AWS Black Belt Online Seminar

Amazon Bedrock Series #01

Amazon Bedrock Overview

森下　裕介

Solutions Architect

2024/06

内容についての注意点

- 本資料では資料作成時点（2024 年 6 月下旬）のサービス内容および価格についてご説明しています。AWS のサービスは常にアップデートを続けているため、最新の情報は AWS 公式ウェブサイト (https://aws.amazon.com/) にてご確認ください
- 資料作成には十分注意しておりますが、資料内の価格と AWS 公式ウェブサイト記載の価格に相違があった場合、AWS 公式ウェブサイトの価格を優先とさせていただきます
- 価格は税抜表記となっています。日本居住者のお客様には別途消費税をご請求させていただきます
- 技術的な内容に関しましては、有料の AWS サポート窓口へお問い合わせください
- 料金面でのお問い合わせに関しましては、カスタマーサポート窓口へ
お問い合わせください (マネジメントコンソールへのログインが必要です)

森下 裕介 Morishita Yusuke

アマゾンウェブサービスジャパン合同会社
ソリューションアーキテクト

製造業のお客様を中心に
クラウドの技術支援を担当しています。

好きな AWS サービス：
- Amazon Bedrock
- Amazon Rekognition

苦手なもの：
- ハルシネーション
- ミニトマト

本動画の対象者とゴール

- 主な対象者
    - これから AWS で 生成 AI の活用を始めたいと考えている方
    - Amazon Bedrock のことを初めて学ぶ方
    - Amazon Bedrock がもつ各種機能について知りたい方

- 本動画のゴール：

  Amazon Bedrock の Overview について理解する！

  Amazon Bedrock の個別機能の詳細については後続の Blackbelt でご紹介予定です。

アジェンダ

- 生成 AI 活用をはじめる上での難しさ
- Amazon Bedrock の基本
- 生成 AI アプリ開発のよくある課題と Amazon Bedrock の各種機能
- まとめ

生成 AI 活用をはじめる上での難しさ

生成 AI とは

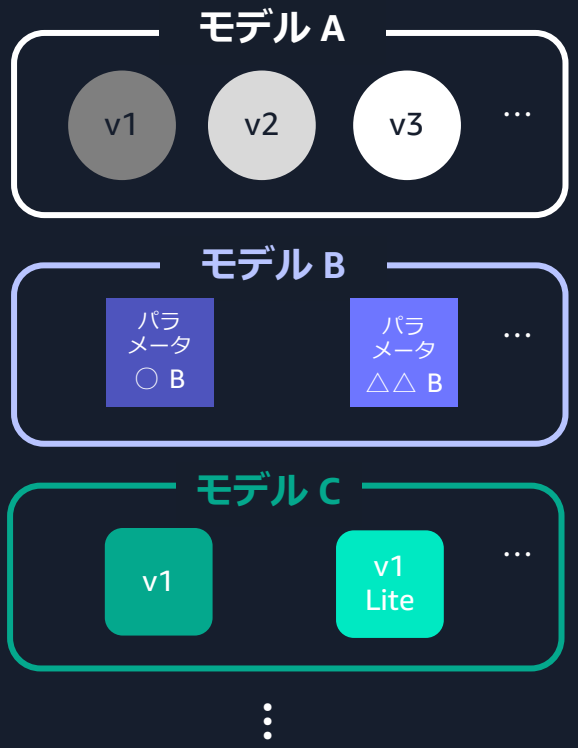
会話、ストーリー、画像、動画、音楽など、
新しいコンテンツやアイデアを創造

基盤モデルと一般的に呼ばれる、膨大なデータで
あらかじめ学習された**大規模なモデルを原動力**とする

生成 AI 活用をはじめる上での難しさ

生成 AI の進化への追従

複数のモデルを簡単に
利用できる環境の構築



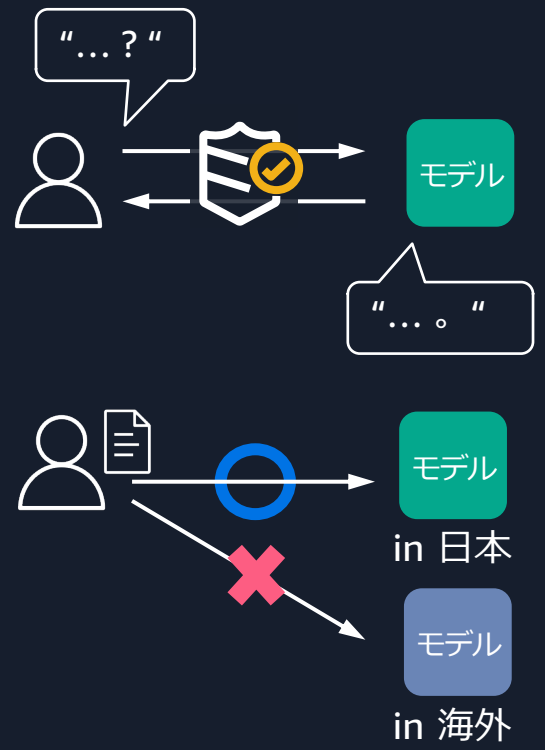
インフラストラクチャ

モデルの実行を支える
インフラの管理

必要なこと
- 各モデルごとの環境整備と機械学習の知識
- モデルのデプロイやカスタマイズのための大規模な計算リソース
- インフラの維持管理
- 柔軟なスケーリング
- …

プライバシーとセキュリティ

データ漏洩の防止、
国内にデータを留めたい



生成 AI 活用をはじめる上での難しさ

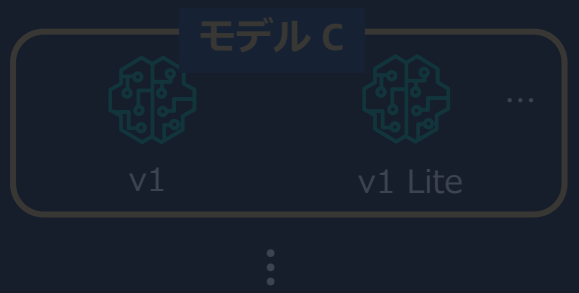
生成 AI の進化への追従

インフラストラクチャ

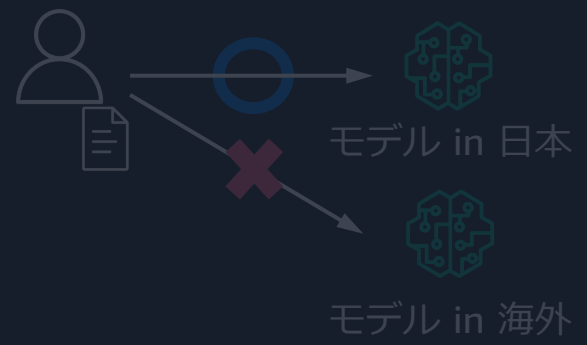
プライバシーと セキュリティ

それ、 Amazon Bedrock で解決できます！

パラメータ○B　パラメータ△B



- 各モデルごとの環境構築と機械学習の知識
- モデルのデプロイやカスタマイズのための大規模な計算リソース
- インフラの維持管理
- 柔軟なスケーリング
- …



Amazon Bedrock

基盤モデルを活用した
生成 AI アプリケーションを
簡単に構築、拡張できる方法

API を介してさまざまな基盤モデルにアクセス、
インフラ管理は不要

お客様の業務用途に適した基盤モデルを選択
Amazon, AI21 Labs, Anthropic, Cohere,
Meta, Mistral, Stability AI, …

データセキュリティやコンプライアンスを実現

エージェント機能、RAG 機能、非公開でのモデル
のカスタマイズなど基盤モデルの効果を高める
さまざまな機能を提供

東京リージョン含む国内外の AWS リージョンで一般提供中

Amazon Bedrock の基本

Amazon Bedrock

2024年6月時点

幅広い基盤モデルの選択肢をご提供

AI21 labs | ANTHROP\C | MISTRAL AI_ | co:here | Meta AI | stability.ai | amazon

JURASSIC

テキスト

Jamba-Instruct
Jurassic-2 Ultra
Jurassic-2 Mid

CLAUDE

テキスト & ビジョン

Claude 3.5 Sonnet
Claude 3 Opus
Claude 3 Sonnet
Claude 3 Haiku

テキスト

Claude 2.1
Claude 2.0
Claude Instant

MISTRAL & MIXTRAL

テキスト

Mistral Large
Mistral Small
Mistral 7B
Mixtral 8X7B

COMMAND & EMBED

テキスト

Command R+
Command R
Command
Command Light

埋め込み

Embed - Multiligual
Embed - English

LLAMA

テキスト

Llama 3 70B
Llama 3 8B
Llama 2 70B
Llama 2 13B

SDXL

画像

Stable Diffusion XL 1.0

AMAZON TITAN

テキスト

Titan Text Premier
Titan Text Express
Titan Text Lite

画像

Titan Image Generator

埋め込み

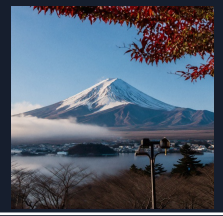
Titan Multimodal Embeddings
Titan Text Embeddings V2
Titan Text Embeddings

aws

基盤モデルの種類

2024 年 6 月時点

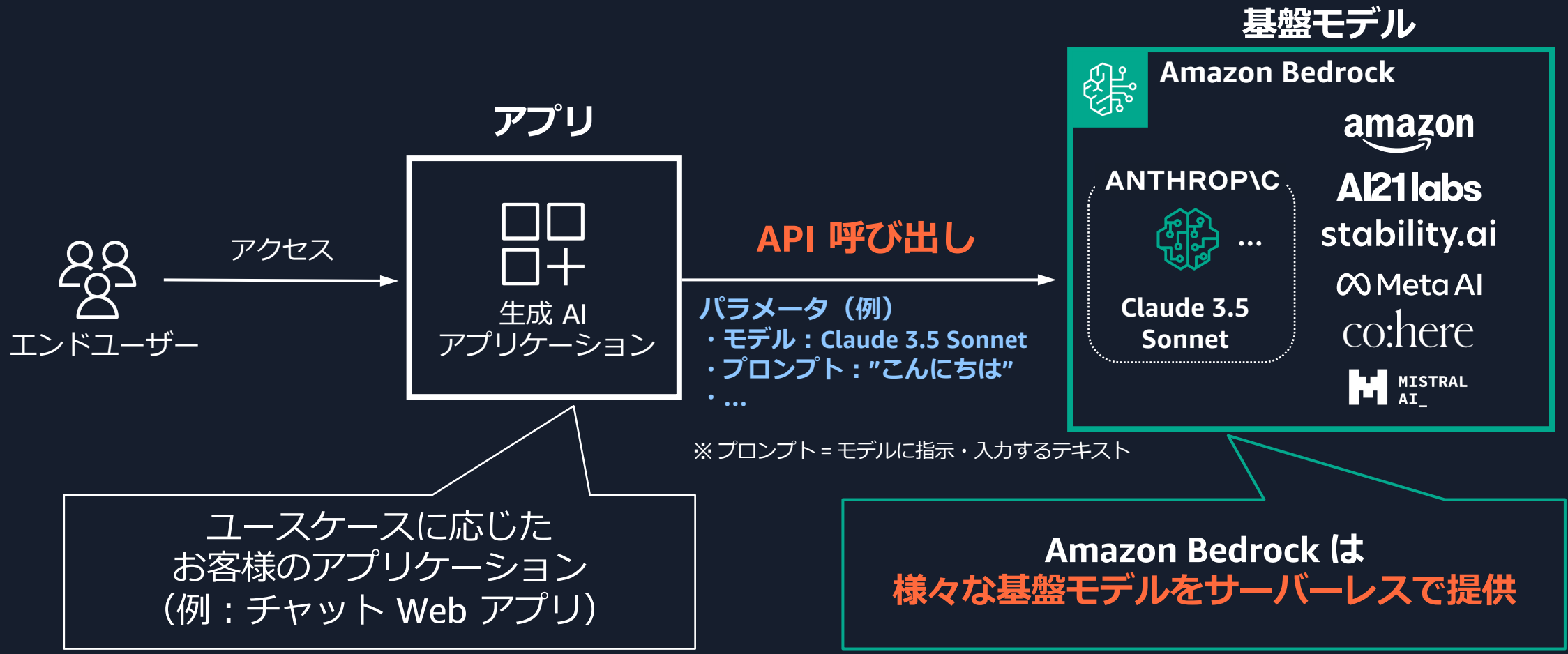
様々なモダリティのデータを扱える基盤モデルの種類を提供

入力（プロンプト）	基盤モデル	出力	
“日本で 1 番高い山は？”	テキスト	“富士山です。...”	入力テキストに対応するテキストを生成
 + “画像には何が写っていますか？”	テキスト & ビジョン	“これは富士山です。...”	テキスト入力だけでなく画像も入力可能なモデル
“富士山”	画像	 by Amazon Titan Image Generator	入力テキストに対応する画像を生成
 or “富士山”	埋め込み	[-0.00268, -0.03386,...]	入力を数値表現に変換検索用途などで利用

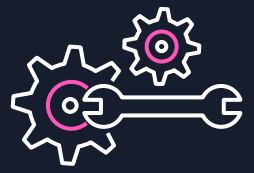
モデルのモダリティ定義は 2024 年 6 月時点の Model Access ページの表記に準拠

Amazon Bedrock の利用イメージ

基盤モデルをサーバーレスで提供しており
アプリケーションから API を通じて利用



データセキュリティ・コンプライアンス



お客様のデータが
モデルの学習や
AWS およびサードパーティーの
モデルプロバイダーに
共有されることは無い

モデル入出力内容に対し
人間による検閲が
なされることは無い

お客様のデータはすべて作成
されたリージョンに留まる

AWS PrivateLink により
Amazon VPC と
Amazon Bedrock 間の
プライベート接続を実現

全ての転送・保管されるデータ
は常に暗号化

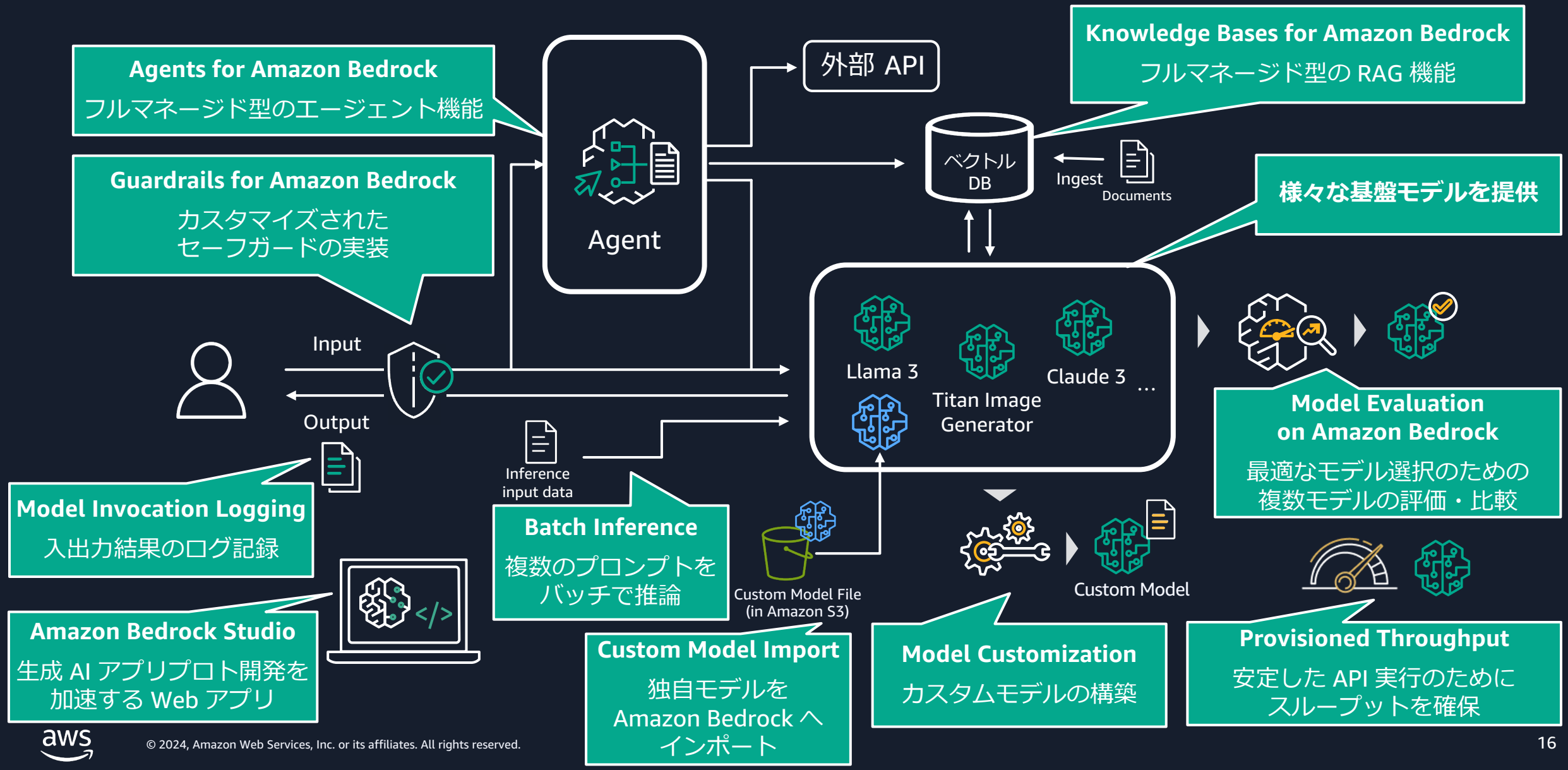
GDPR (一般データ保護規則)

HIPPA コンプライアンス等
標準規格に準拠

ISMAP の言明対象範囲に
登録

Amazon Bedrock の主な機能の全体像

2024 年 6 月 ver.



生成 AI アプリ開発をサポートする Amazon Bedrock の各種機能

生成 AI アプリ開発でよくある課題

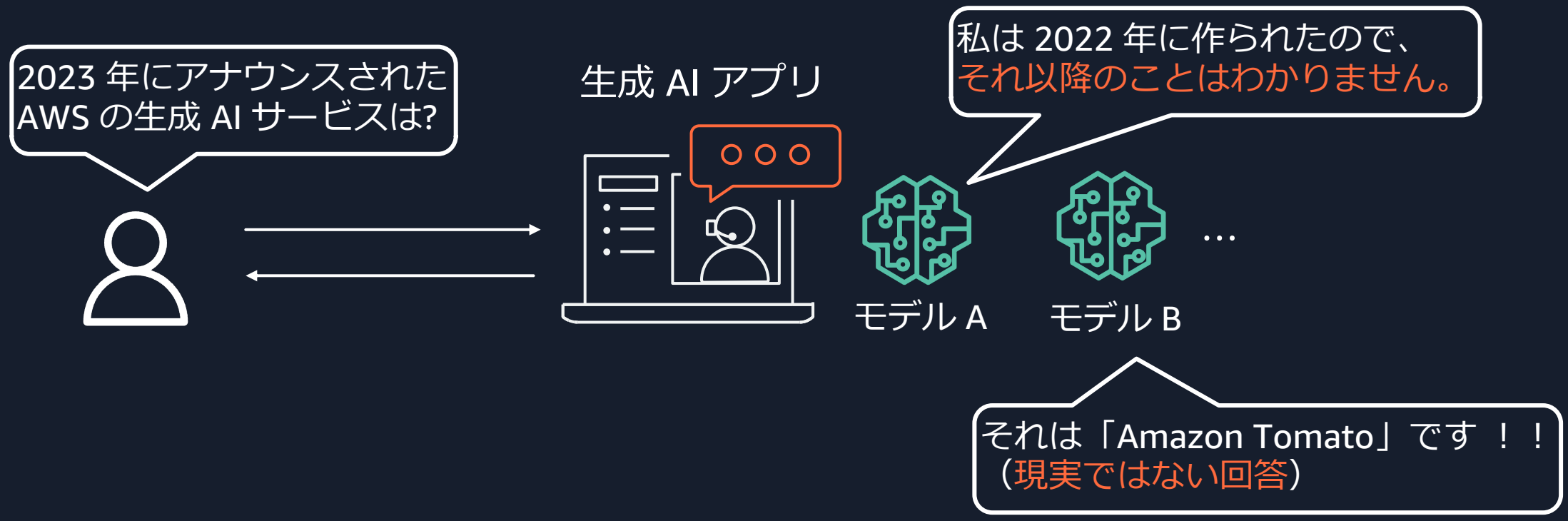
- 課題 1 ：独自データの活用
- 課題 2 ：データや外部 API との連携の実現
- 課題 3 ：基盤モデルの評価と最適なモデルの選択
- 課題 4 ：生成 AI アプリの適切な利用状況の維持と安定的稼働
- 課題 5 ： 生成 AI 活用のアイデアの迅速な実験と評価

生成 AI アプリ開発でよくある課題

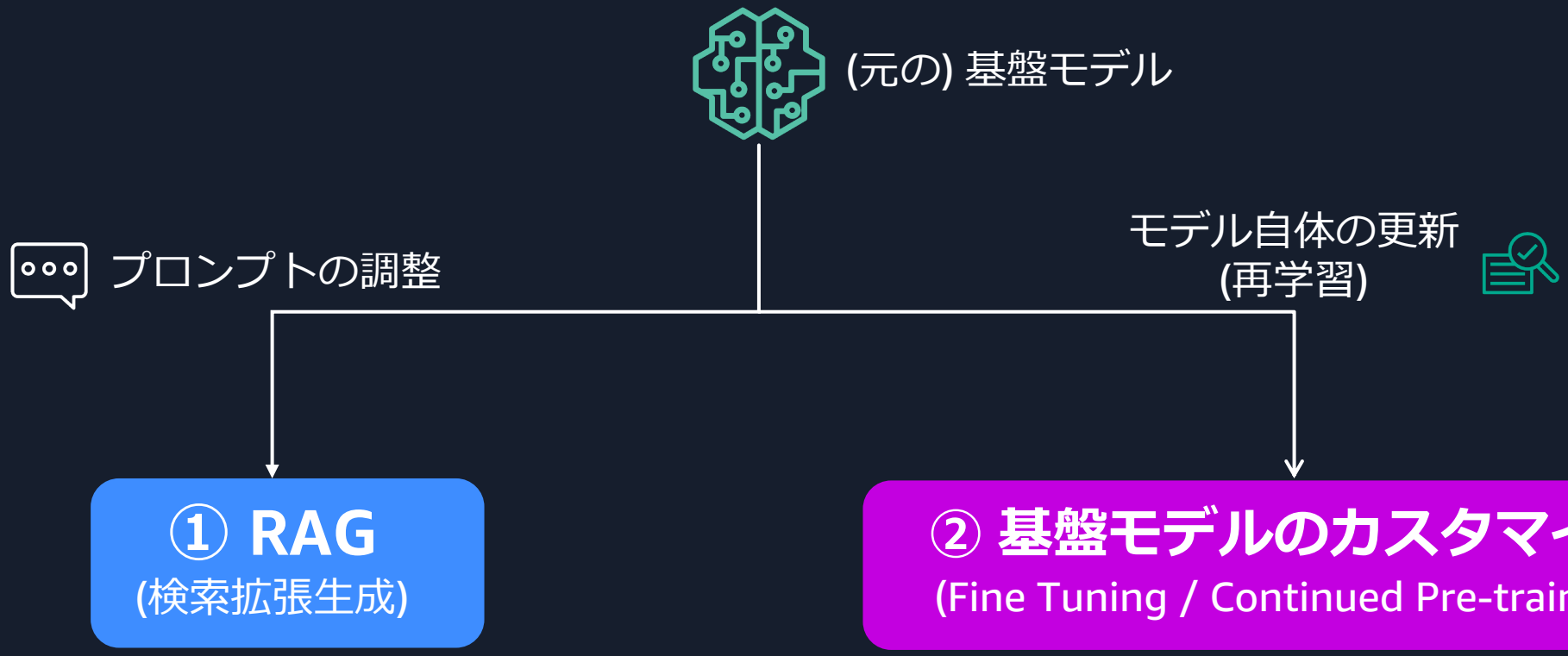
- 課題 1 ：独自データの活用
- 課題 2 ：データや外部 API との連携の実現
- 課題 3 ：基盤モデルの評価と最適なモデルの選択
- 課題 4 ：生成 AI アプリの適切な利用状況の維持と安定的稼働
- 課題 5 ： 生成 AI 活用のアイデアの迅速な実験と評価

よくある課題 1：
独自データの活用

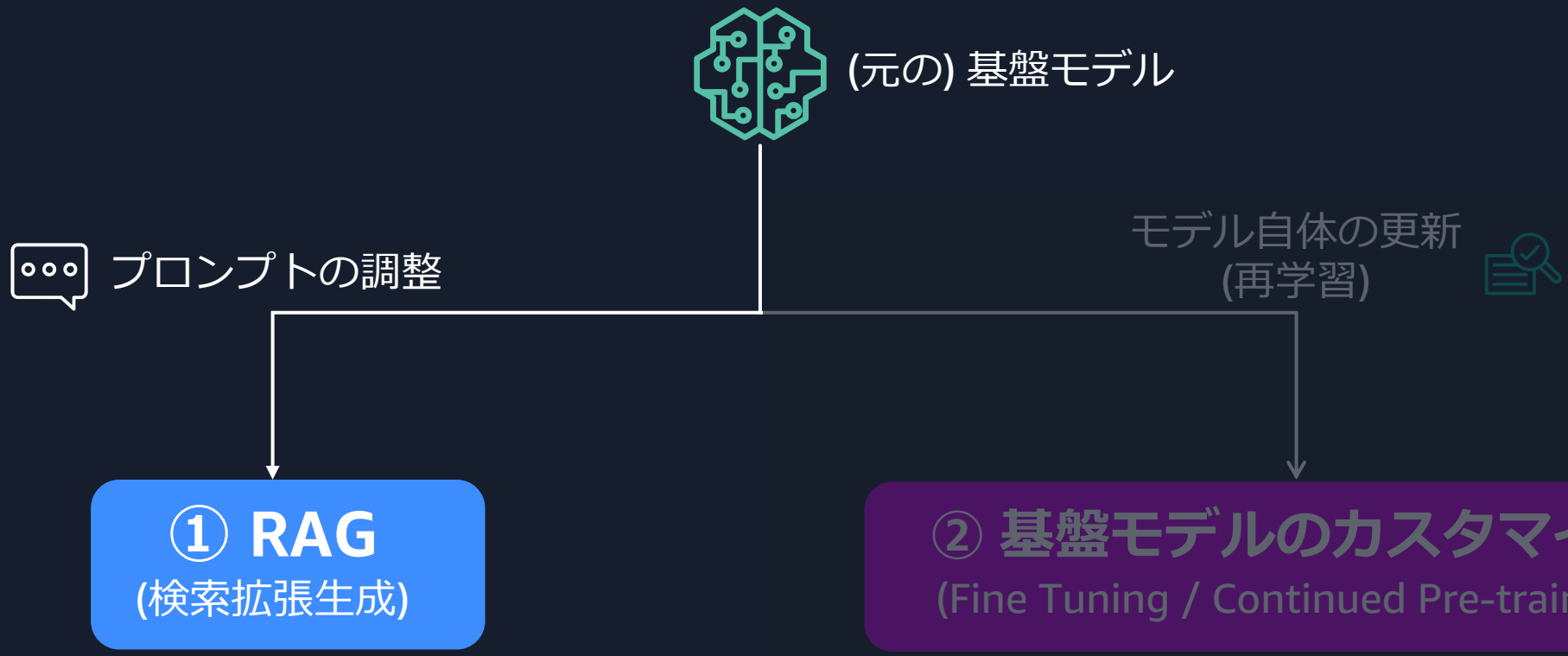
基盤モデルは学習データに含まれていないような情報は回答できない。
場合によっては不正確な回答をしてしまう（ハルシネーション）。



独自データを基盤モデルに活用する方針



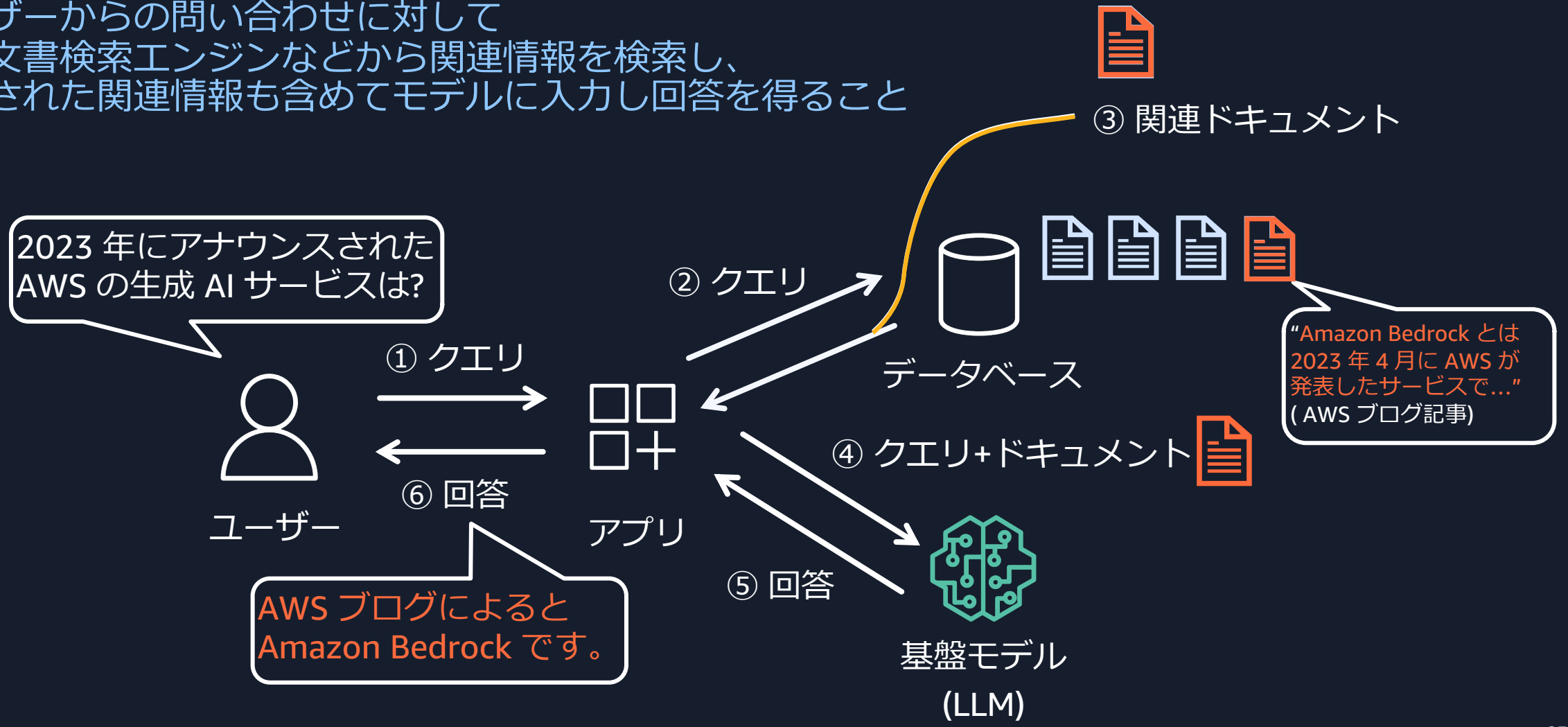
独自データを基盤モデルに活用する方針



① RAG とは

Retrieval-Augmented Generation (検索拡張生成)

ユーザーからの問い合わせに対して
社内文書検索エンジンなどから関連情報を検索し、
取得された関連情報も含めてモデルに入力し回答を得ること

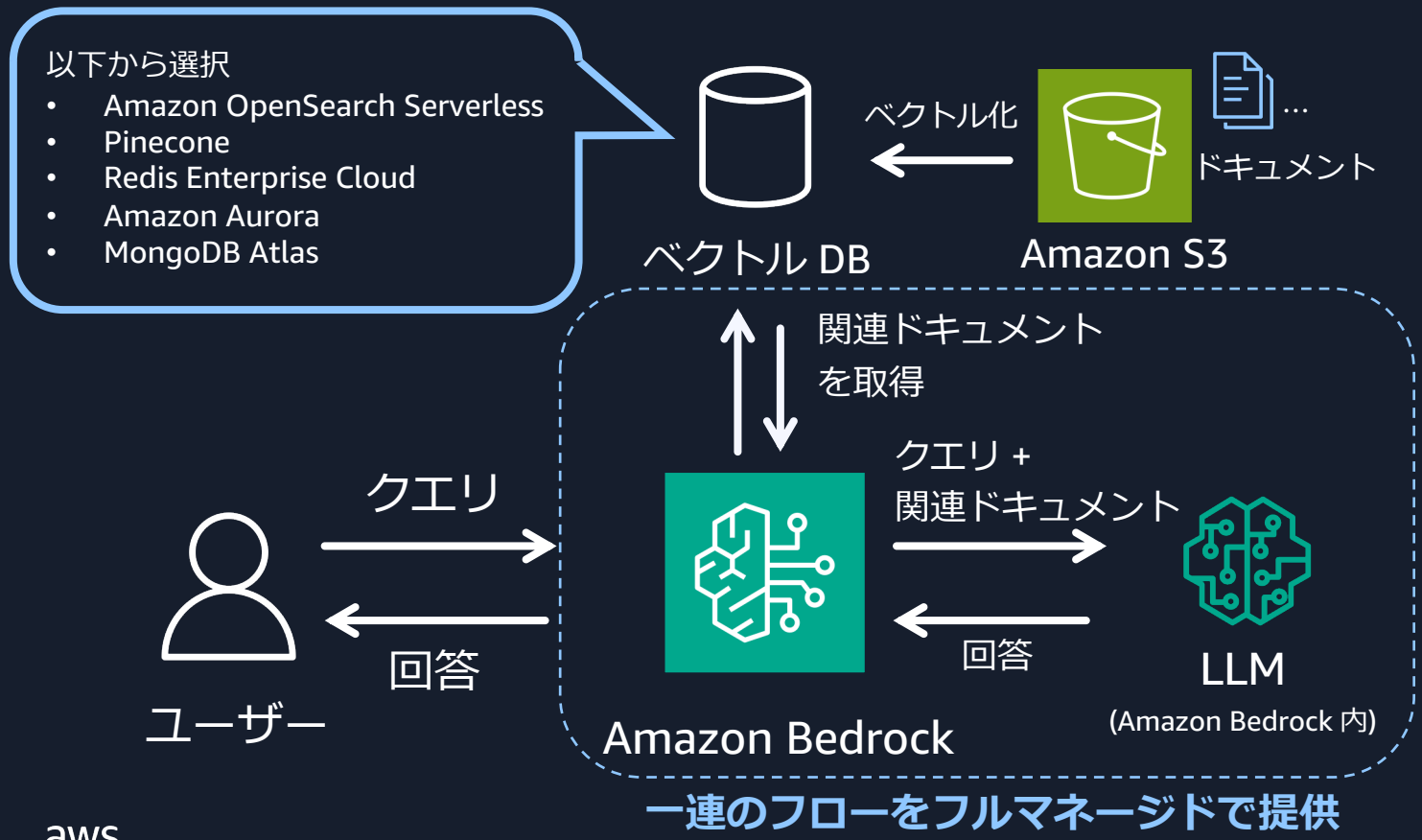


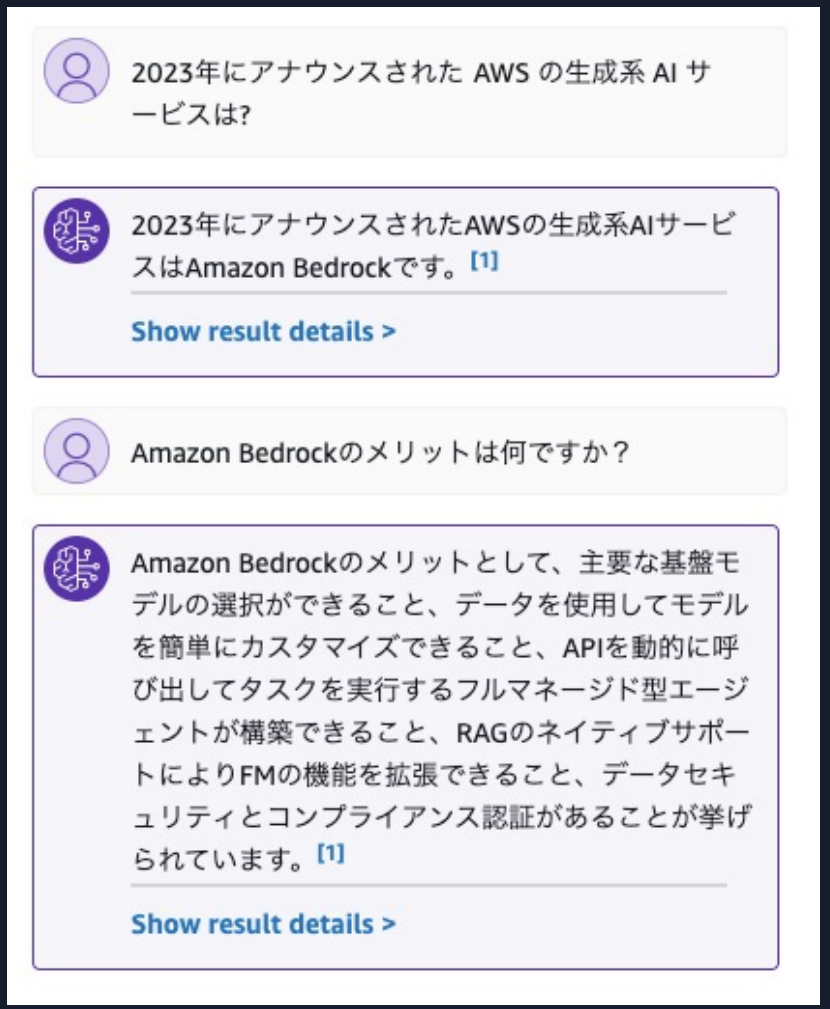
※ LLM = 大規模言語モデル

① Knowledge Bases for Amazon Bedrock

フルマネージド型の RAG 機能を提供

基盤モデルと自社データソースを組み合わせた
RAG (検索拡張生成)をフルマネージドに実現可能に



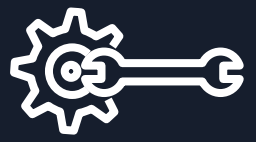


一連のフローをフルマネージドで提供

独自データを基盤モデルに活用する方針

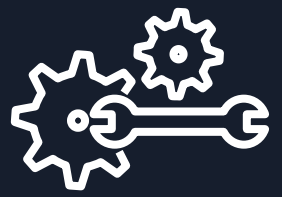
② 基盤モデルのカスタマイズ

自身のデータセットを学習して基盤モデル自体をカスタマイズする機能



Fine Tuning

ラベル付きデータを新たに学習して
特定のタスクの精度を高める手法



Continued Pre-training

大量のラベル無しデータを用いて
モデルに新たなドメイン知識を習得させる手法

再学習用のコード不要でマネージドな環境でモデルのカスタマイズを実行

Custom Model Import for Amazon Bedrock

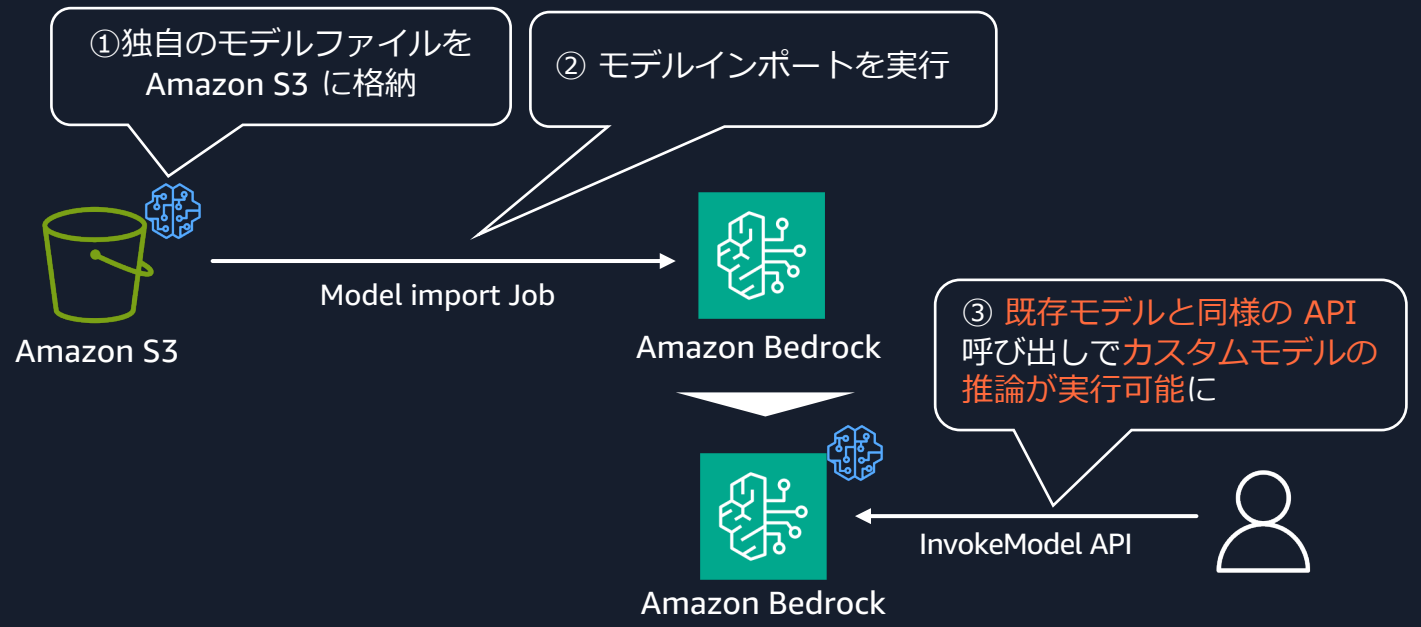
Amazon Bedrock とは他の環境でトレーニングされたカスタムモデルを
Amazon Bedrock にインポートすることが可能に

Amazon SageMaker やローカル環境のような別の環境で学習された基盤モデルをインポートし
Amazon Bedrock の既存モデルと同様の API 呼び出しで推論が実行可能に

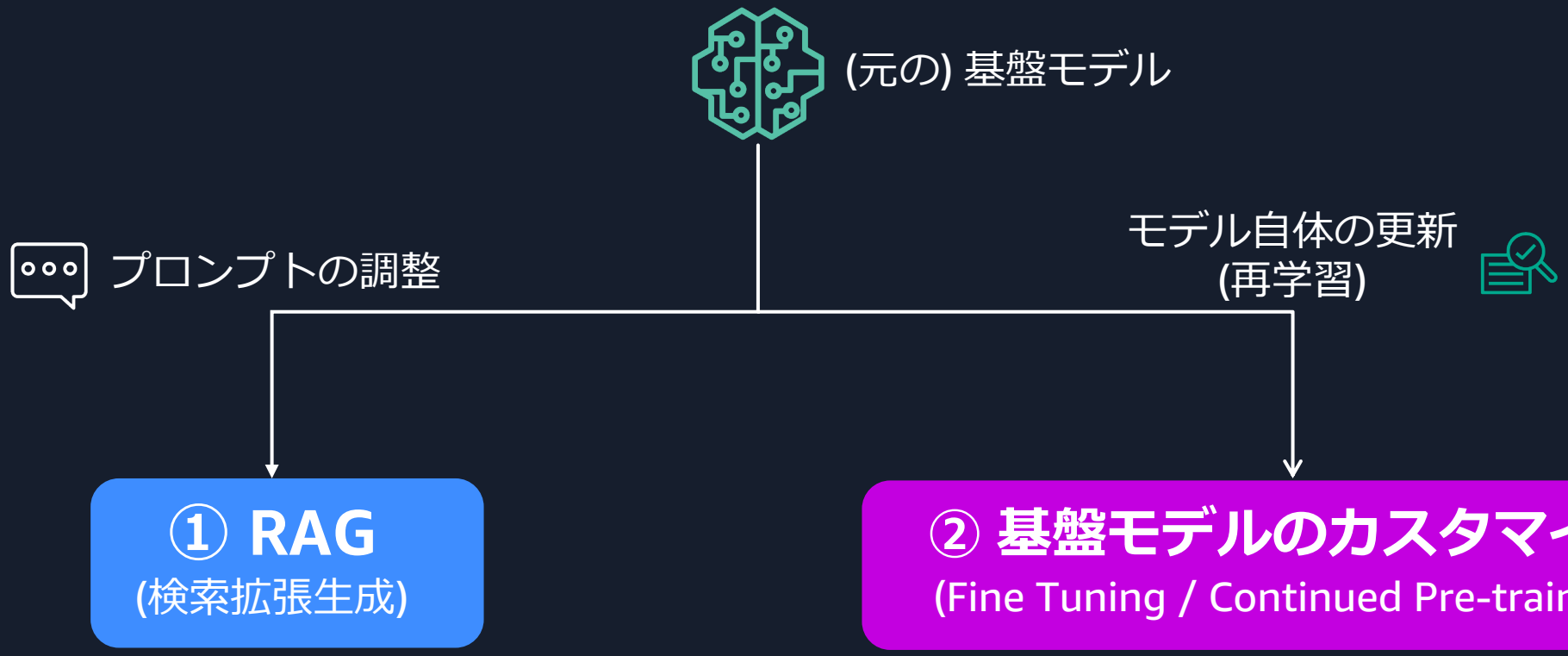
On-demand 方式で API 呼び出し可能

以下のアーキテクチャに対応 (2024 年 6 月時点)

- Mistral
- Flan
- Llama 2 & Llama 3



独自データを基盤モデルに活用する方針



使い分け (例)

- 最新情報を出力に反映させたい場合
- プロンプト内で与えられた知識のみで十分対応可能な場合

- 未知のドメイン知識の追加や出力スタイルの制御など、モデルの振る舞いをカスタマイズしたい場合

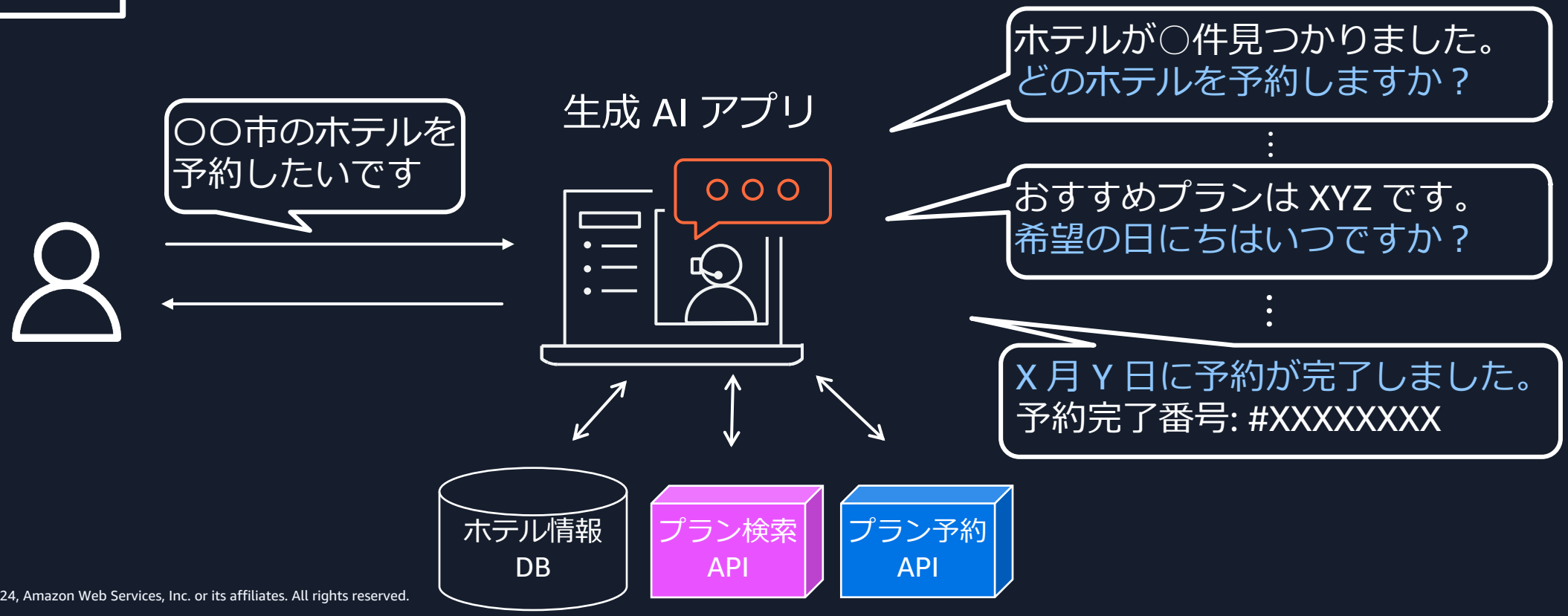
生成 AI アプリ開発でよくある課題

- 課題 1 ：独自データの活用
- 課題 2 ：データや外部 API との連携の実現
- 課題 3 ：基盤モデルの評価と最適なモデルの選択
- 課題 4 ：生成 AI アプリの適切な利用状況の維持と安定的稼働
- 課題 5 ： 生成 AI 活用のアイデアの迅速な実験と評価

よくある課題２：
データや外部 API との連携の実現

基盤モデル単体では外部との連携が必要なタスクの実行は難しい。
自身のデータソースの活用や外部 API 連携と組み合わせてタスクの完了までを実現したい。

実現イメージ



Agent とは？

ユーザーの入力を複数の小さなタスクに分割し、
タスクごとに適切な API を呼び出すことで回答を生成させるアプローチ

ユーザー

〇〇市のホテルを予約したいです。

Agent (LLM)

まずはデータベースから〇〇市のホテルを探そう。
各ホテルにはおすすめのプランもあるみたいだ。
これも追加の情報として提示しよう。
まずはどのホテルに宿泊したいかを確認しないと。

Agent とは？

ユーザーの入力を複数の小さなタスクに分割し、
タスクごとに適切な API を呼び出すことで回答を生成させるアプローチ

ユーザー

〇〇市のホテルを予約したいです。

Agent (LLM)

〇〇市のホテルは 8 件見つかりました。
ホテル A は朝食付きプラン、ホテル B は素泊まりプランがおすすめです。ホテル C は…

どちらのホテルへのご宿泊を希望されますか？

Agents for Amazon Bedrock

API を呼び出しタスクを実行する Agent 機能をフルマネージドで提供

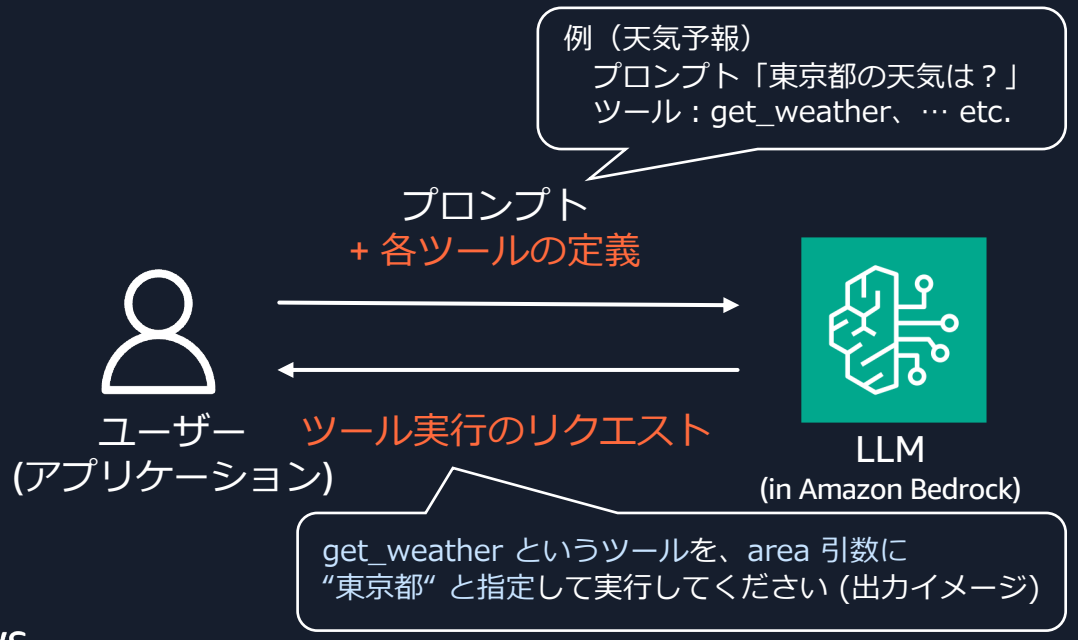
基盤モデルを使ってユーザのクエリを理解し、登録された情報を
Knowledge Base から検索したり、タスク完了に必要なアクションを実行

その他関連実装パターン

Tool use (Function calling)

外部ツールや関数などを定義し呼び出すことで
モデルの能力を拡張する機能

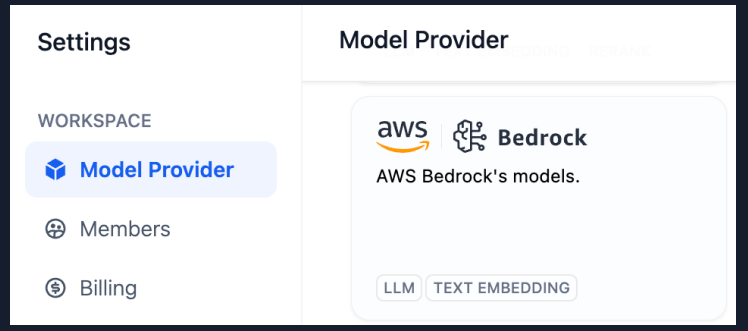
「どのツールをどのような引数で呼ぶべきか」を
モデルの出力として返し、ユーザー側 (アプリ側)
でツールを実行



OSS からの Amazon Bedrock 活用

Dify LangGenius 社提供のオープンソースの
LLM アプリケーション開発プラットフォーム

Agent 機能、AI ワークフロー構築、RAG パイプラインなどの
機能が搭載。モデルプロバイダーとして Amazon Bedrock を
利用可能。



LangChain LLM アプリケーション開発のための
オープンソースフレームワーク

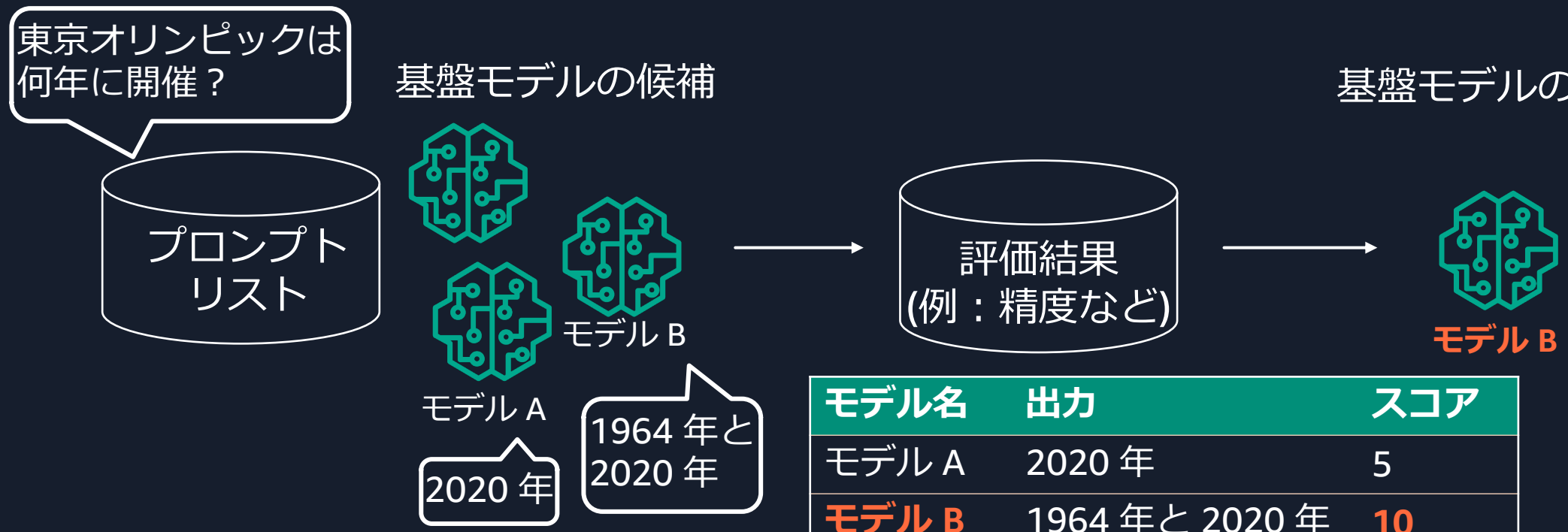
LLM Agent 、チャットボット、RAG、リコメンデーションなど
様々なアプリケーションのパターンが実装可能となる
ライブラリ。LLM として Amazon Bedrock を利用可能。

生成 AI アプリ開発でよくある課題

- 課題 1 ：独自データの活用
- 課題 2 ：データや外部 API との連携の実現
- 課題 3 ：基盤モデルの評価と最適なモデルの選択
- 課題 4 ：生成 AI アプリの適切な利用状況の維持と安定的稼働
- 課題 5 ： 生成 AI 活用のアイデアの迅速な実験と評価

よくある課題 3：
お客様のユースケースに適した基盤モデルの選択の必要性

最適な基盤モデルを選択するためには、ユースケースに沿ったプロンプトを用いて
評価の実施および評価プログラムの実装などの工数がかかる。



モデル名	出力	スコア
モデル A	2020 年	5
モデル B	1964 年と 2020 年	**10**

Model Evaluation on Amazon Bedrock

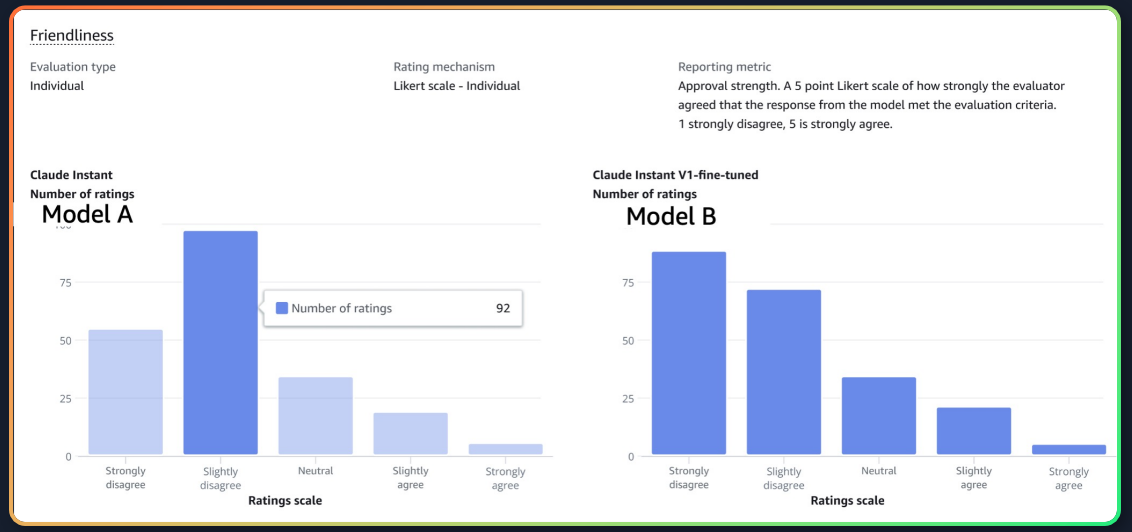
複数の基盤モデルをコード不要で比較・評価し
ユースケースに最適なモデルを選択可能に

評価方法：
自動評価 (Automatic Evaluation)、または
人間による評価 (Human Evaluation)を選択

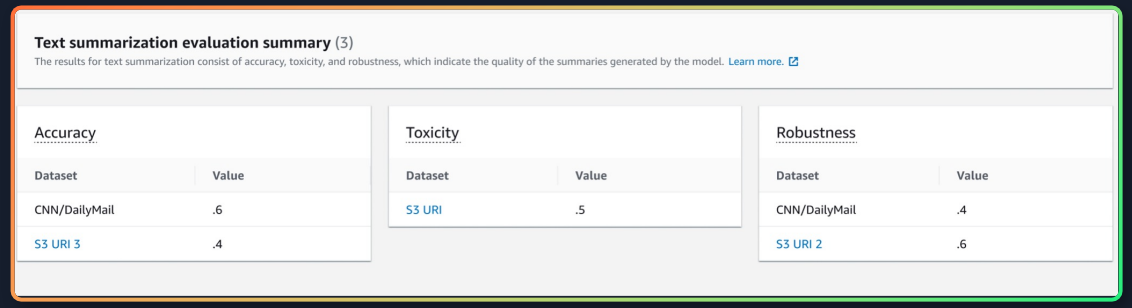
評価用データセット：
お客様独自のデータセット、または
ビルトインのデータセットを指定

評価指標：
事前定義済みの複数の指標や
お客様独自のカスタム指標も指定可能

人間による評価についてのレポート



自動評価レポート



バッチ推論

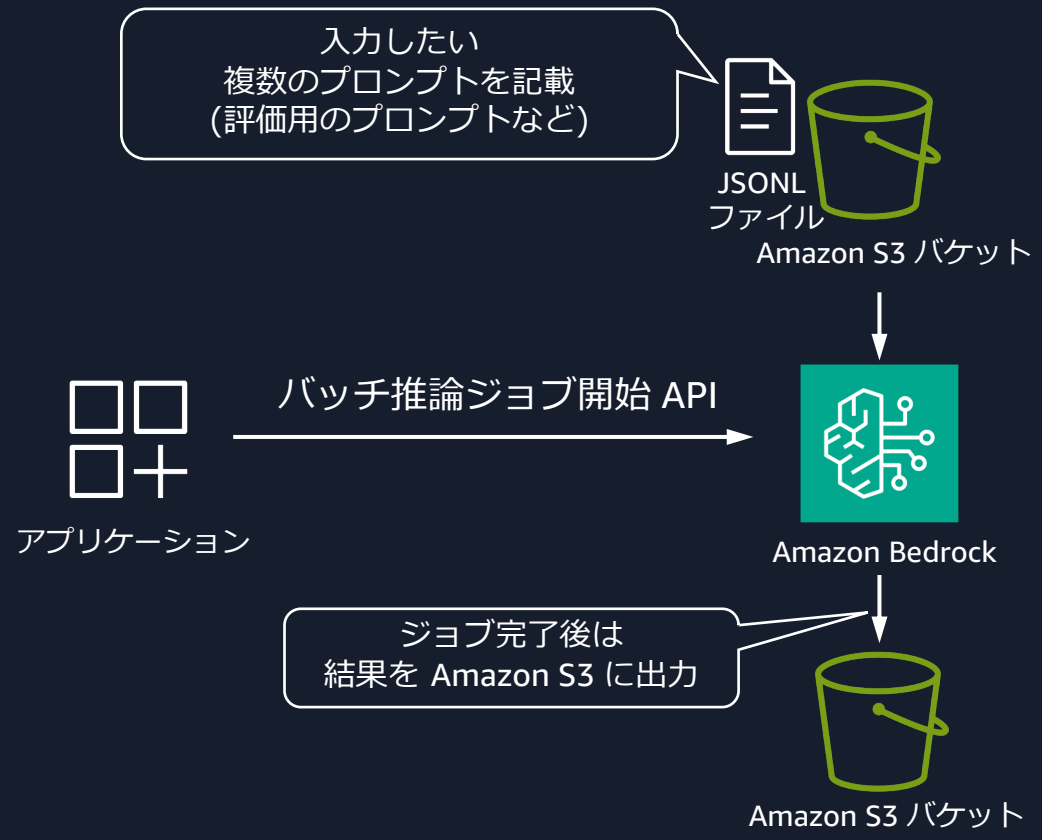
複数のプロンプトを非同期にバッチで推論

スロットリングを回避し

大規模なジョブを確実に処理

その他ユースケース

- 定期的なオフラインでの推論
  - 例：ユーザーごとにパーソナライズされたメール配信の文面作成
- RAG などの用途での埋め込みベクトルの一括作成

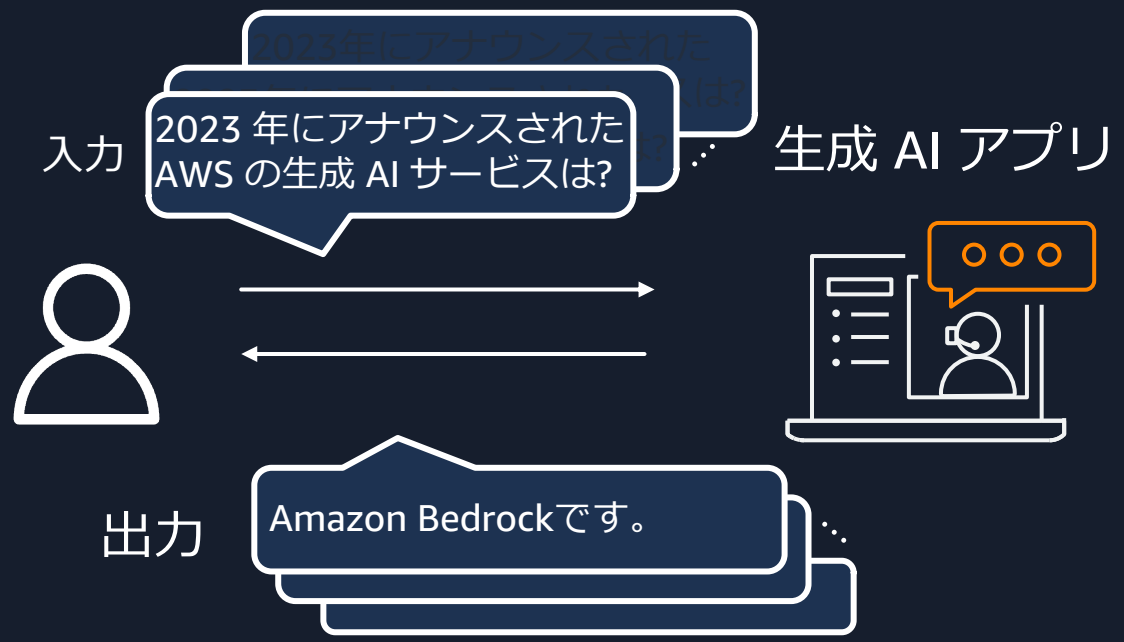


生成 AI アプリ開発でよくある課題

- 課題 1 ：独自データの活用
- 課題 2 ：データや外部 API との連携の実現
- 課題 3 ：基盤モデルの評価と最適なモデルの選択
- 課題 4 ：生成 AI アプリの適切な利用状況の維持と安定的稼働
- 課題 5 ： 生成 AI 活用のアイデアの迅速な実験と評価

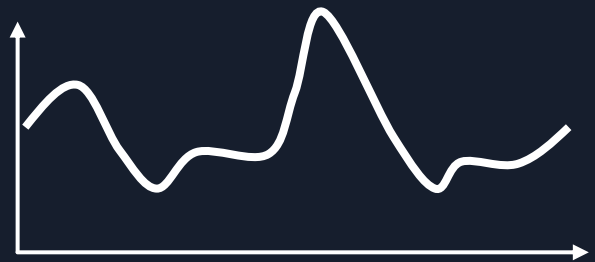
よくある課題 4 :
生成 AI アプリの適切な利用状況の維持と安定的稼働

生成 AI の不適切な利用や有害な結果の出力を避けられているか、
ユーザーのリクエストに安定的に応えられているか、管理し維持する必要がある。



スケーリング

リクエスト



プロンプトのモニタリング

- ログ管理・分析
- 精度評価

Guardrails for Amazon Bedrock

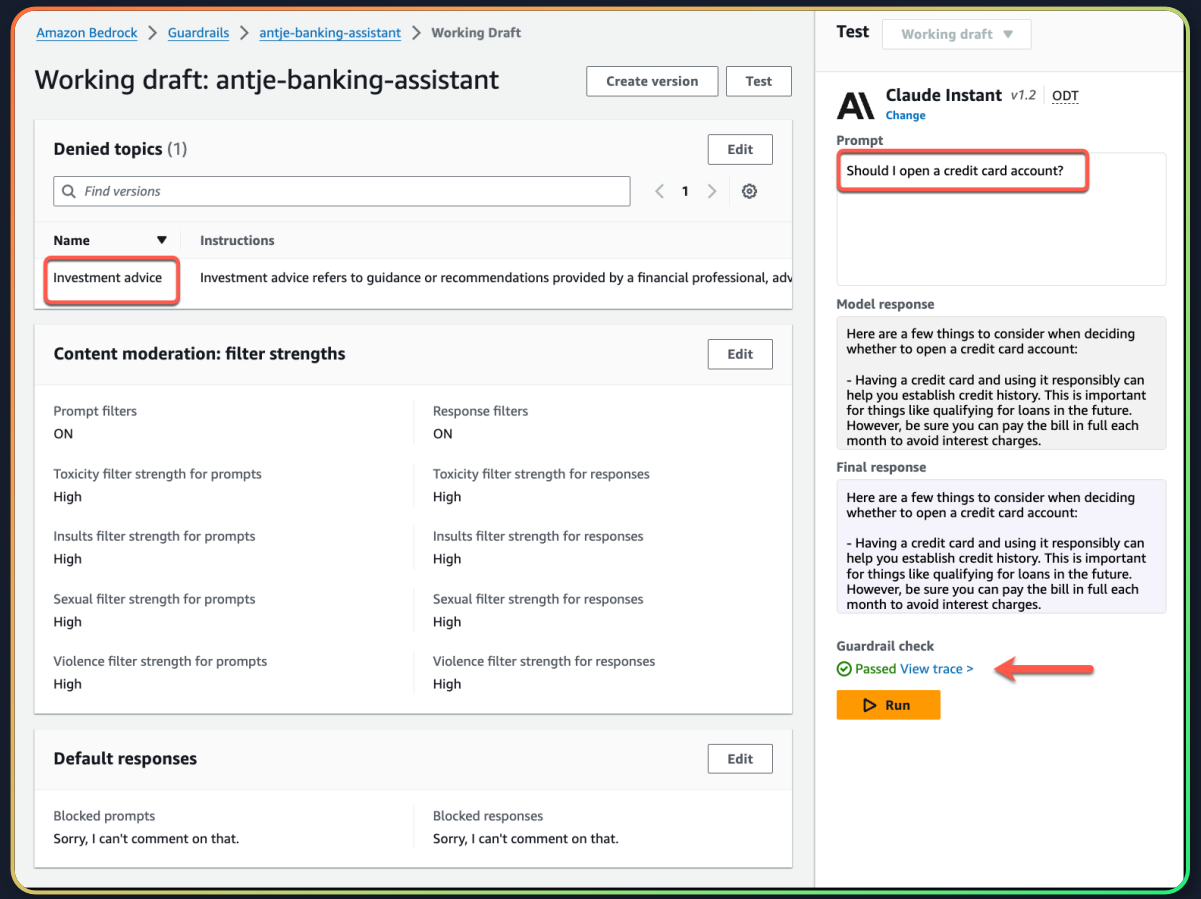
基盤モデルの不適切な入出力をブロックし、責任ある AI ポリシーの実現

自社ポリシーに従って、ユーザーの入力や基盤モデルの出力から望ましくないコンテンツを制御するためのガードレールを定義

- 拒否トピック
- コンテンツフィルター
- PII リダクション
- ワードフィルター

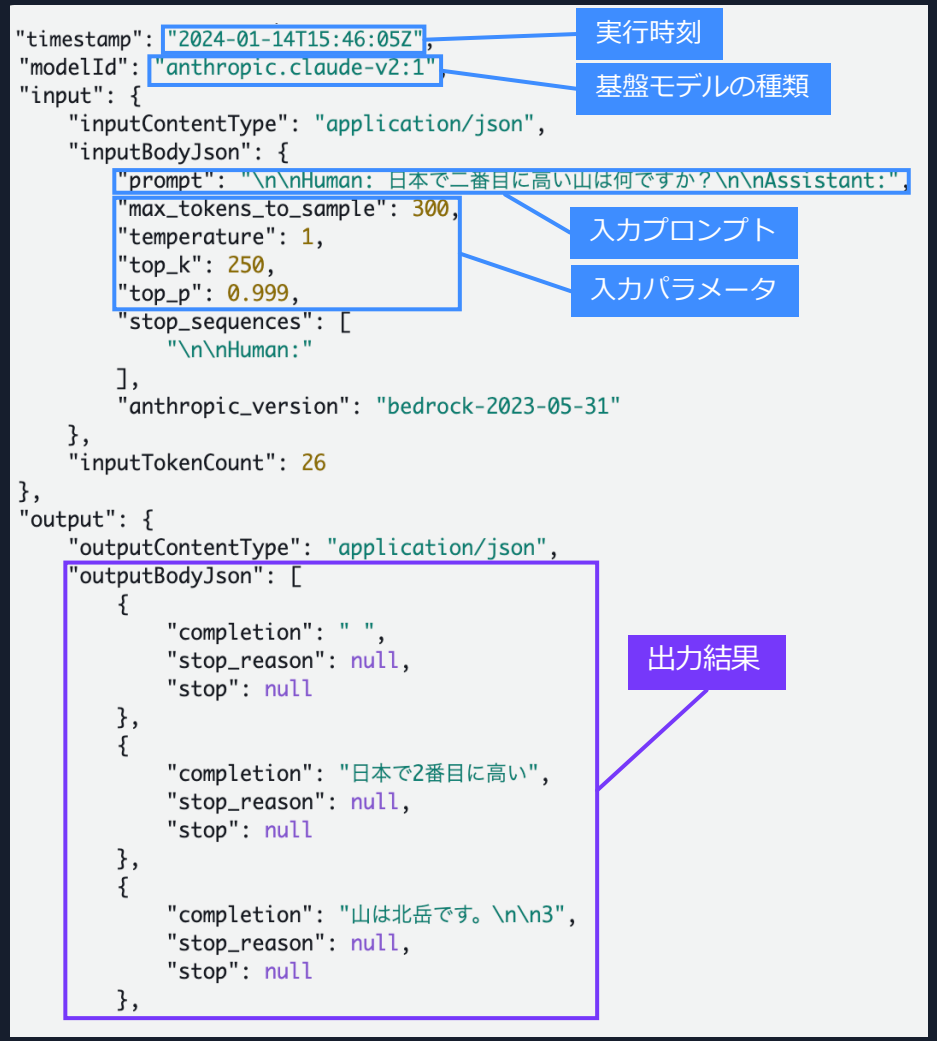
基盤モデルがネイティブで提供する保護機能より 85% 多くの有害コンテンツをブロック

ファインチューニングされたモデルや Agents / Knowledge Bases for Amazon Bedrock に対しても適用可能



実行履歴の記録: Model Invocation Logging

基盤モデルの入出力結果をログとして記録



- デフォルトでは AWS CloudTrail によって API 実行時刻と基盤モデルの種類が記録される

- Model Invocation Logging を有効化することで基盤モデルの入出力内容や入力パラメータ設定値も含めた実行履歴を記録することが可能に

- ログの保存先には Amazon S3 や Amazon CloudWatch を指定

Provisioned Throughput

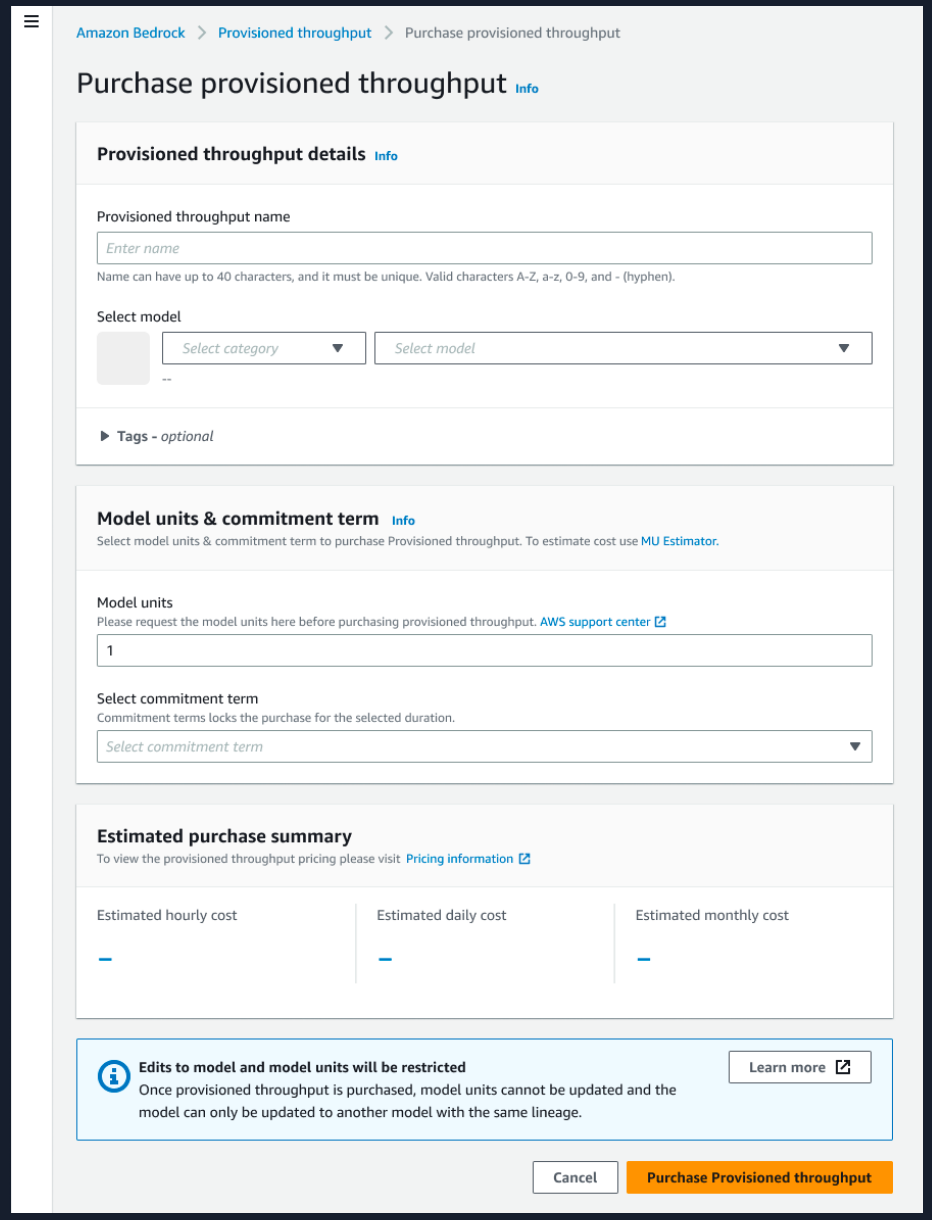
スループットを確保し安定した API 実行が可能に

通常の方式（オンデマンド方式）の場合、
1 分間当たりのリクエスト数およびトークン数に上限がある

→ モデルユニット（ 1 分間に処理可能な入出力トークン数に基づく処理単位）を購入し
一定のスループットを確保

トラフィックの急増に対応し、
一貫したユーザー体験を確保

コミット期間なし or コミット期間付き
（1 ヶ月または 6 ヶ月）での購入

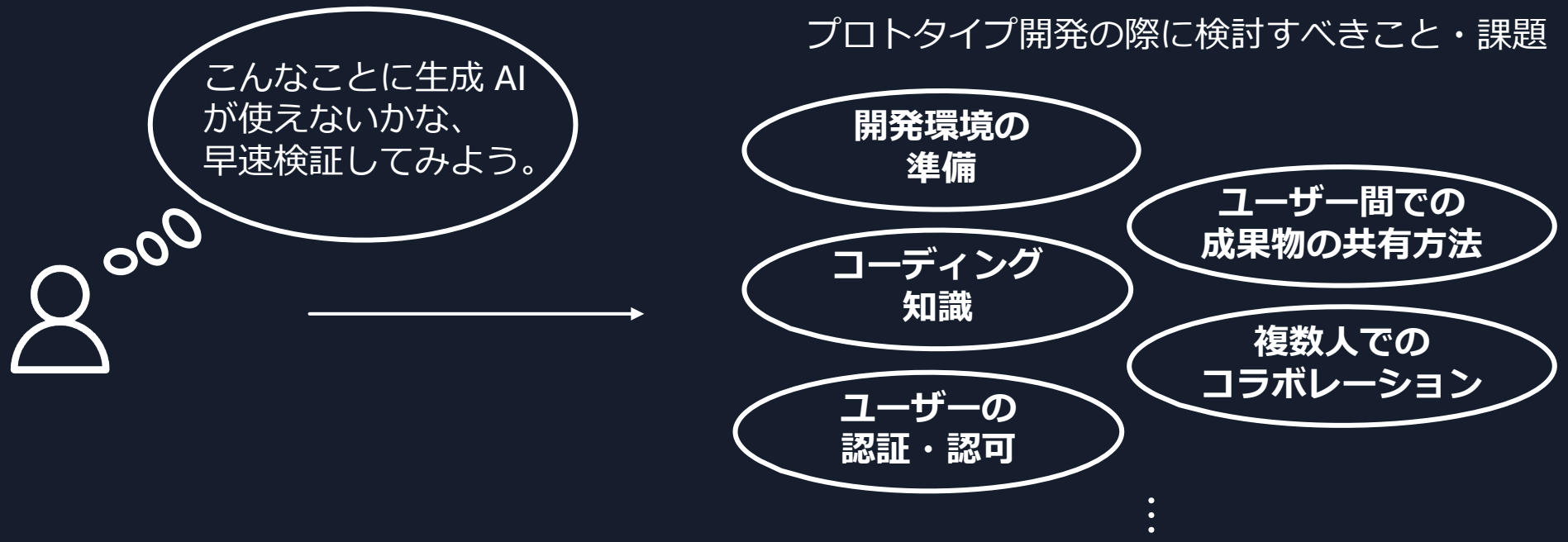


生成 AI アプリ開発でよくある課題

- 課題 1 ：独自データの活用
- 課題 2 ：データや外部 API との連携の実現
- 課題 3 ：基盤モデルの評価と最適なモデルの選択
- 課題 4 ：生成 AI アプリの適切な利用状況の維持と安定的稼働
- 課題 5 ： 生成 AI 活用のアイデアの迅速な実験と評価

よくある課題 5：
生成 AI 活用のアイデアの迅速な実験と評価

生成 AI ユースケースのアイデアの検証や評価のスピードを加速したい。
その際、コーディングや認証認可など、本来検証したい生成 AI 以外の部分がアイデア検証のスピードを妨げる。



Amazon Bedrock Studio

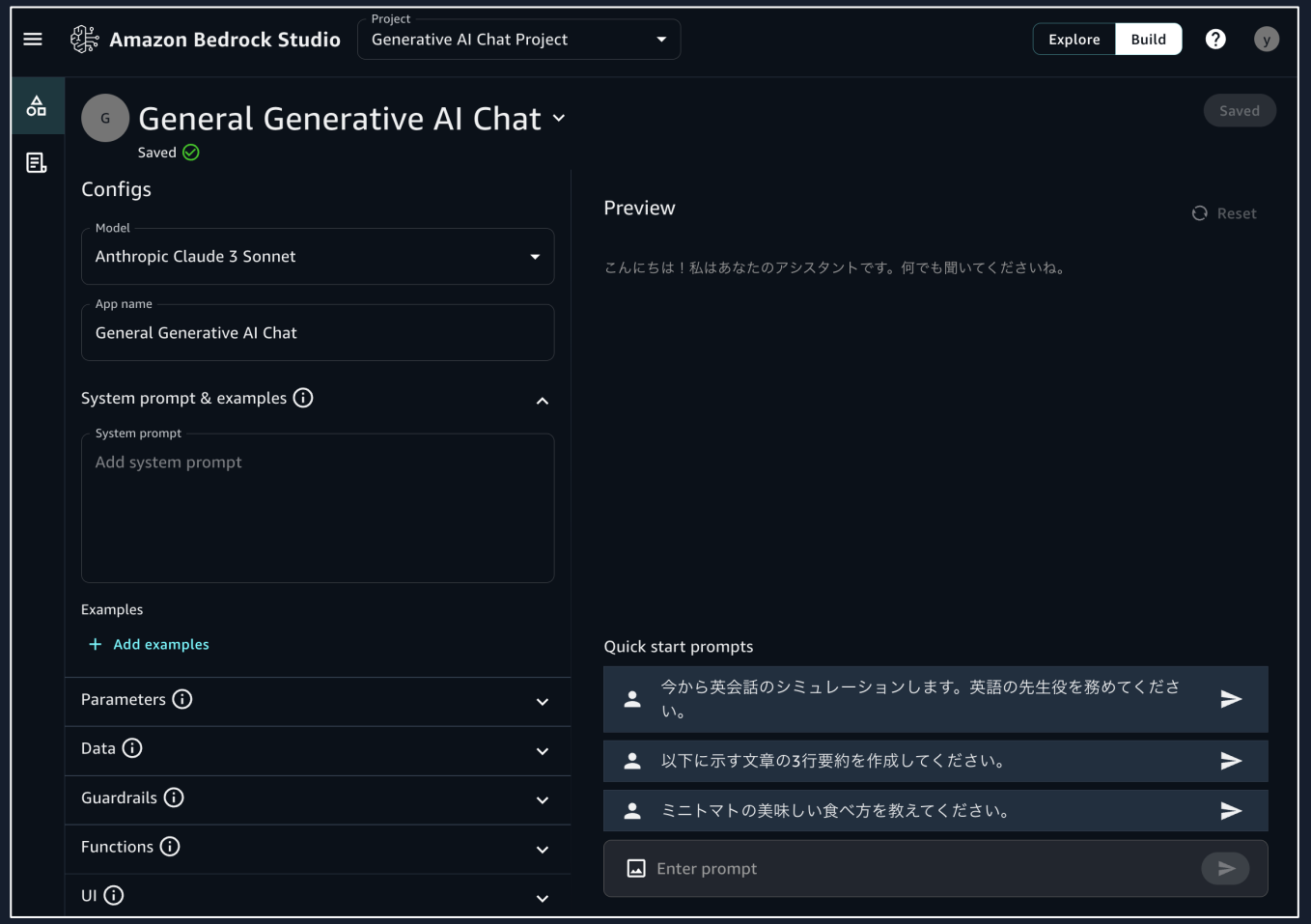
生成 AI アプリをコード不要で簡単にプロトタイプ化でき
組織内で共有できる SSO 対応 Web アプリケーション

Amazon Bedrock の各種機能を活用した
生成 AI アプリケーションのプロトタイプ開発

- 複数の基盤モデルの選択・呼び出し
- ナレッジベース
- エージェント（Function）
- ガードレール

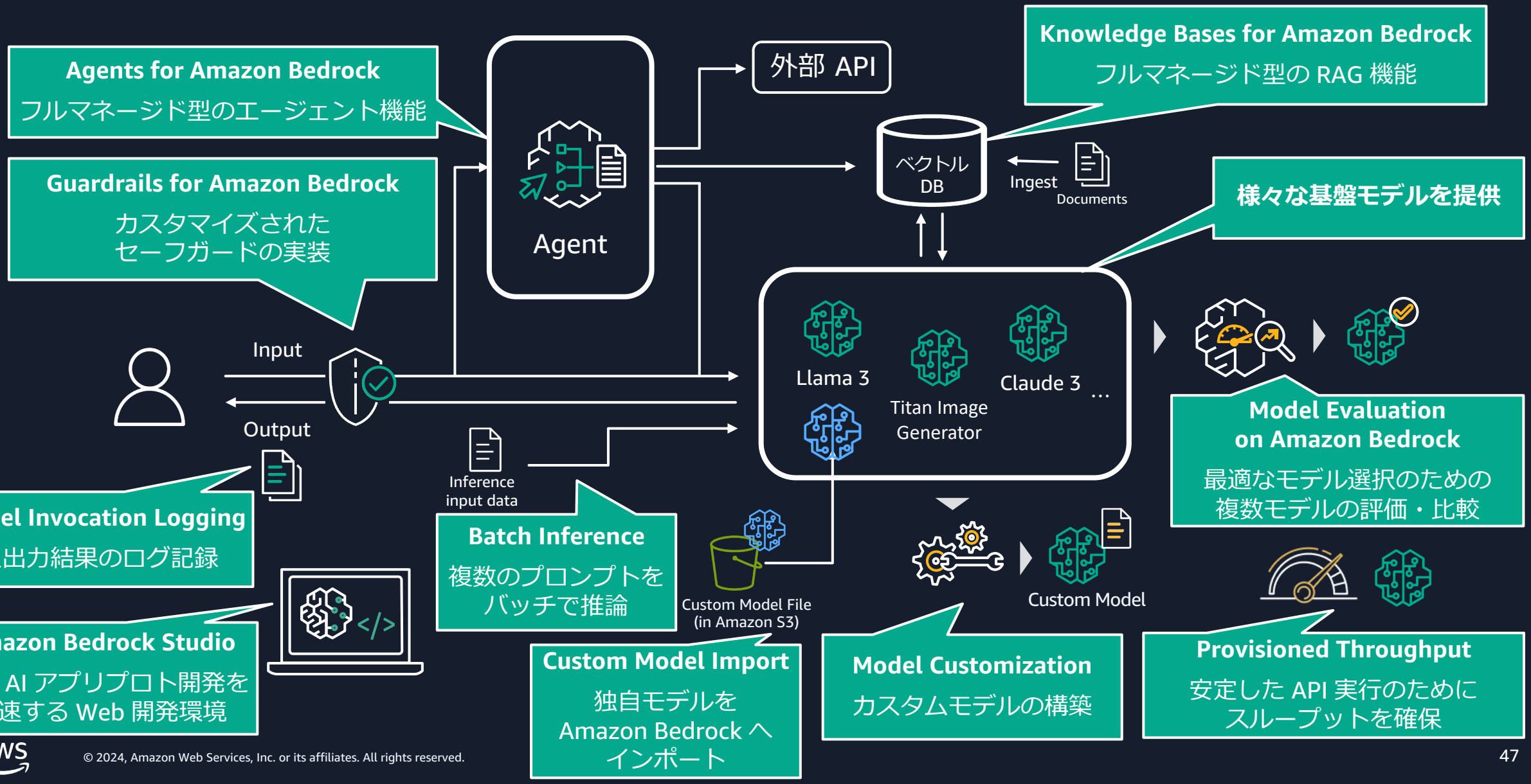
SSO でのログイン
(AWS IAM Identity Center)

作成したプロトタイプアプリケーションを
組織間で簡単に共有



（再掲）Amazon Bedrock の主な機能の全体像

2024 年 6 月 ver.



まとめ

- Amazon Bedrock は基盤モデルを活用した生成 AI アプリケーションを簡単に構築可能なマネージドサービス
- 様々な基盤モデルをユースケースに応じて選択し API を通じて利用可能
- 生成 AI アプリケーションの開発・運用をサポートする様々な機能を提供
    - RAG (検索拡張生成)
    - Fine Tuning / Continued Pre-training
    - Custom Model Import
    - エージェント
    - 基盤モデル評価
    - ガードレール
    - Amazon Bedrock Studio
    - … etc.

AWS Black Belt Online Seminar とは

- 「サービス別」「ソリューション別」「業種別」などのテーマに分け、アマゾン ウェブ サービス ジャパン合同会社が提供するオンラインセミナーシリーズです
- AWS の技術担当者が、 AWS の各サービスやソリューションについてテーマごとに動画を公開します
- 以下の URL より、過去のセミナー含めた資料などをダウンロードすることができます
  - https://aws.amazon.com/jp/aws-jp-introduction/aws-jp-webinar-service-cut/
  - https://www.youtube.com/playlist?list=PLzWGOASvSx6FIwIC2X1nObr1KcMCBBlqY

ご感想は X (Twitter) へ！ハッシュタグは以下をご利用ください
#awsblackbelt

Thank you!