講じることができる手段がいくつかあります。

モデルの削除は、プライバシーの問題に対処する方法の 1 つです。不適切に使用されたデータを見つけしだいすぐに削除して、そのデータが FM のどのコンポーネントにも影響を与えないようにします。

もう 1 つのアプローチはシャーディングです。シャーディングでは、トレーニングデータを小さく分割し、そのうえで個別のサブモデルをトレーニングして、最終的にサブモデルを組み合わせて FM 全体を形成します。この方法により、非公開の情報を流出している、または流出する危険性のある FM をはるかに簡単に修復できます。モデル全体を再トレーニングするのではなく、不要なデータや不適切に使用されたデータをシャードから削除して、そのサブモデルを再トレーニングするだけで済むからです。

フィルタリングとブロッキングも効果的なアプローチです。これらの手法では、生成されたコンテンツをユーザーに表示する前に、保護された情報と明確に比較を行います。2 つの内容があまりにも似ている場合は、公開しないようコンテンツが非表示になるか、置換されます。また、特定のコンテンツについてトレーニングデータに登場する回数を制限することも有益であることが立証されています。

説明可能性と可監査性を強化する

責任ある AI をさらに補強するには、アプリケーションの出力に影響している方法論と主要な要素を説明する必要があることも考えておきましょう。可監査性もまた、責任ある AI の重要な要素です。生成系 AI アプリケーションの開発と運用を追跡してレビューできるメカニズムを実装しましょう。問題の根本原因を突き止め、ガバナンス要件を満たすのに役立ちます。

開発ライフサイクル全体を通じて、関連する設計上の決定や意見を文書化することも検討してください。追跡可能な記録を作成することにより、社内外のチームが生成系 AI アプリケーションの開発と機能を評価しやすくなります。

責任を果たす

最後に、責任ある AI ポリシーを継続的に順守するために役立つ方法を考えてみましょう。学んだ教訓と得た経験を活かして、セキュリティとプライバシーの取り組みを強化します。安全で安心な生成系 AI を実践するうえで守るべき義務について、組織内のすべての従業員に定期的な教育を行ってください。責任ある AI の文化を育み、適切なツールでモデルのパフォーマンスを監視して、リスクを発見し、チームが必要に応じてモデルとそのコンポーネントを検査できるようにします。テストにテストを重ね、疑わしい場合は再度テストしてください。

開始方法

質問 4:

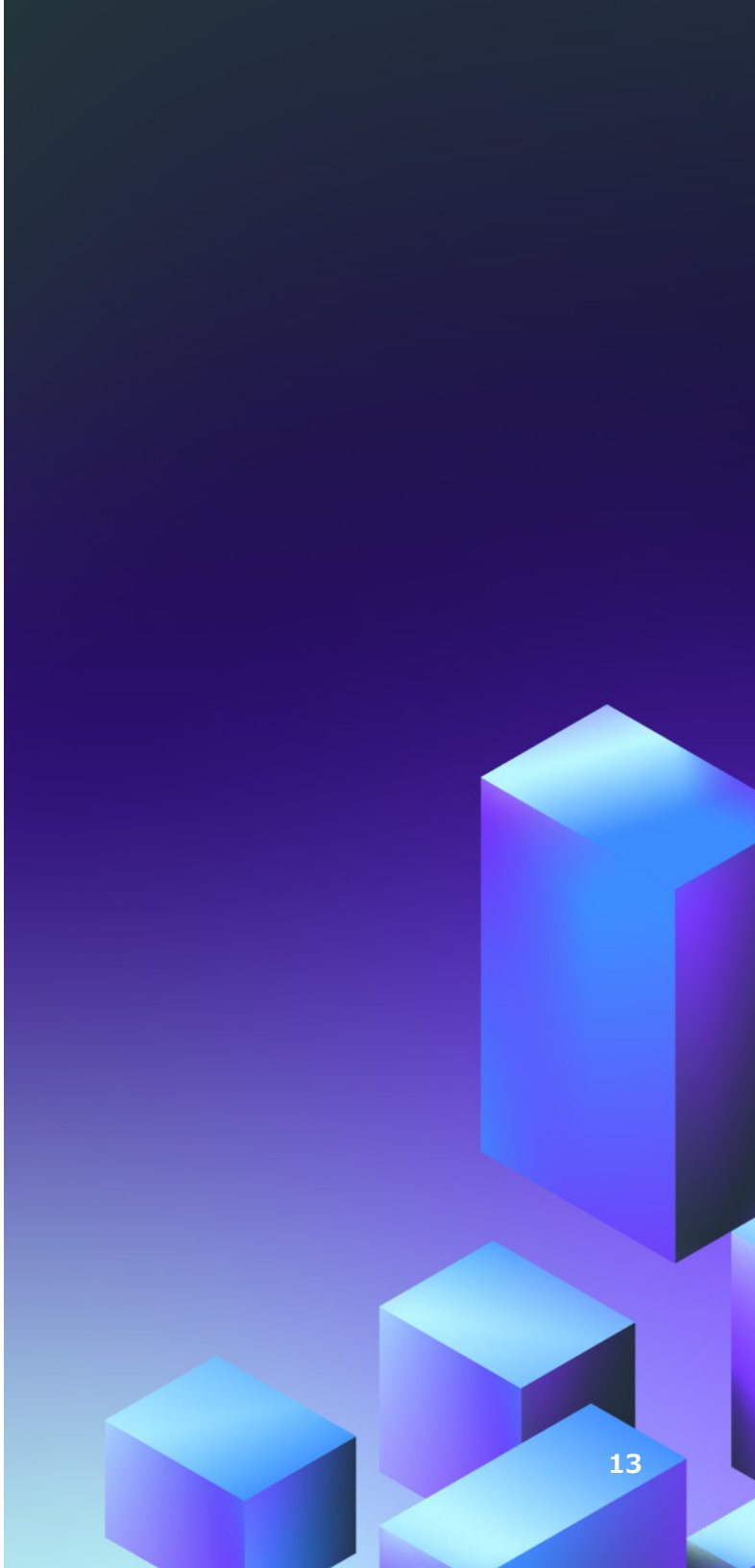
何から始めるべきか

生成系 AI アプリケーションのセキュリティを確保することは簡単ではなく、その実現に向けて実行できる万能なアクションはありません。しかし、適切なベンダーと協力して適切なツールをデプロイすれば、成功への道すじはより明確なものになります。

例えば、**Amazon Bedrock** を使用すると、安全な生成系 AI アプリケーションの開発プロセスが大幅に簡素化され、加速されます。Amazon Bedrock は、Amazon や主要な AI スタートアップによる FM を API で利用できるフルマネージドサービスです。

Amazon Bedrock でモデルをカスタマイズすると、特定のタスクに合わせてモデルのファインチューニングができ、大量のデータのアノテーションは必要ありません。そして Amazon Bedrock は、元の FM からお客様だけがアクセスできるコピーを作成し、この非公開のモデルコピーをトレーニングします。お客様のデータが、元の基本モデルのトレーニングに使用されることはないため、独自データを非公開かつ安全に保つことができます。

また、**Amazon VPC** を設定して Amazon Bedrock API にアクセスできるようにし、ファインチューニングしたデータを安全な方法でモデルに提供できます。お客様のデータは、転送中も保存中も、サービス管理キーによって常に暗号化されます。さらに、**AWS PrivateLink** を使えば、AWS ネットワークだけを使用して AWS クラウドデータを Amazon Bedrock に送れます。パブリックインターネットを経由することはありません。



AWS でプライバシーを強化する

お客様が生成系 AI アプリケーションの構築に Amazon Bedrock やその他のサービス (**Amazon SageMaker** など) を利用している場合でも、もしくは独自のツールを使用している場合でも、AWS でアプリケーションの実行と管理を行えば、業界トップレベルのプライバシー保護と制御が可能になります。

AWS は 143 のセキュリティ標準とコンプライアンス認証に対応しており、各要件を満たすよう世界中のお客様をサポートしています。すべてのデータは、お客様専用の **AWS Key Management Service** (Amazon KMS) キーを使用して保存時に暗号化できるため、データと FM の保存方法とアクセス方法を完全に制御し、可視化できます。

まとめ

次のステップ

AWS は、ビジネスを成長させる生成系 AI アプリケーションの構築をお手伝いし、さらにお客様がセキュリティ、プライバシー、コンプライアンスの目標を達成できるよう全力で支援します。

AWS は、生成系 AI アプリケーションは安全に設計、開発、運用できるという強い信念を持っています。また、これらの技術に関するセキュリティとプライバシーに関する懸念の妥当性も認識しています。データプライバシー、知的財産、立法監督、平等、透明性に関する問題の定義、測定、軽減において、**生成系 AI は新たな課題をもたらしています**。

新しい製品の登場、ソリューションの複雑化と大規模化、新しいトレーニングパラメータ、増大を続けるデータセットにより、生成系 AI のセキュリティは今後さらに不可欠なものになるでしょう。生成系 AI ワークロード向けの効果的かつ包括的なセキュリティ戦略を今すぐ策定すれば、競争上の優位性を最大限に高め、急速に迫り来る未来に備えることができます。

さいわいにも、生成系 AI アプリケーションを安全に設計、開発、実行するために必要な基本的な統制は以前から導入されており、**AWS Well-Architected フレームワーク**に見られるような、信頼性が高く実証済みのクラウドセキュリティの原則に則っています。

生成系 AI ワークロードの保護に向けた第一歩を、この日本語ガイドに概説されているプラクティスを知るところから踏み出してください。

AWS で次のステップに進みましょう。当社は、お客様のデータ、顧客、ビジネスを保護するとともに、新たに登場するテーマに対応できるよう深いインサイトと具体的なガイダンスを提供し、お客様の固有の課題から考えを深め、生成系 AI の利点を最大限に引き出します。

AWS の生成系 AI について詳しく知る ›

Amazon Bedrock で今すぐ始める ›

Amazon SageMaker で FM を構築、カスタマイズする ›

AWS でクラウドのセキュリティを向上する ›

責任ある AI を理論から実践に移す ›