生成系 AI on AWS

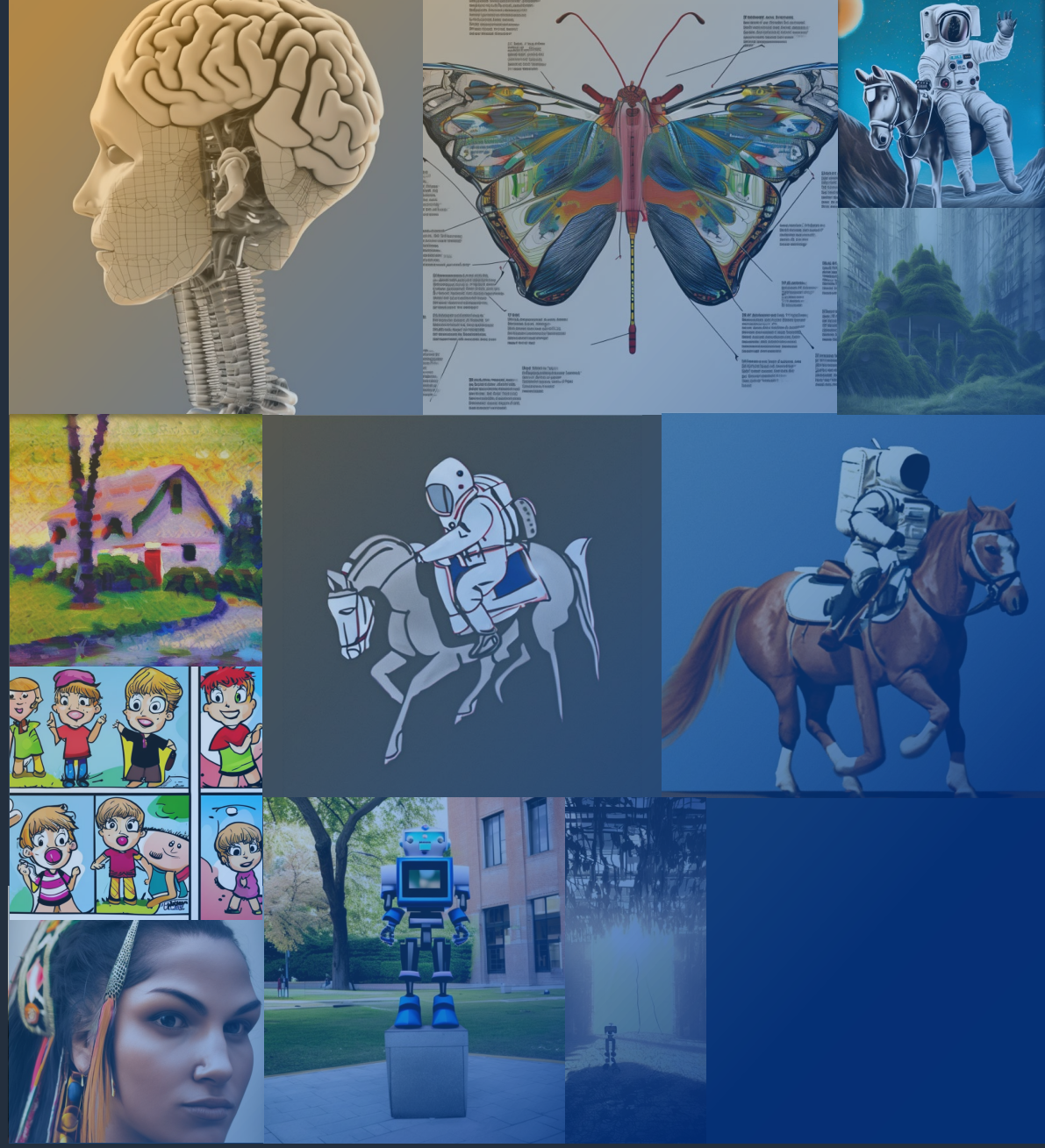
山口 賢人

インダストリー事業開発部 (メディア & エンターテインメント)
アマゾン ウェブ サービス ジャパン合同会社

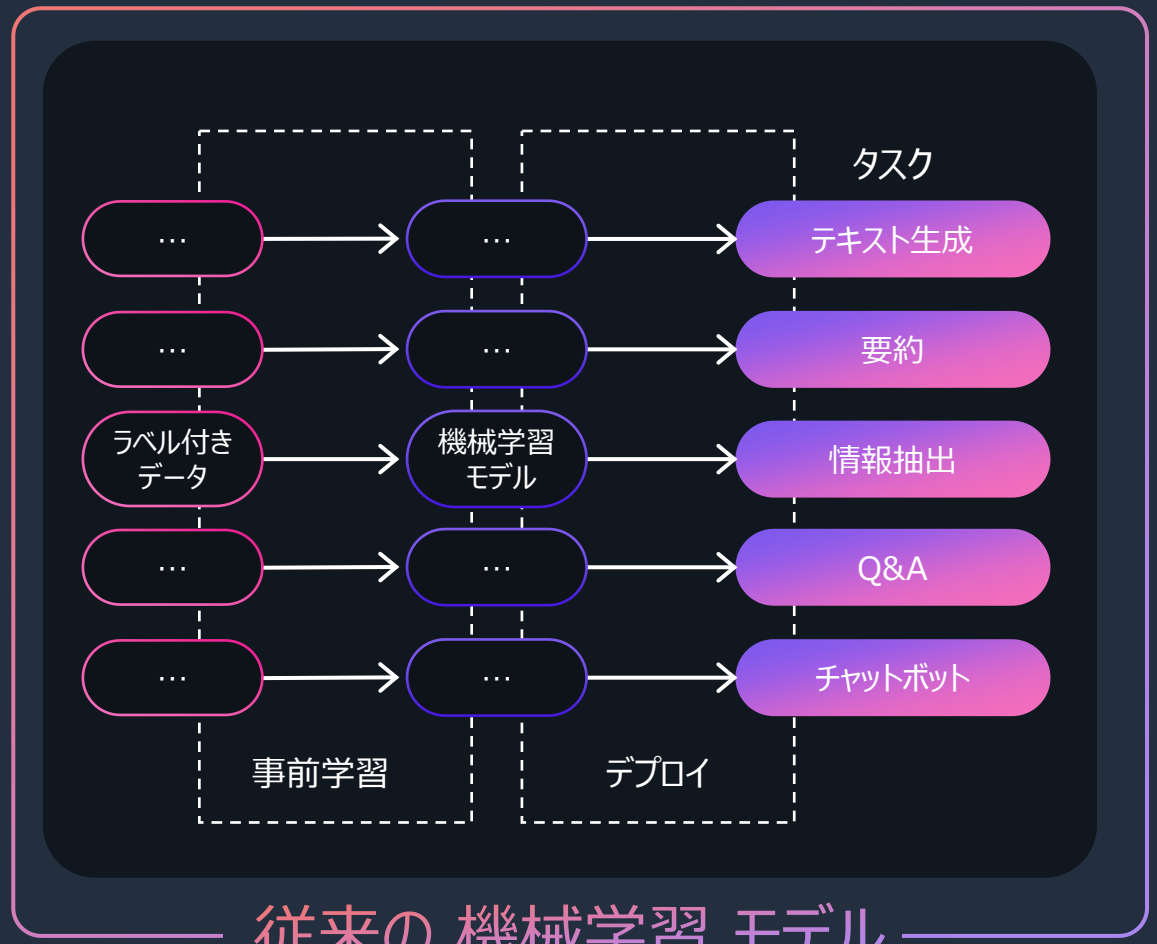
Generative AI
(生成系 AI) とは？

- 新たなコンテンツやアイデアを生成する
(会話、ストーリー、画像、映像、音楽)

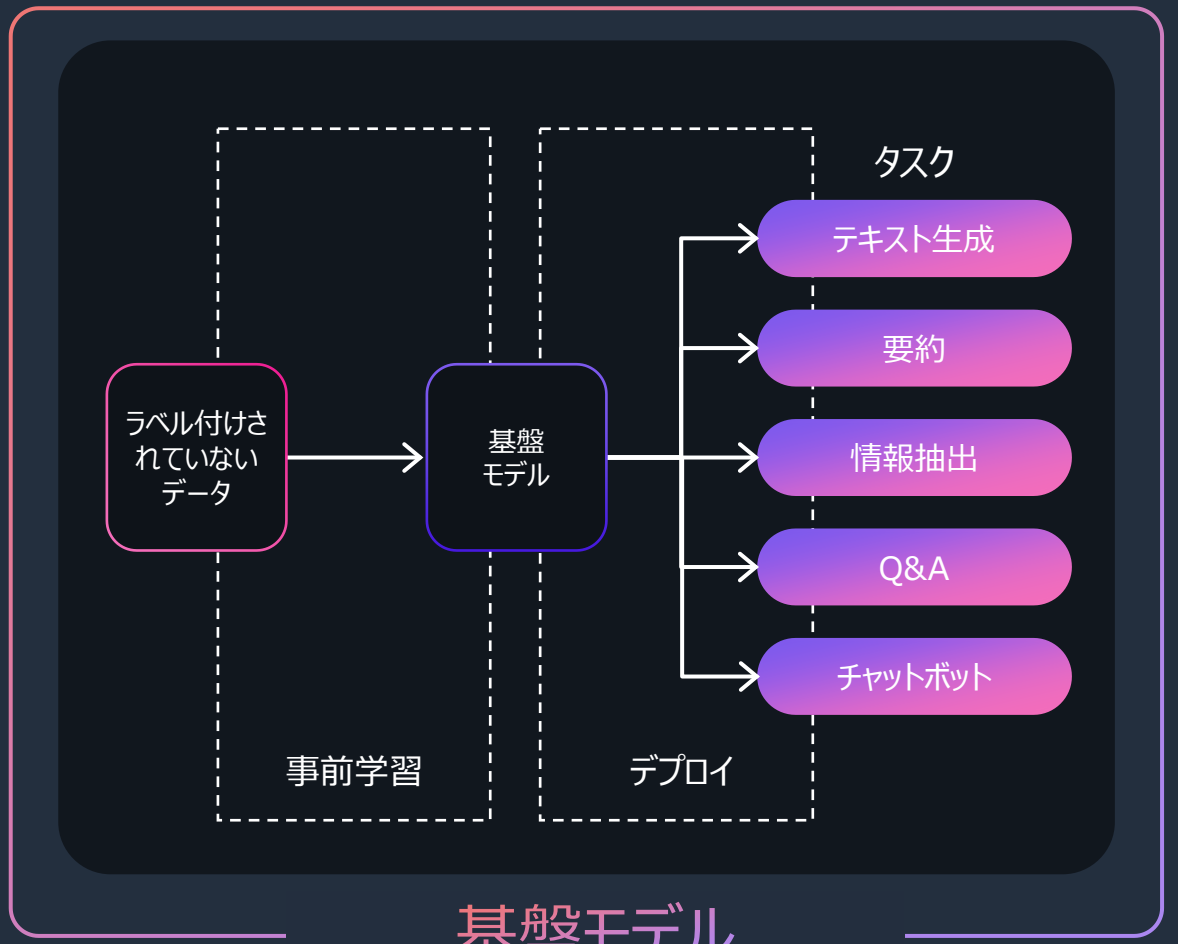
- 膨大なデータで事前学習された大規模な
モデルによって実現される
(一般に、基盤モデル/Foundation models:
FMs と呼ばれる)



なぜ基盤モデルなのか



従来の 機械学習 モデル

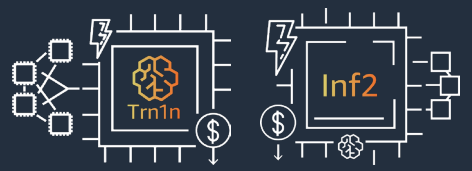


基盤モデル

AWSの取り組み

Amazon
Bedrock

お客様の業務課題を解決する
基盤モデルによる生成系 AI アプリを
サーバレスで素早く開発可能に



Amazon EC2 Trn1n インスタンス
Amazon EC2 Inf2 インスタンス

トレーニングと推論に
高いコストパフォーマンスを発揮する
専用設計の機械学習アクセラレータを搭載

Amazon
Sagemaker

機械学習モデルを高速に開発、学習、デプロイする
ためのフルマネージド型サービス
Jumpstartというモデルカタログの仕組みで生成系AIの利用可

何を実現するのか？

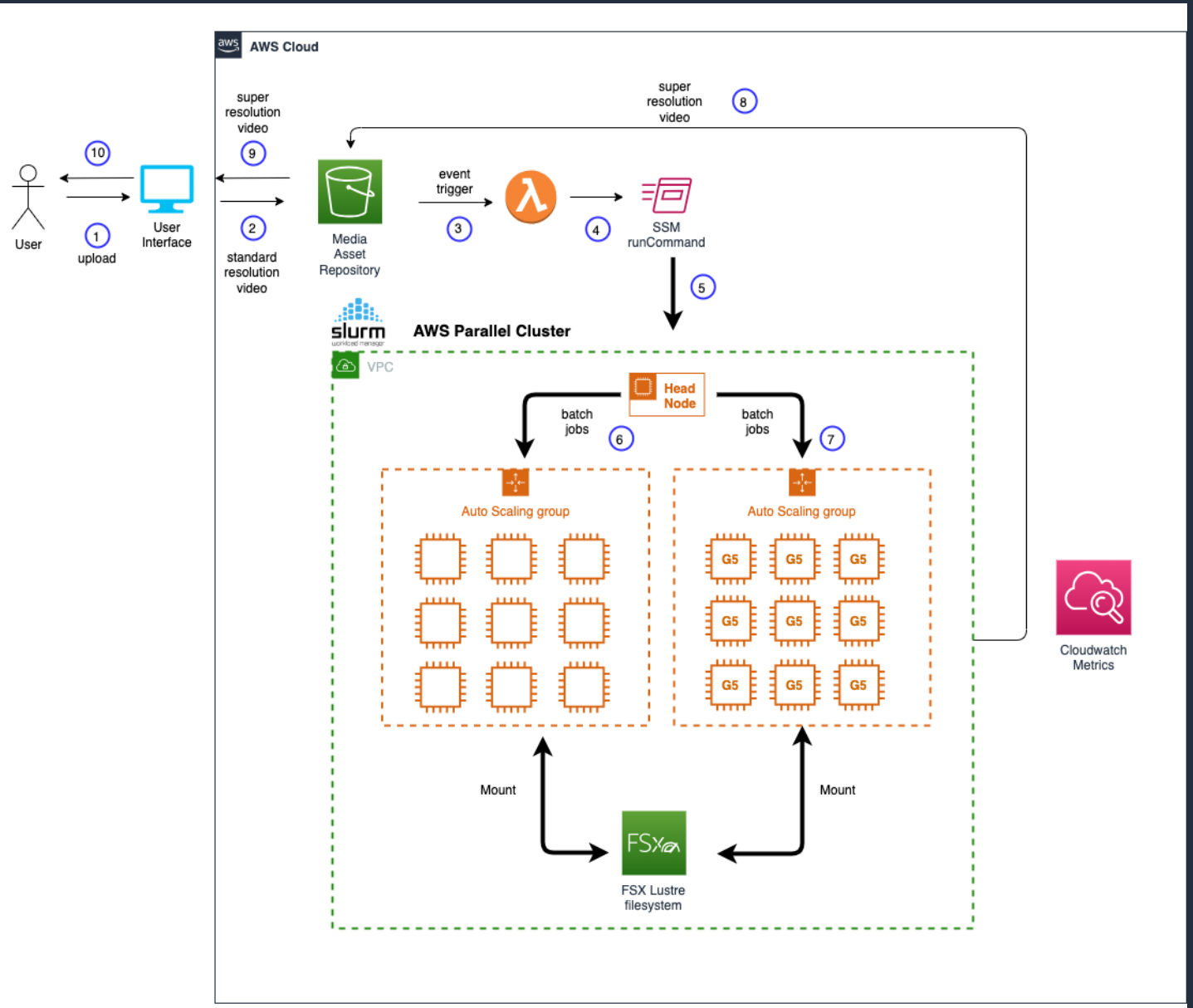
競争力向上

新規事業の開拓

社会への貢献

Demo: Super resolution

Super resolution video creation



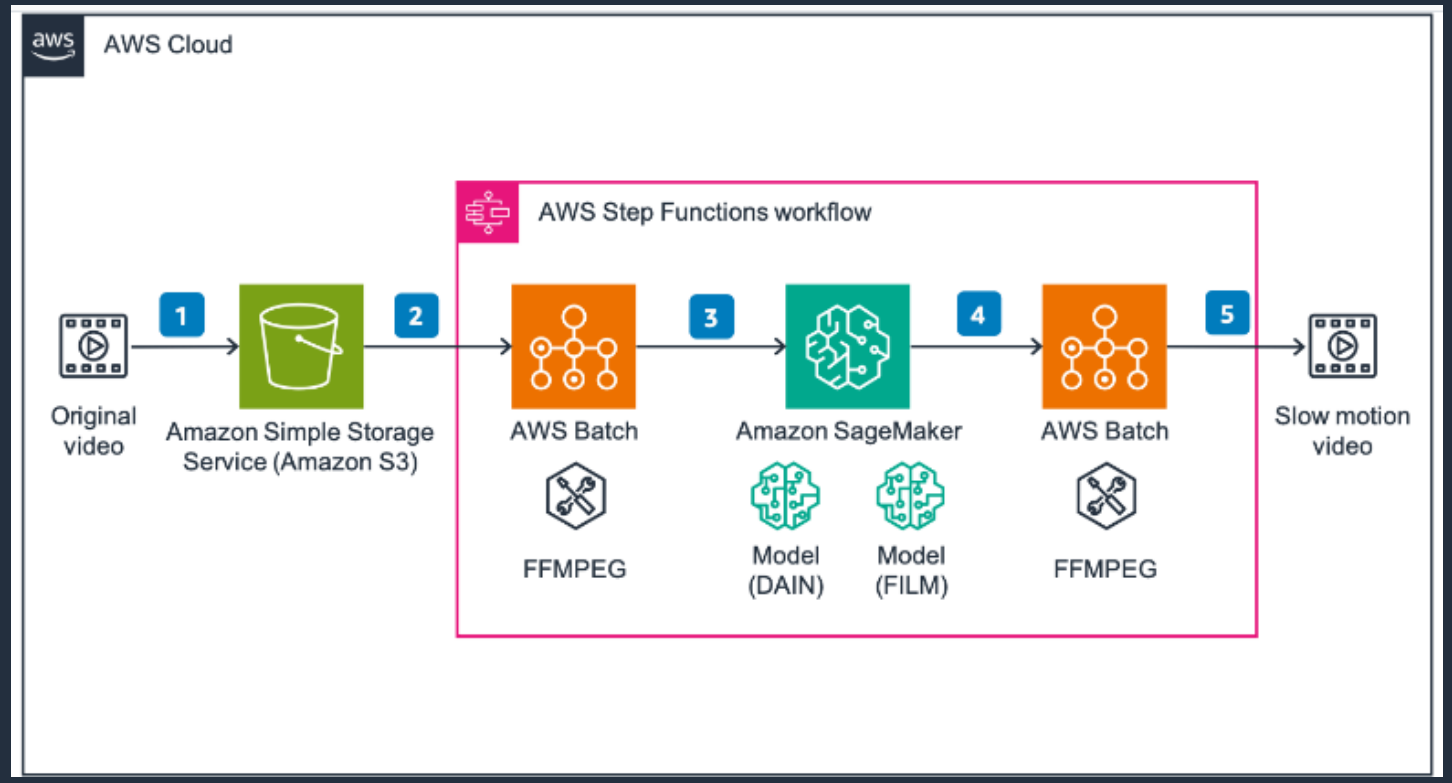
処理の流れ:

1. 映像アップロードとフレーム抽出
2. SlurmスケジューラとAWS Parallel Clusterによりオーケストレーション
3. AIモデルによるアップスケーリング
4. 超解像ビデオの出力

基盤モデル:
Real-ESRGAN / SwinIR

Demo: Super slow motion video creation

Super slow motion video creation



処理の流れ:

1. 映像アップロード
2. AWS BatchでFFMPEGを利用しフレーム抽出
3. Amazon Sagemaker上にホストされているモデルで補完処理
4. FFMPEGでスーパースロモ映像出力

基盤モデル:

DAIN / FILM

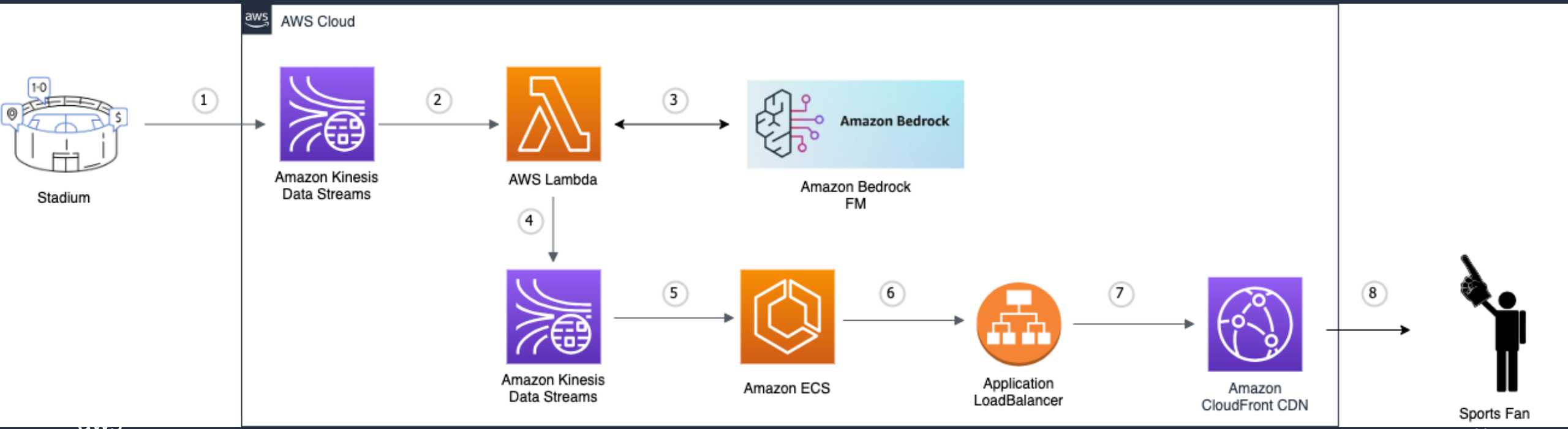
Demo: Live Sports Commentary

Live Sports Commentary

処理の流れ:

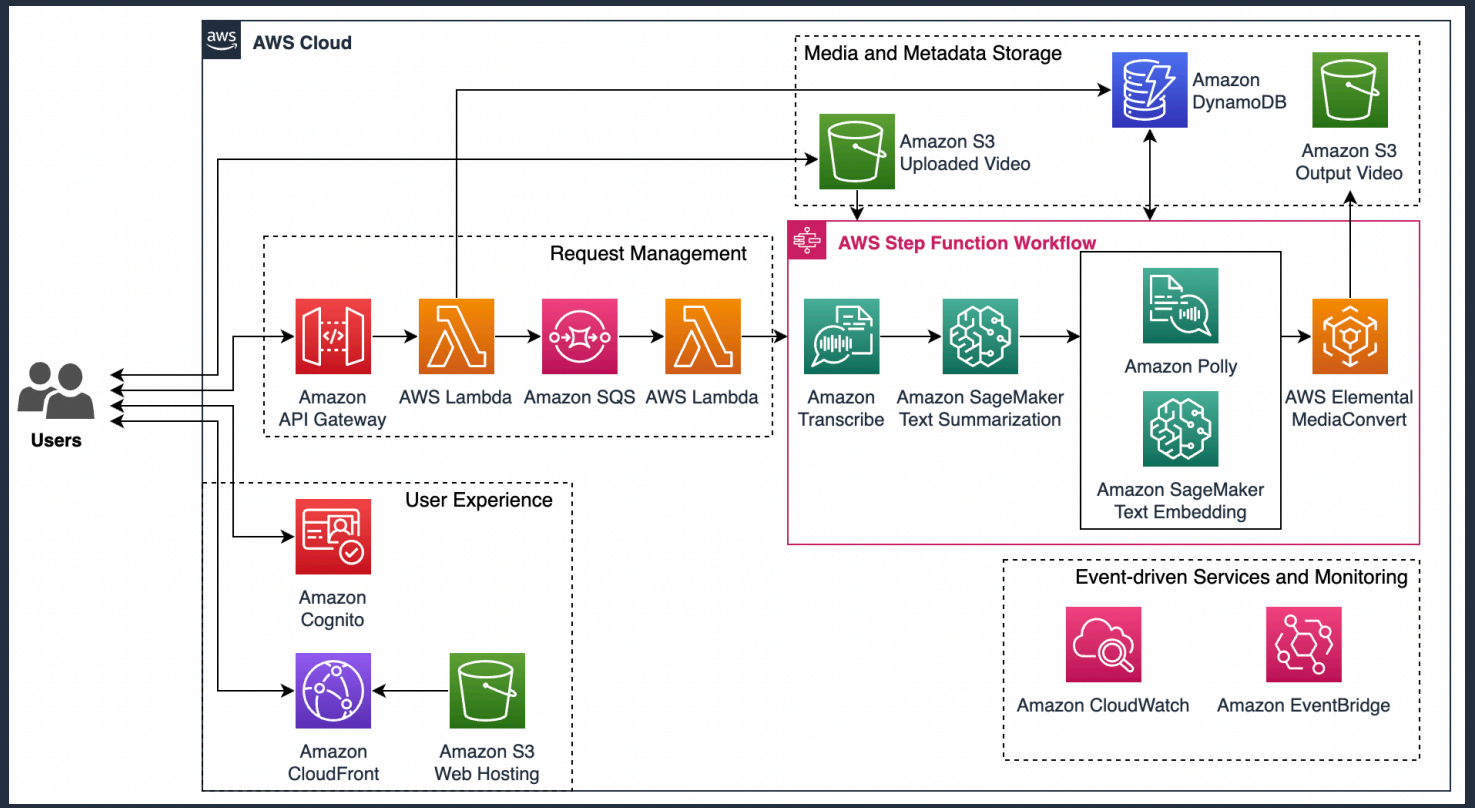
1. スポーツデータフィードのインジェスト
2. 抽出したデータから、プロンプトを生成
3. Amazon Bedrockを使い基盤モデルに投入し、処理を実行
4. コメンタリーデータを取得

基盤モデル:
Jurassic-2 Ultra



Demo: Video Summarization with AWS AI and ML Services

Video Summarization with AWS AI and ML Services



処理の流れ:

1. ビデオ音声からテキストを取得し、タイムコード付きの字幕ファイルを生成
2. Amazon SageMaker Jumpstart のデプロイされている基盤モデルで上記テキストを要約
3. Amazon Pollyで上記テキストを音声化
4. 要約コンテンツの文章が該当するビデオシーケンスを選択しビデオを作成

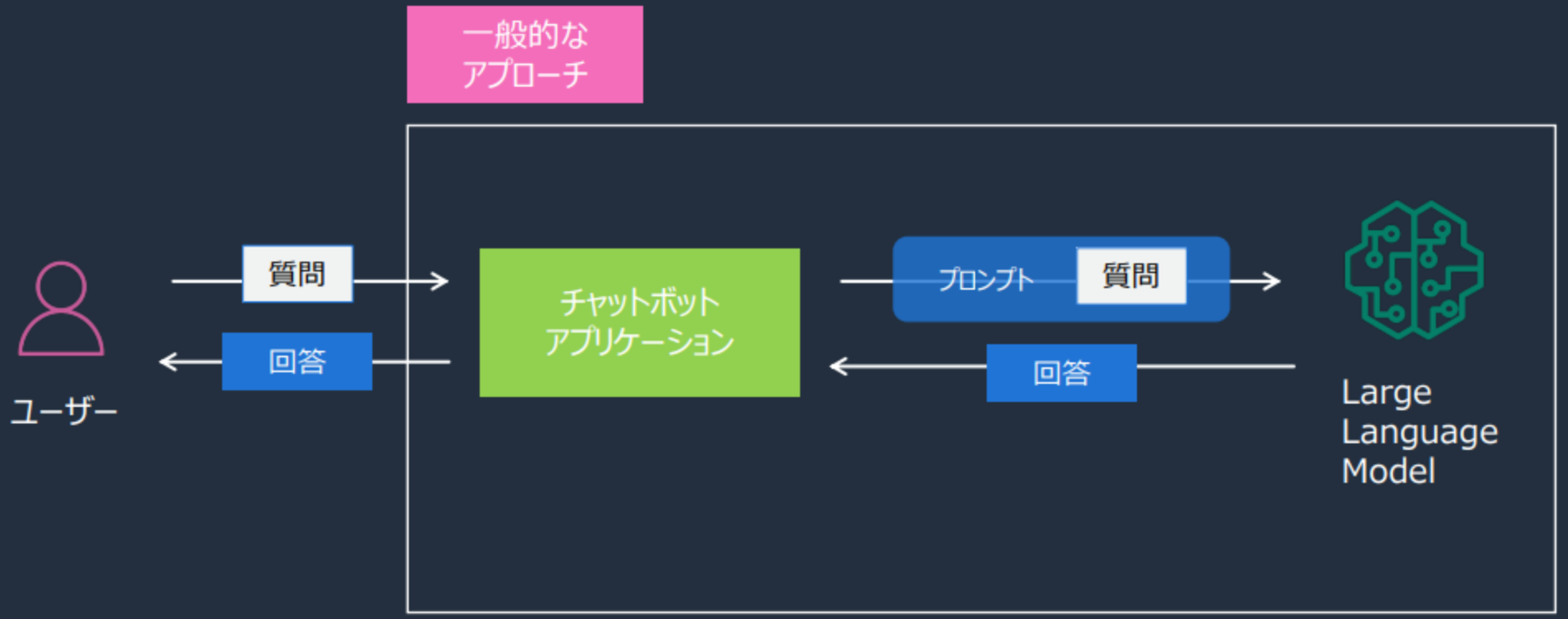
基盤モデル:

Llama2 / all-MiniLM-L6-v2

https://aws.amazon.com/jp/blogs/media/video-summarization-with-aws-artificial-intelligence-ai-and-machine-learning-ml-services/

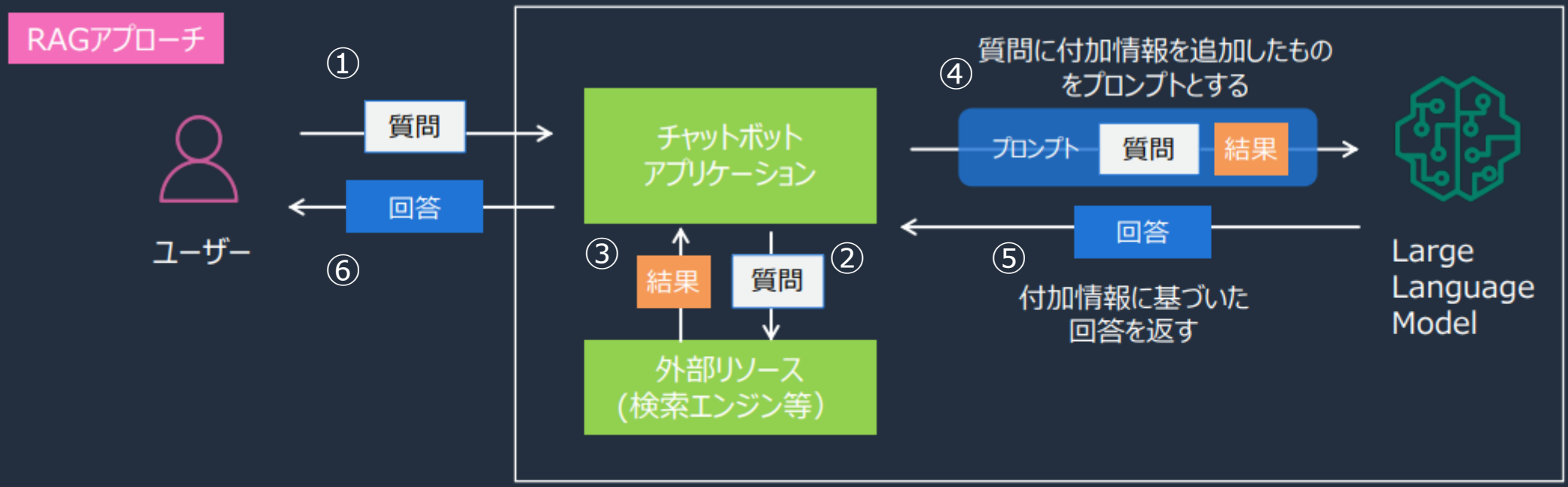
Demo: Amazon KendraとLLMを使ったチャットボット

LLM を使ったチャットボットアプリ



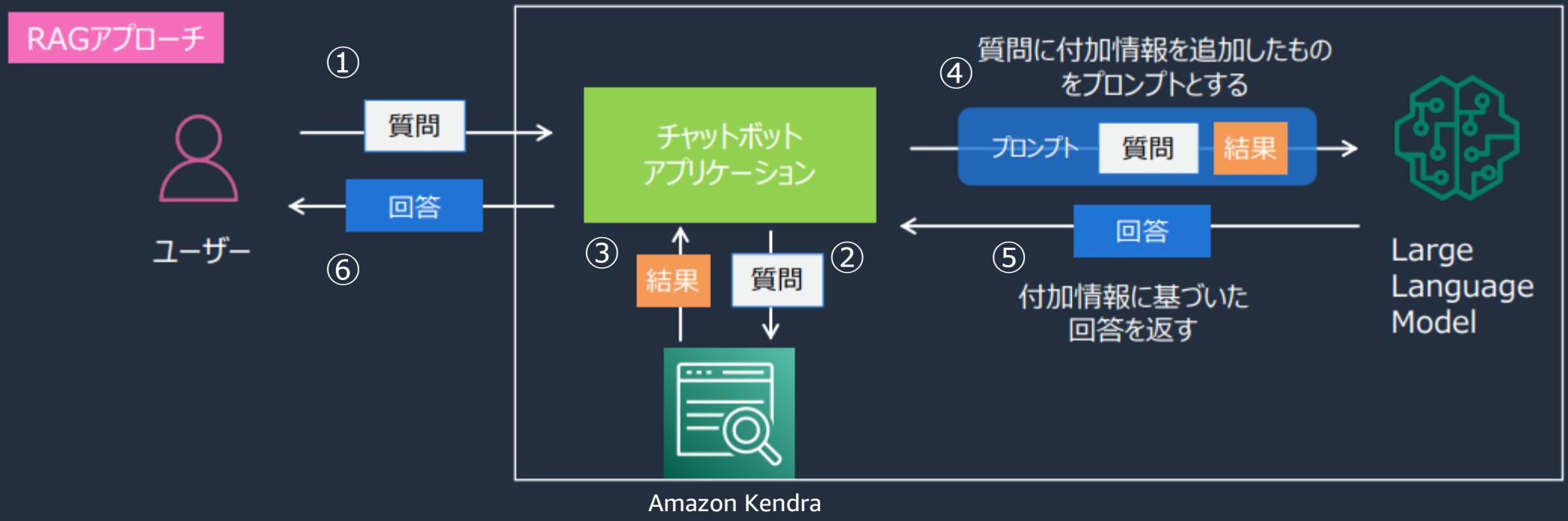
RAG (Restrieval Augmented Generation) とは

社内検索エンジン等のリソースからデータを取得した結果を
Generative AI への入力となるプロンプトに付与することで、
言語モデルの正確性を強化したり制限する方法。



RAG (Restrieval Augmented Generation) とは

社内検索エンジン等のリソースからデータを取得した結果を
Generative AI への入力となるプロンプトに付与することで、
言語モデルの正確性を強化したり制限する方法。



Amazon Kendra の特徴

1. セットアップが簡単

ML モデルの**自身での管理は不要**（AWS が管理）、
Amazon S3 や RDS、Gmail、Google Drive、Salesforce、Jira などのデータを取得する **Connector の用意**

2. 見つけやすい

ML ベースのドキュメントランキング、質問への応答、該当箇所のハイライト、FAQ
Amazon S3 や RDS（データベース）、Gmail、Google Drive、Salesforce、Jira **などのデータを横断検索**

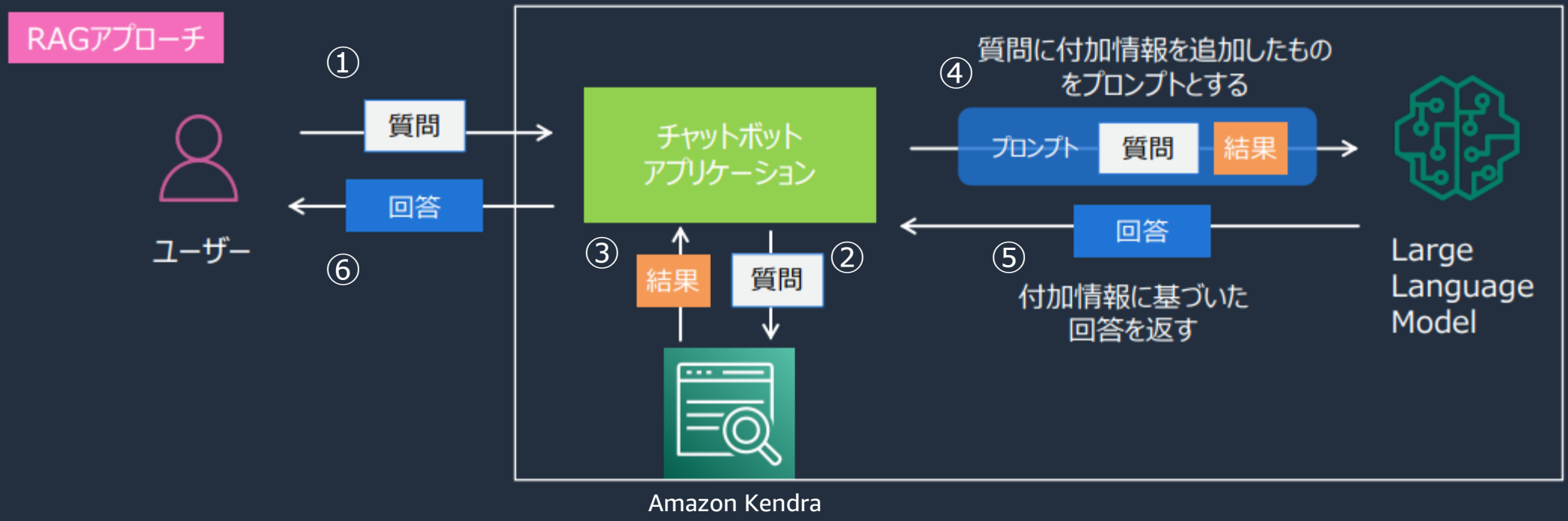
3. 分析と改善が簡単

セキュリティ、**チューニング**、Search Analytics Dashboard、Custom Document Enrichment、スケール

Demo: Amazon KendraとLLMを使ったチャットボット

RAG (Restrieval Augmented Generation) とは

社内検索エンジン等のリソースからデータを取得した結果を
Generative AI への入力となるプロンプトに付与することで、
言語モデルの正確性を強化したり制限する方法。



デモ - LLM へのプロンプト例

Kendra から取得した
ドキュメントの内容

以下は人間とAIの間の友好的な会話です。
AI はおしゃべりで、コンテキストから多くの具体的な詳細を提供します。
AI は、質問に対する答えが分からない場合、正直に「分かりません」と答えます。

Document Title: AWS278_JRFU_1222_6.pdf
Document Excerpt:
能にし、スタジアム中継車の専門機器の削 減や、制作コストの最適化を進めることも 検討しています。
「今回のプロジェクトで、さまざまな映像サービスを提供するための基盤が整備できまし た。
これからは、新たなファンエンゲージメ ント創出の支援や試合映像制作を効率化する仕組みを
提供していきます。 システムの 利用範囲も日本代表や学生ラグビーなどに 広げていきますので、
AWS には引き続き さまざまな支援を期待しています」（室口氏）
2022 年 1 月のリーグ戦開始に合わせて、試合映像を蓄積してチーム、
放送局、競技 関係者等と共有する映像管理システムを、アマゾン ウェブ サービス（AWS）を活用し
……
……

指示: 上記の文書に基づいて、Kendra とは何ですか？ に対する詳細な回答を提供してください。
文書に記載されていない場合は「分かりません」と回答してください。
回答:≈

白い文字が、プロンプトのテンプレート

AWS が提供できる価値

素早い生成系 AI
アプリケーション開発
API を通じて
基盤モデルを利用
インフラ管理が不要

利用用途に合った
基盤モデルを
選択可能

お客様のデータを
外部に出すことなく
基盤モデルを
カスタマイズ・活用可能

Thank you!