AWS で始める！
出版業界向け生成 AI 活用のご紹介

大規模言語モデルから画像生成まで

アマゾン ウェブ サービス ジャパン合同会社

2023/06/29

アジェンダ

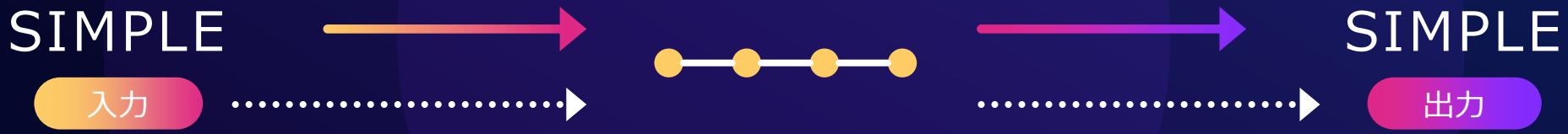
- 登壇者紹介
- AWS の生成 AI に関する取り組み
- AWS を活用した出版業界での生成 AI のユースケース
- 早速始めるために
- まとめ

AWS の生成 AI に関する取り組み

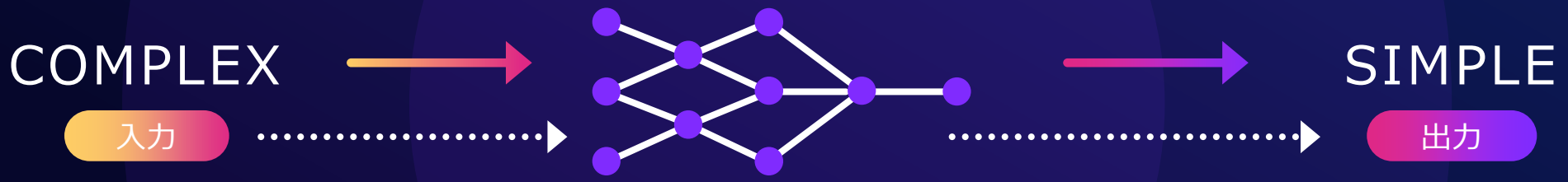
生成系 AI とは？

- 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
- 一般に 基盤モデル（Foundation Model) と呼ばれる
  膨大なデータに基づいて事前にトレーニングされた大規模モデルを利用

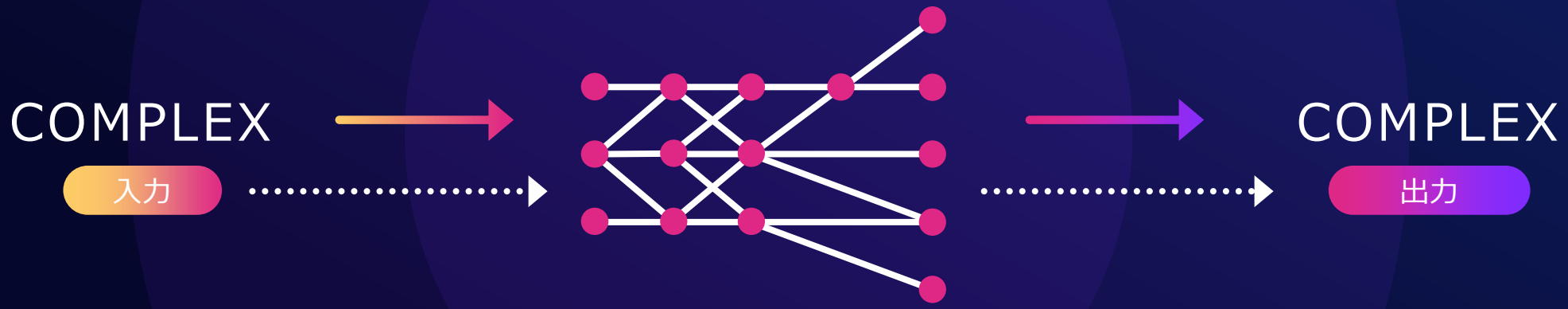
機械学習



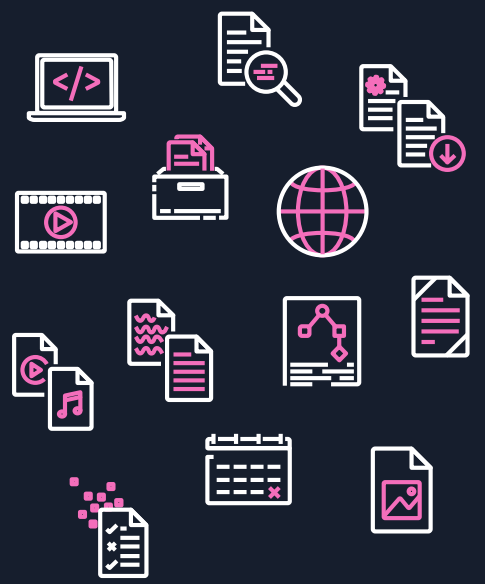
深層学習



基盤モデル



基盤モデルの仕組み



膨大なラベル無しデータ

事前学習

幅広い知識

基盤モデル

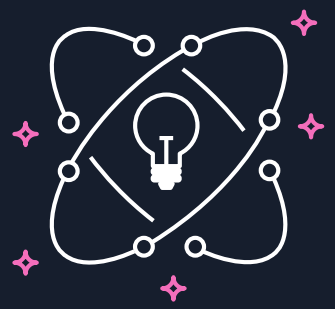
ファインチューン

特定業務の正確性

- テキスト生成
- 要約
- 情報抽出
- Q&A
- チャットボット

少数のラベル付きデータ

生成系 AI が創るビジネス価値の広がり



創造性

新しいコンテンツや
アイデアの創出

会話、物語、画像、映像、
音楽等

生産性

あらゆる事業部門の
生産性を大幅に向上

経済成長

今後 10 年間で
世界の GDP が 7 兆ドル増加

生成系 AI の適用例

コミュニケーション
チャットボット、QA 対応、検索

医療・ヘルスケア
タンパク質構造過程、医薬品開発、個別化医療、
医療画像の高解像度化

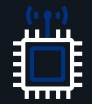
メディア・エンターテイメント
ビデオゲーム生成、コンテンツの高解像度化、
顔合成、フィルム保存＆着色

モビリティ
自動運転車両、燃費向上のための設計支援

金融サービス
リスク管理、不正検知

小売
価格最適化、在庫最適化、
ブランド・カテゴリ別商品フラグ付け適正化

エネルギー・公益
地域別再生可能エネルギー配置の最適化、予知保全



ハードウェア技術
半導体デザイン、ロボティックス

AWS 上の生成系 AI の活用

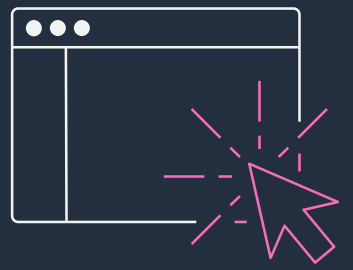
AWS 上で生成系 AI を活用し、ビジネス価値をより早く勝ち得るために

基盤モデルの選択肢

柔軟な基盤モデル選択
安全にカスタマイズ
迅速にアプリ反映して差別化

コスパが良い

最適なパフォーマンスと
コストを両立する
マネージド・インフラ



成果まで早い

通常業務タスクから
生成系 AI を組み合わせ

AWS で生成系 AI を動かすには

- Amazon Bedrock（生成AI開発/限定プレビュー中）
  - 基盤モデルを選択し生成系 AI アプリをサーバレスで開発
- Amazon SageMaker（機械学習特化の PaaS）
  - SageMaker JumpStart ではモデルを SageMaker 上にコピーし Fine Tuning 可能
- Amazon EC2（IaaS/4 月に新インスタンス発表）
  - 基盤モデルをホスト。EC2上で独自開発する。学習/推論特化のインスタンスを4月に発表。

ユースケースに応じてアプリ開発しやすいクラウド環境を用意

Amazon Bedrock

最も簡単に

基盤モデルと生成系 AI アプリケーションを

開発、横展開する方法

プレビュー

お客様にとっての価値

- 生成系 AI アプリケーション開発の加速
  各種基盤モデルを単一 API から利用可能
- インフラ管理不要
- 複数基盤モデルからの選択可能
  Amazon, AI21 Labs, Anthropic, Stability AI
- 非公開に自社データを使用して
  基盤モデルをカスタマイズ
- セキュリティ最重視の AWS 運用に沿った
  包括的なセキュリティ管理

Amazon Bedrock で始める最先端の基盤モデル

Amazon Titan

テキスト要約・生成・分類
自由形式の Q&A、情報抽出
Embedding、検索

Jurassic-2

テキスト生成用 LLM
多言語対応

Claude

会話、質問応答、ワークフロー自動化用 LLM
誠実かつ責任ある AI 学習を最重視

Stable Diffusion

リアルな高品質画像生成、高精細化
ユニークなアート、ロゴ、デザイン生成

注記：上記以外の基盤モデルも、 Amazon SageMaker JumpStartでご利用いただけます

Amazon Titan

責任ある AI のベストプラクティスとして Amazon 自身が提供する高性能基盤モデル

Titan Text

自然言語処理 (NLP) に特化

Titan Embeddings

エンタープライズ検索
パーソナライゼーション

お客様にとっての価値

- 20 年以上の Amazon の機械学習活用経験と自社事業における稼働実績
- Amazon Titan Text は、テキスト要約・生成などの言語に関わるタスクを自動化
- Amazon Titan Embeddings は、検索精度を向上と顧客毎のパーソナライズ・レコメンデーション結果を改善
- 不適切・有害なコンテンツを制限することで、生成系 AI の責任ある利用をサポート

Amazon SageMaker JumpStart

今日から始める生成系 AI

最先端の複数の基盤モデルにアクセス！

お客様にとっての価値

- 人気の基盤モデルにアクセスできます。
  公開済みのオープン・ソース、
  企業・団体が提供するプロプライエタリを
  選ぶことができます。

- 選択の柔軟性があります。
  業務に必要な基盤モデルのサイズを選び、
  正確性、性能、コストを最適化できます。

- 数クリックで、基盤モデルを試したり、
  稼働させることができます。

- Amazon SageMaker は、5 年以上弊社が投資を
  続ける機械学習ライフサイクル全般を支える
  基幹サービスです。
  世界中の生成系 AI の企業利用を促進しています。

基盤モデルが SageMaker JumpStart で動く仕組み

Amazon SageMaker JumpStart

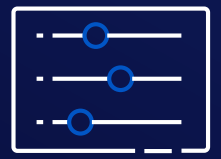
さまざまな基盤モデルに
アクセスし、試用可能
選んだ基盤モデルと
自社アプリを組み合わせて
生成系 AI アプリとして
統合可能

ブラウズ

増え続ける
優れた基盤モデルを
一覧可能

実験

本格的に
導入する前に
複数の基盤モデルで
実験
自社に合うか
試すことが可能



カスタマイズ

選んだ基盤モデルを
自社データセットで
カスタマイズ
一からのトレーニングは
不要

実装

モデルの実装
お客様生成系アプリを
本番稼働
（推論実行）

基盤モデルが SageMaker JumpStart でご利用いただけます

公開済みモデル

stability.ai

モデル

テキスト→画像
高精細化

タスク

テキスト入力から
フォト・リアルな
画像生成
生成画像の品質改善

特徴

Stable Diffusion 2.1
ファイン・チューン対応

alexa

モデル

AlexaTM
20B

タスク

機械翻訳
質問回答
テキスト要約
注釈付与
データ生成

Hugging Face

モデル

Flan T-5 models
(8 種類)
DistilGPT2, GPT2
Bloom models
(3 種類)

タスク

機械翻訳
質問回答
テキスト要約
注釈付与
データ生成

プロプライエタリ・モデル

co:here

モデル

Cohere
generate-med

タスク

テキスト生成
情報抽出
質問回答
テキスト要約

Lighton

モデル

Lyra-Fr
10B

タスク

テキスト生成
キーワード抽出
情報抽出
質問回答
テキスト要約
意味分析
(Sentiment analysis)
テキスト分類

AI21labs

モデル

Jurassic-1
Grande 17B

Tasks

テキスト生成
長文生成
テキスト要約
言い換え
チャット
情報抽出
質問回答
テキスト分類

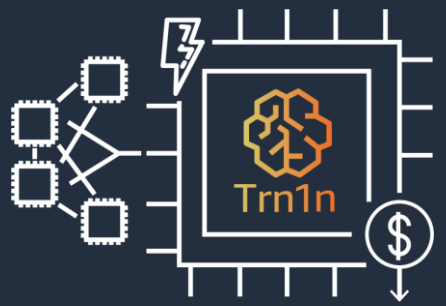
生成系 AI のための特化型アクセラレーター

AWS Inferentia

ディープラーニングモデルの
推論処理を最も安価に実現

推論処理あたりのコストを
最大 70% 削減*

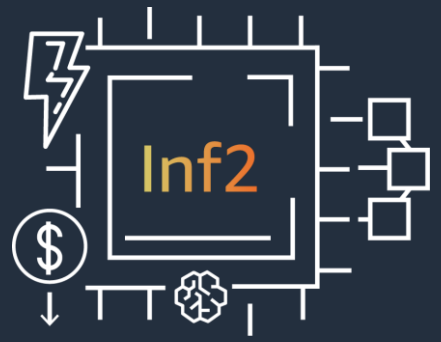
AWS Trainium



大規模言語モデルや
拡散モデルのトレーニングを
高コスト効率・高性能で実行

トレーニングコストを
最大 50% 削減 *

AWS Inferentia2



大規模言語モデルや
拡散モデルの推論を
高性能・低コストで実行

最大 40% 高い
コストパフォーマンス *

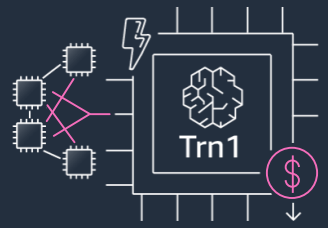
*同等の EC2 インスタンスと比較

まとめ AWS 生成系 AI サービスとインフラストラクチャ

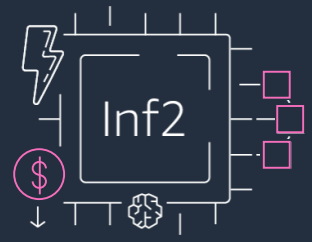
再掲

Amazon Bedrock

Amazon SageMaker



AWS Trainium



AWS Inferentia2

生成系 AI を活用したユーズケースとデモ

＋すぐに始められる解説ブログ& 一部サンプルコード付き

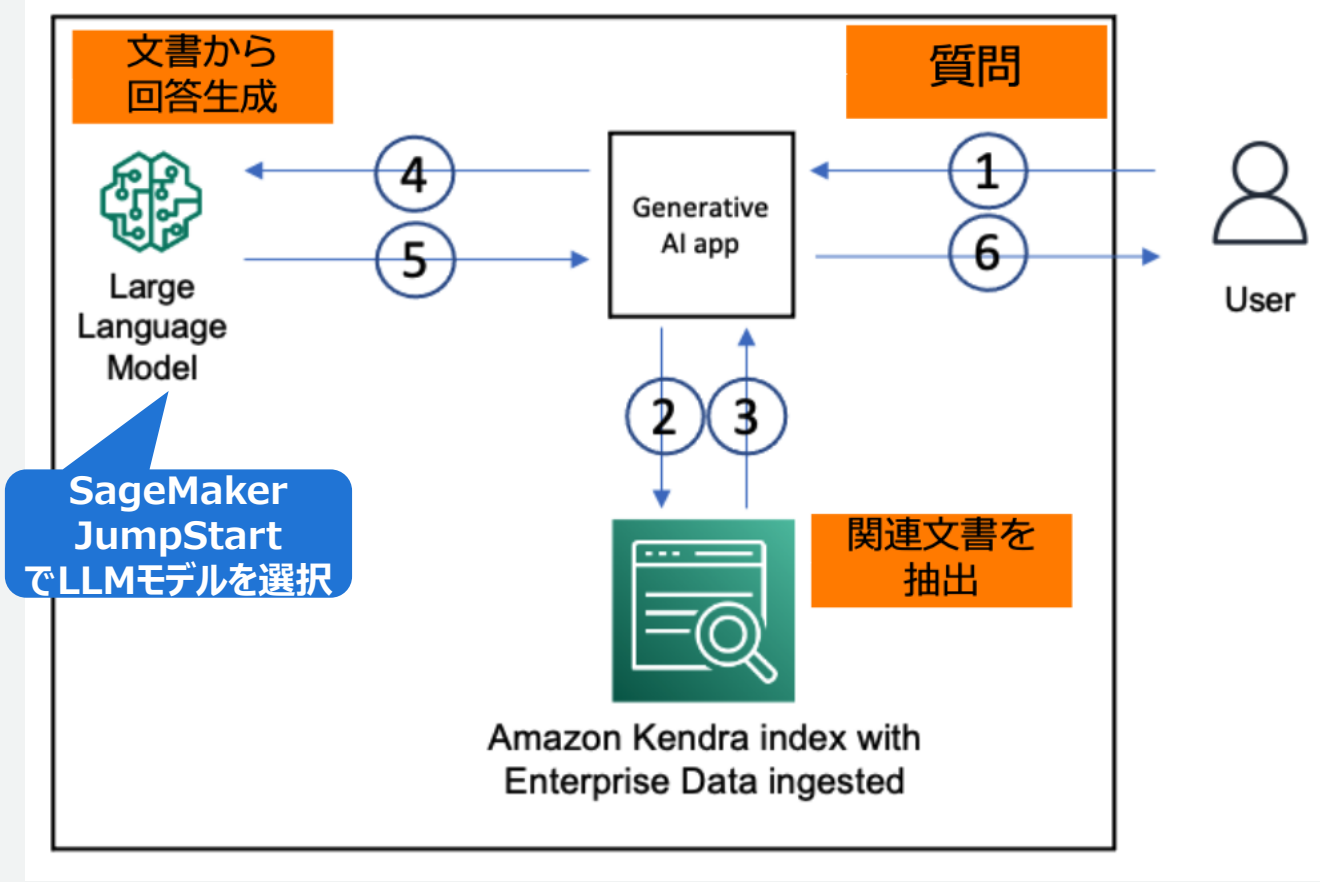
生成系 AI を活用したユーズケース1

ドキュメントからの回答文生成

出版業界での活用例
・自社書籍など言語アセットをデータソースとしたチャットボット
・編集部ごとに分割管理された社内資料をソースにした社内チャットボット

ユーズケース概要 Retrieval Augmented Generation (RAG) = 検索拡張生成

ユーザーからの質問に対して、社内ドキュメントを基にした回答を生成する



1. ユーザーが質問文をアプリケーションに入力

2. 3. アプリケーションは Amazon Kendra を利用して、関連するドキュ メントを抽出

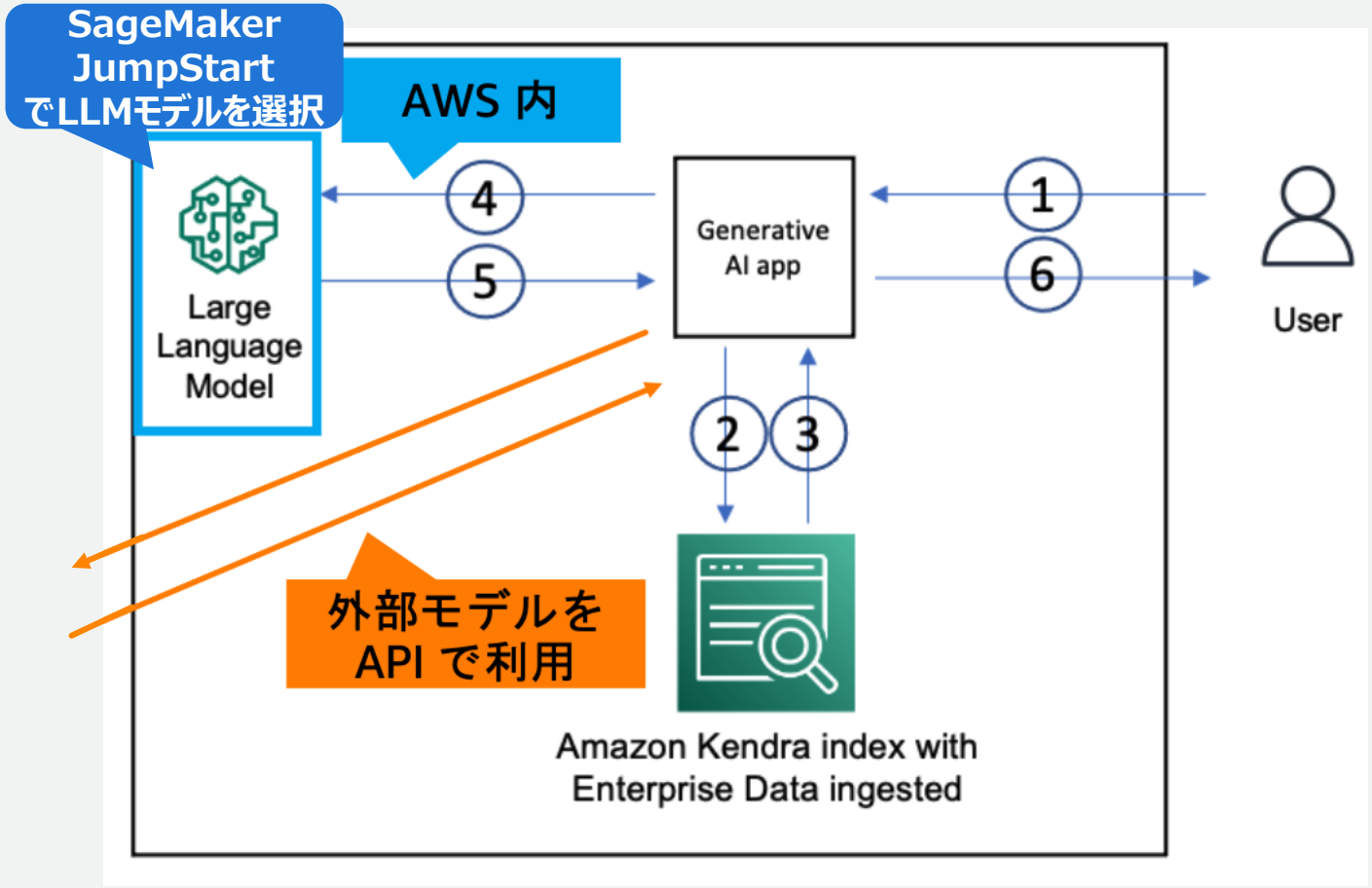
4. 5. アプリケーションは抽出された ドキュメントを生成系 AI にインプットし、生成系 AI が生成した回答を受けとる

6. ユーザーはアプリケーションから生成された回答と、参考ドキュメントリンクを得る

高精度な生成系 AI アプリケーションを Amazon Kendra、LangChain、大規模言語モデルを使って作る

生成系 AI の利用パターン

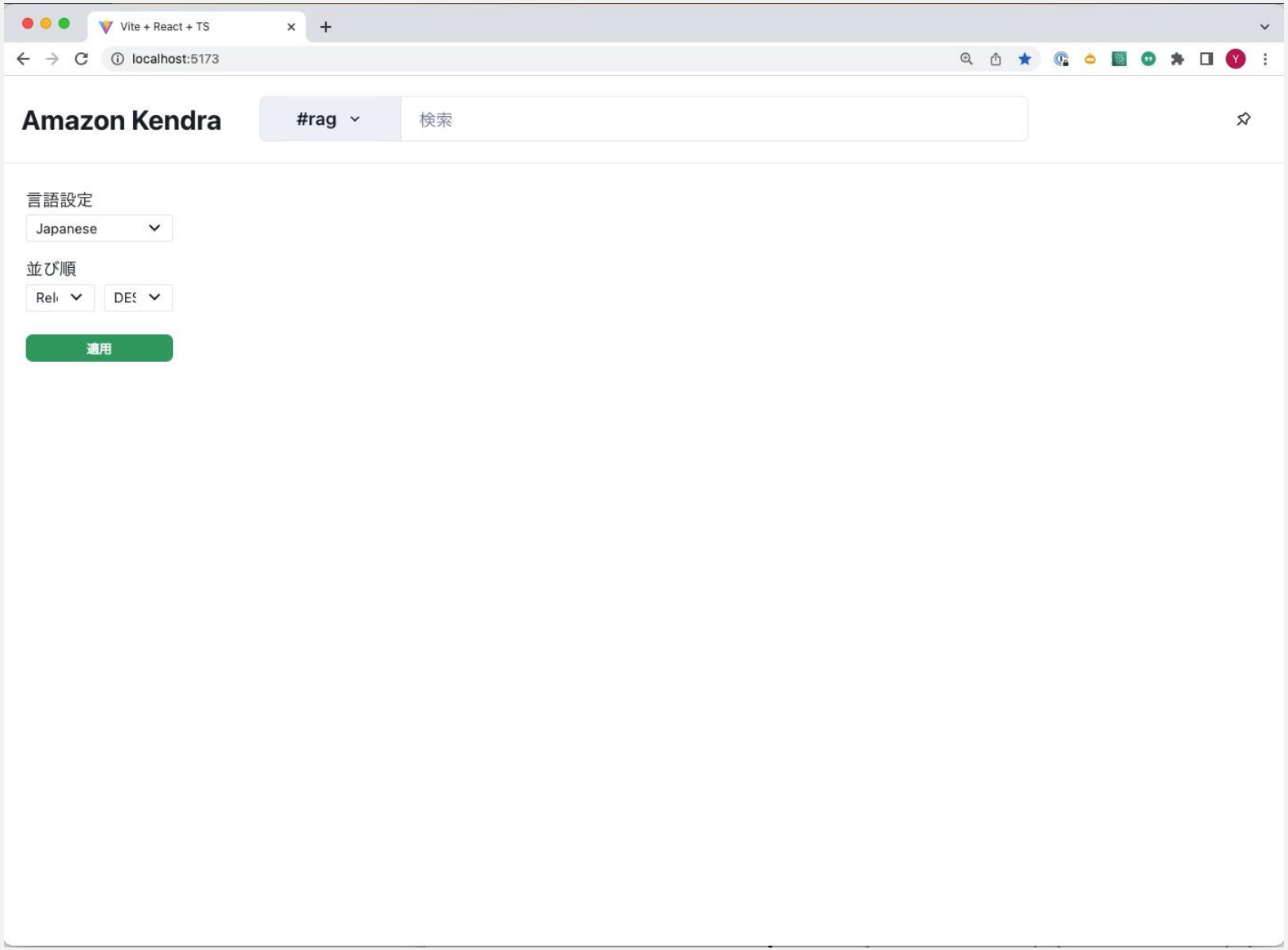
デプロイされた自社環境のモデルを利用したり、外部モデルを API 経由で利用する



利用パターン1
自社 AWS 環境内の SageMaker Jumpstart / カスタムモデルをエンドポイントにデプロイした生成系 AI を使う

利用パターン2
Bedrock もしくは外部の生成系 AI を、API 経由で使う

【デモ】ドキュメントからの回答文生成



LLMの限界

1. 事実と異なる、または古い情報
LLMの知識は、ある時点までのトレーニングデータに限定されているため、誤った情報や古い情報を生み出す可能性があります。

2. トレーニングデータとの過剰な類似
LLMは、学習データを過剰に使用したり、過剰に適合させたりすることがあり、学習した例とあまりにも類似したアウトプットを生成することがあります。

3. 理解・推論の欠如
LLMは、処理・生成するテキストを本当に理解しているわけではありません。学習データから学んだパターンや関連付けに依存しているため、誤った回答や無意味な回答をする可能性があります。

LLM をカスタマイズするには

- Zero-Shot Learning, Few Shot Learning （モデルの重み変更なしプロンプトエンジニアリング）
  - + CoT などのプロンプトエンジニアリングにより、パフォーマンスを向上が可能
  - - コンテキスト内に収めることができるものには限界がある
  - - 複雑なタスクでは、おそらく Fine Tuning が必要になる
- Instruction finetuning （教師データを用いた対話型モデルへ）
  - + シンプルでわかりやすい汎用言語モデル。
  - - 多くのタスクのデモを収集するのはコストがかかる
  - - LM の目的と人間の嗜好のミスマッチ
- Reinforcement Learning （人手のフィードバックでさらに高精度なモデルへ）
  - + 嗜好を直接モデル化し（言語モデリングなど）、ラベル付きデータを超えて一般化する
  - - RL は正しく行うのが非常に難しい
  - - 人間の嗜好は誤りやすい、人間の嗜好のモデルはさらに誤りやすい
- RAG （検索拡張生成）
  - + 外部から知識を得るので低コストに最新情報を反映できる
  - - 回答の精度は元データと検索の正確性に依存する

本ユーズケースで紹介したパターン

例：社内ドキュメントをデータソースとし信頼度の高いAIチャットボットを作る

Stanford CS224 Lecture 11: Prompting, Instruction Finetuning, and RLHF

Amazon Kendra について

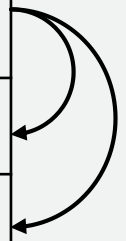
Kendra のメリットと機能

1. セットアップが簡単
   - AWS Managed な Index 、 Connector
2. 見つけやすい
   - ML ベースのドキュメントランキング、質問応答、該当箇所のハイライト、 FAQ
3. 分析と改善が簡単
   - Search Analytics Dashboard 、 Custom Document Enrichment 、チューニング、スケール、セキュリティ

LLM の選定について

- モデルごとに得意不得意がありコストも異なります。
- ユーズケースにより適切なサイズのモデルを使い分けることでパフォーマンスとコストを最適化できます。

モデル	パラメータ数	コスト	質疑応答1000件の解答にかかるコスト
gpt-3.5-turbo	非公開 (175B?)	入力: $0.0015 / 1K tokens 出力: $0.002 / 1K tokens	$0.09
OpenCALM 7B	7B	g5.xlarge の場合 $1.006 / 時	$0.0045
OpenCALM 3B	3B	g5.xlarge の場合 $1.006 / 時	$0.0025



20 倍

37 倍

※ 冒頭のデモは Rinna 3.6 B を g4dn.xlarge にデプロイしたもので月額400ドル以下で運用可能。

日本語大規模言語モデル OpenCALM の知識でクイズ王に挑戦する

今日から試せるドキュメントからの回答文生成

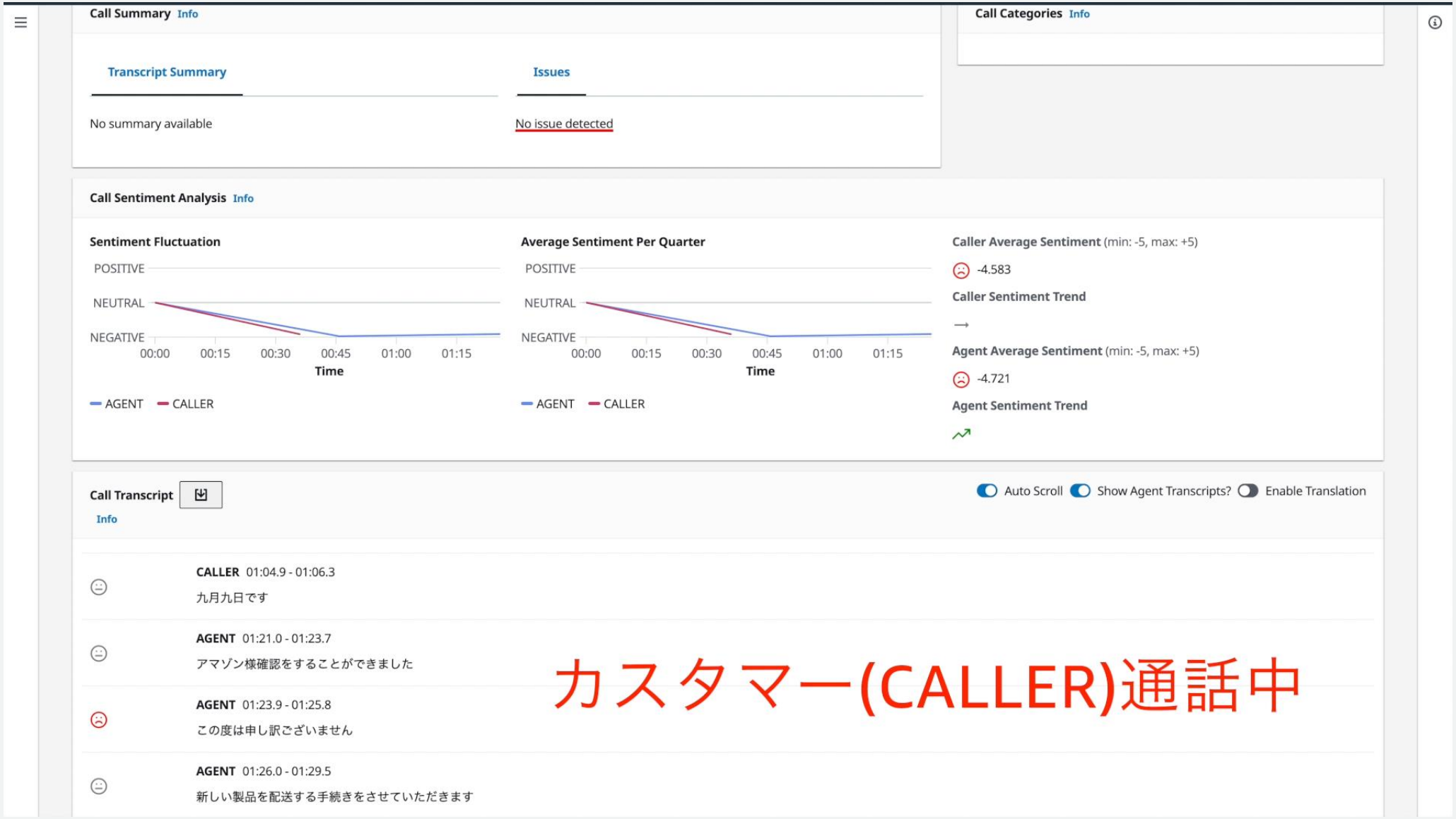
- まずは RAG で社内ドキュメント等を検索できるソリューションをデプロイ
  - 高精度な生成系 AI アプリケーションを Amazon Kendra、LangChain、大規模言語モデルを使って作る
- 軽量なモデルをデプロイする/望まれる回答例でモデルをチューニングする
  - Instruction Tuning サンプルコード

生成系 AI を活用したユーズケース2

音声要約

出版業界での活用例
・取材、インタビューのサマリの自動作成
・音声から記事草案の自動作成

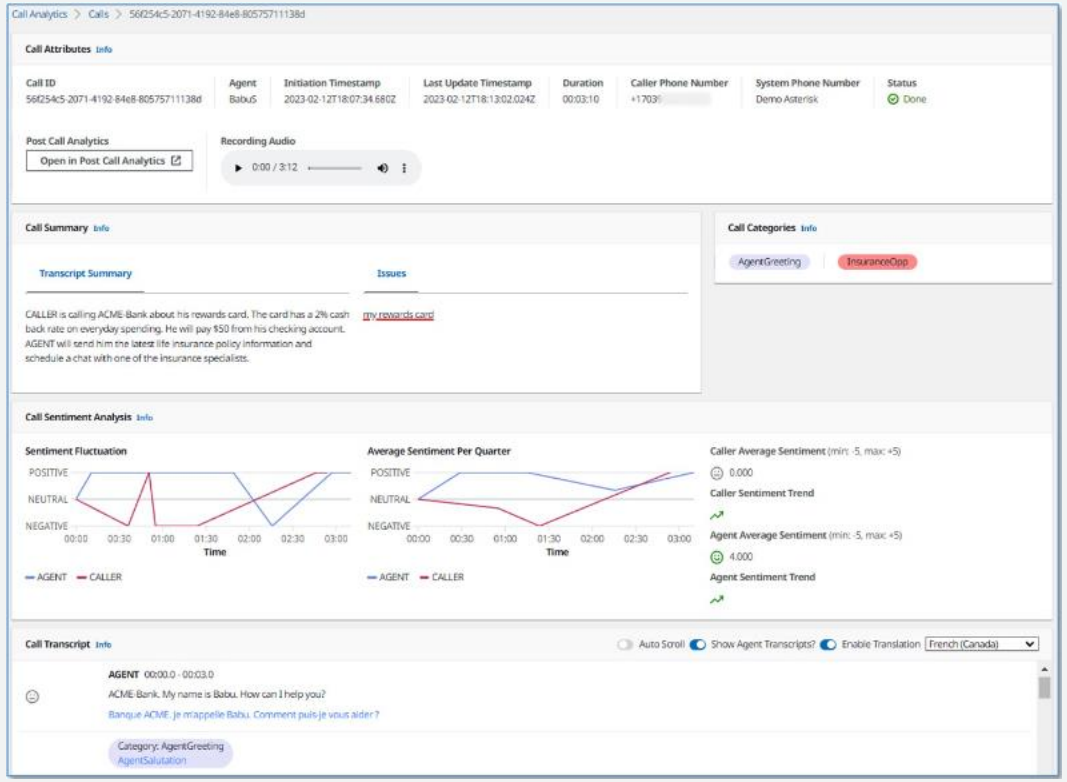
【デモ】通話要約



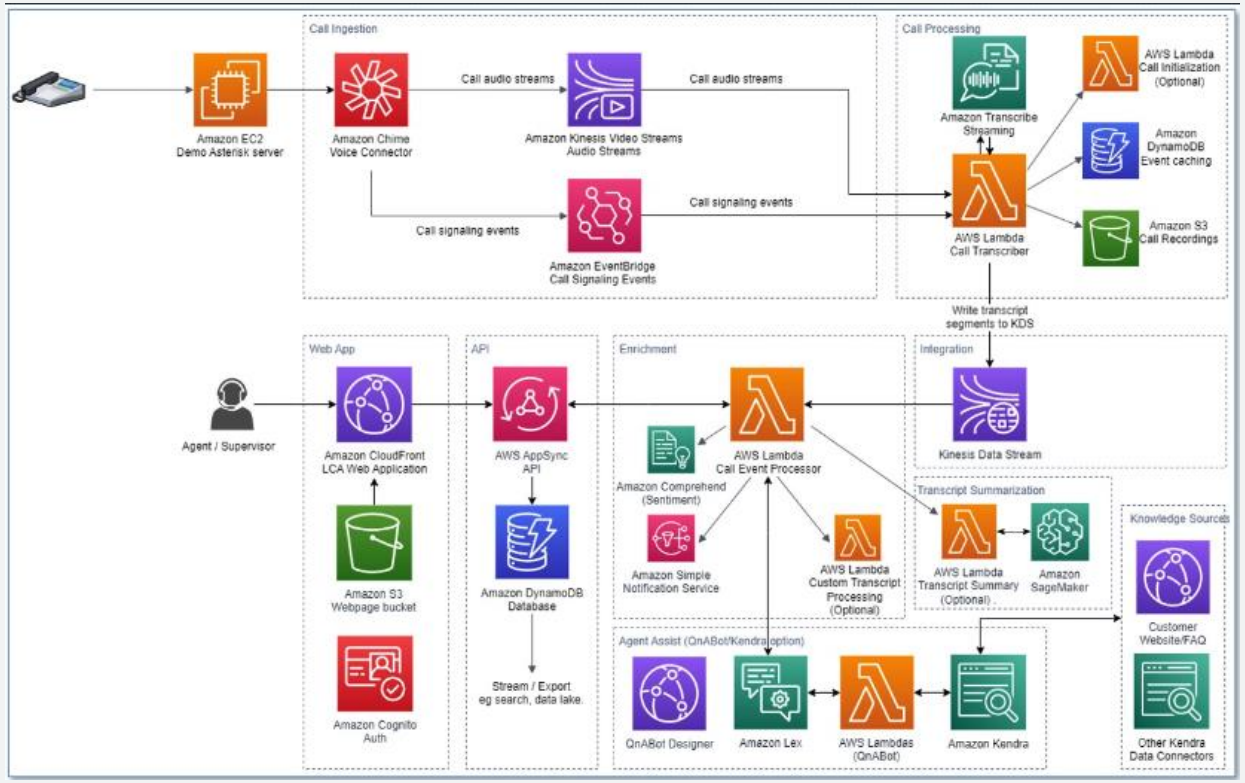
通話要約

通話のリアルタイム書き起こし、感情分析、参考ドキュメント提示、通話内容要約などの機能を提供

分析ダッシュボード



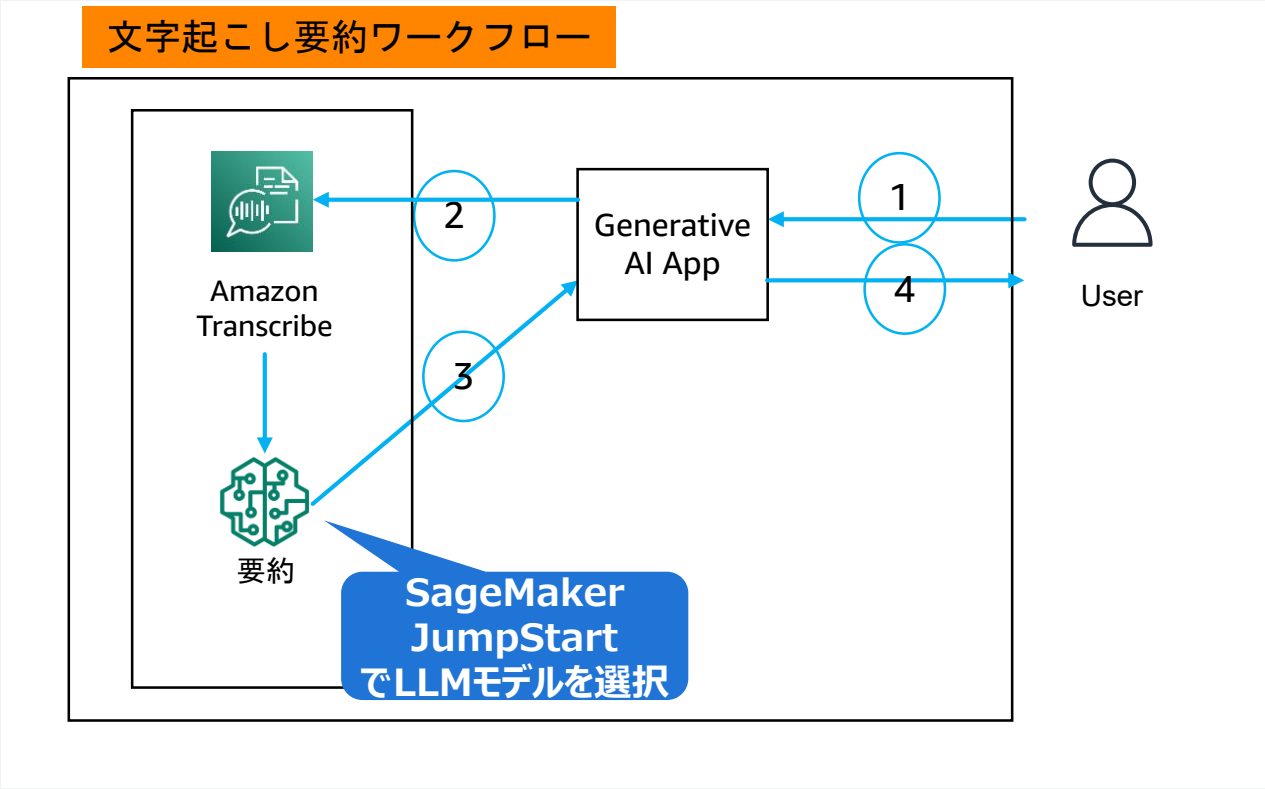
アーキテクチャ



[Amazon言語系AIサービスによるコンタクトセンターのライブ通話分析とエージェントアシスト]

ユーズケース概要 文字起こし要約

音声の文字起こしと要約を実現



1. ユーザーが音声ファイルをアップロード
2. 文字起こし要約ワークフローを実行
3. 文字起こしと要約を取得
4. ユーザーは文字起こしと要約を得る

今日から試せる音声要約

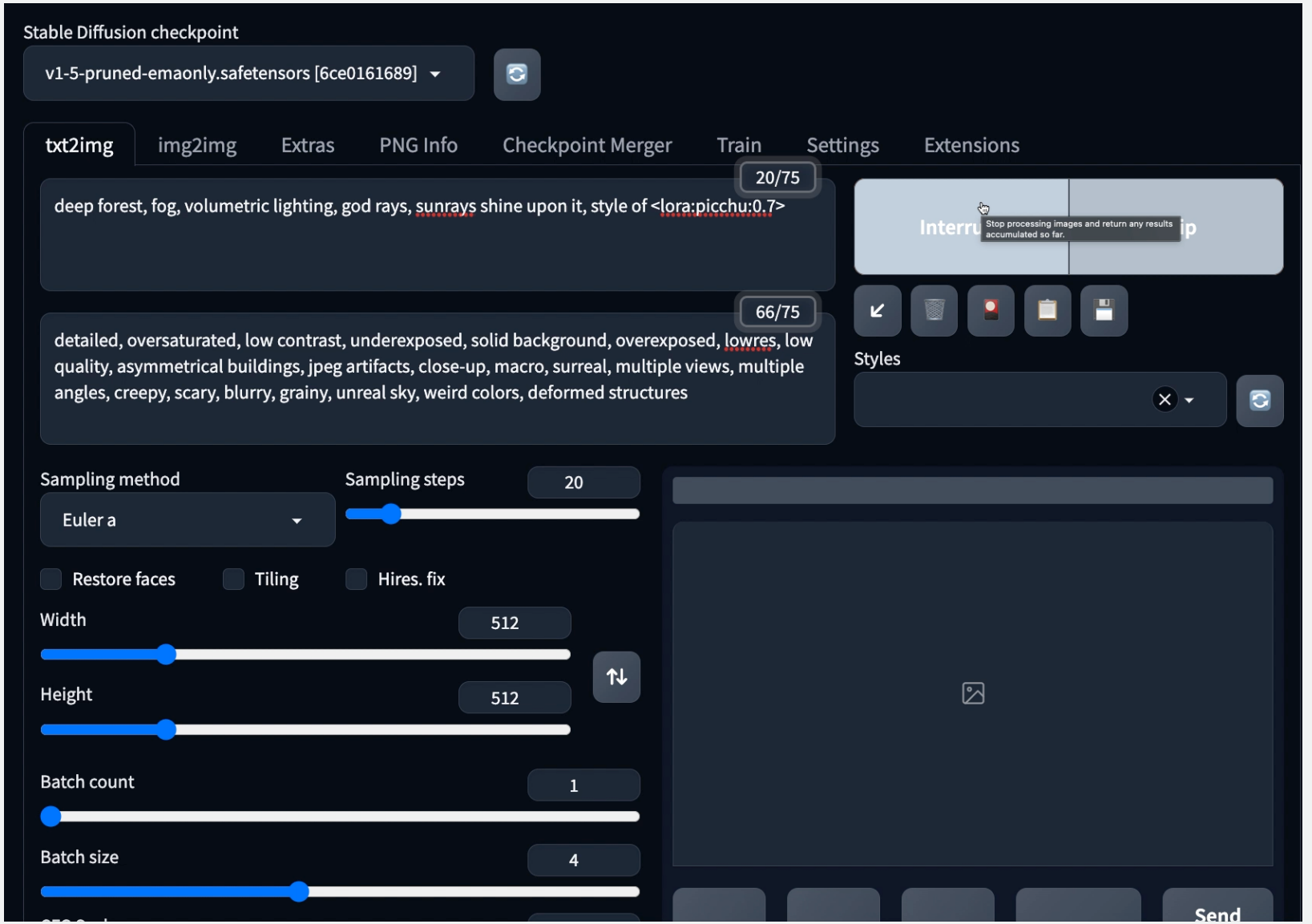
- 音声要約
  - Amazon言語系AIサービスによるコンタクトセンターのライブ通話分析とエージェントアシスト
  - ソリューションからの変更点についての解説

生成系 AI を活用したユーズケース3

画像素材の生成/編集

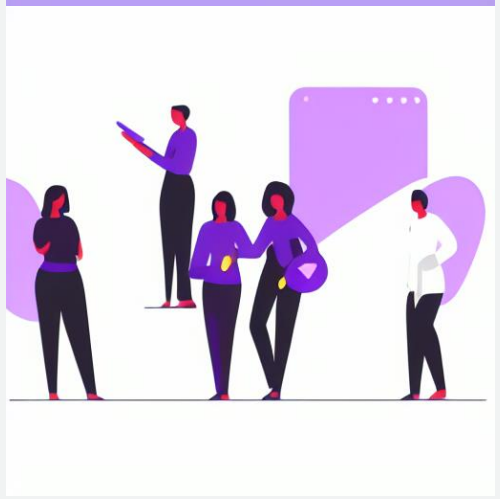
出版業界での活用例
・挿絵、背景画、広告クリエイティブなど画像の自動生成
・古いコンテンツのリマスター（高解像度化）

【デモ】画像素材の生成/編集

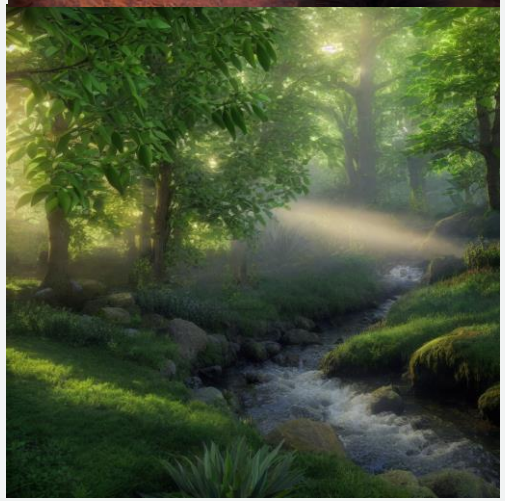


ユーズケース例：画像生成

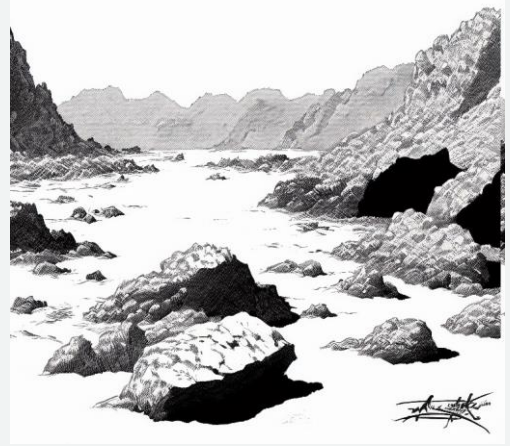
画像素材生成



アニメーション背景生成



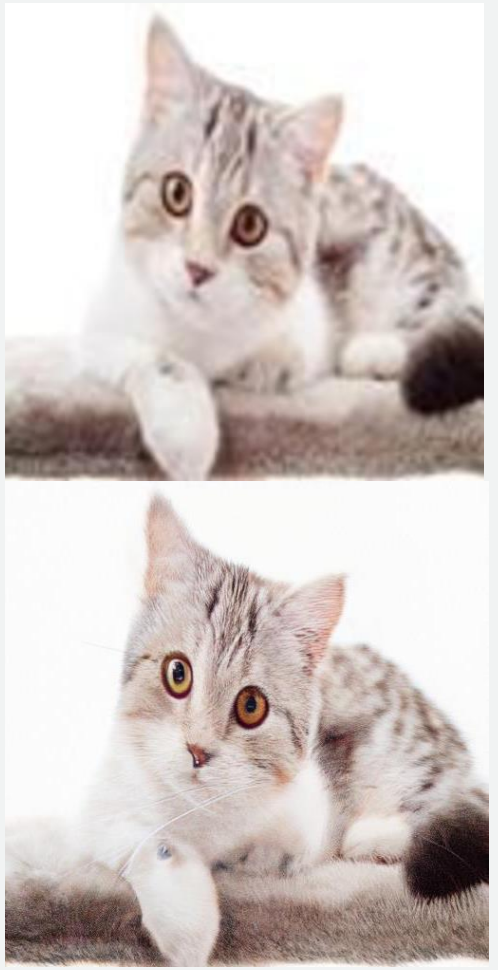
漫画背景生成



ユーズケース例：画像編集

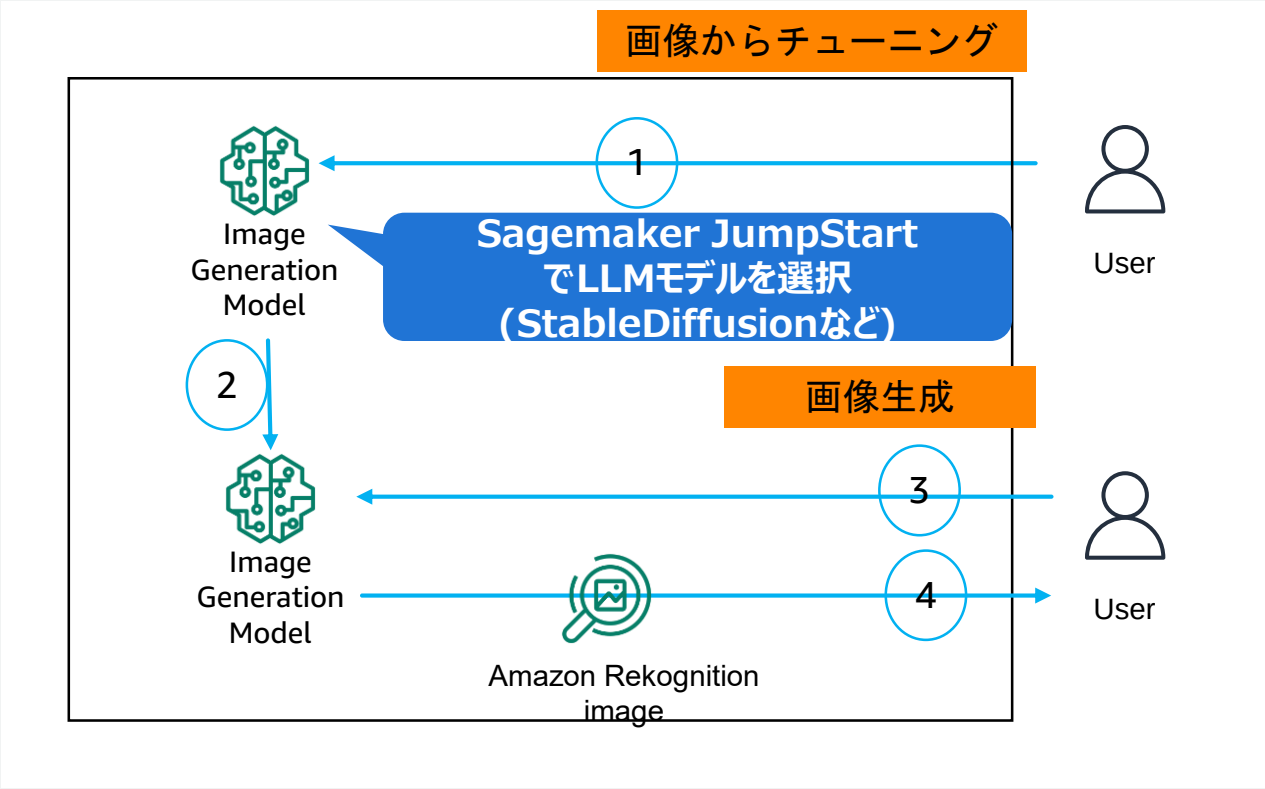
ファッション画像編集

高解像度化



ユーズケース概要 Image Generation

画像から特徴を学習し画像素材を生成/編集する



1. （オプション）ユーザーがアップロードした画像からスタイルや対象物を学習

2. （オプション）学習した差分をロード

3. 4. 文章/画像などから画像を生成。必要に応じて不適切な内容を検知。

事例：Canva オーストラリア

課題: 1 億のユーザーにスケールする安全な画像生成 AI を構築する

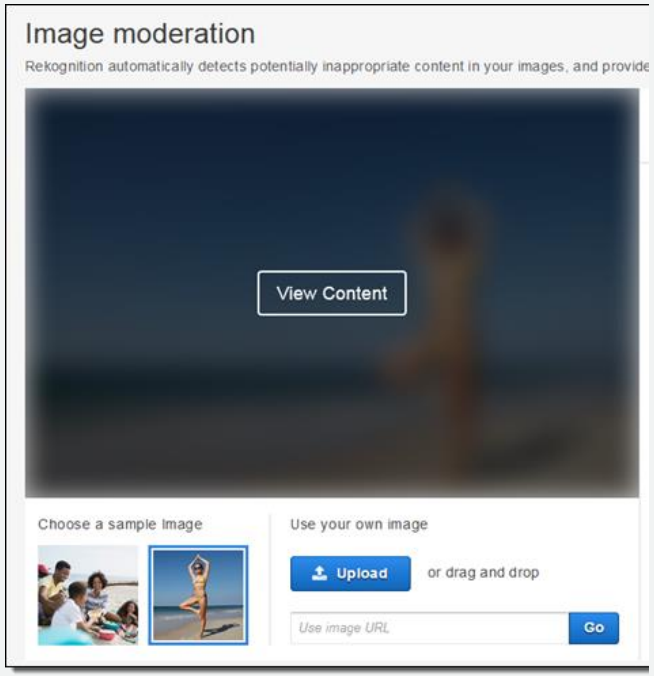
“パンフレットに使用する画像で、たくさんの種類のパンがテーブルに乗っている。” というテキストから画像を生成しパンフレットのデザインに入れている様子

Canva の機械学習部門ディレクター *Glen Pink* 氏
「 *AI* による画像生成は最近までおもちゃ以上のものではありませんでしたが、創造的なデザインプロセスの一部として実際使えるようになりました」

- Stable Diffusion を用いて商材画像を生成
- Stable Diffusion モデルを Amazon SageMaker の Real-time inference でホスト
- 数秒で独自の高品質な画像生成、スケーラブルなシステム運用

Canva が Amazon SageMaker と Amazon Rekognition を使用し 1 億ユーザーにテキストから画像を生成する AI を提供した方法

事例：Canva 生成系 AI の出力結果の不適切画像検査に Amazon Rekognition を活用



有害なコンテンツを作成したり、第三者の権利(著作権や商標など)を侵害しないこと

リリースするスピードだけでなく、ユーザーの信頼と安全性も重要な問題

AI による攻撃的な画像生成をチェックするため 24 時間体制で数百人のモデレーターを雇う必要があると考えていた

Amazon Rekognition モデレーション機能の活用

Canva は 3 週間以下で 1 億人のユーザーへ画像生成機能を提供

[Canva が Amazon SageMaker と Amazon Rekognition を使用し 1 億ユーザーにテキストから画像を生成する AI を提供した方法]

今日から試せる画像生成/編集

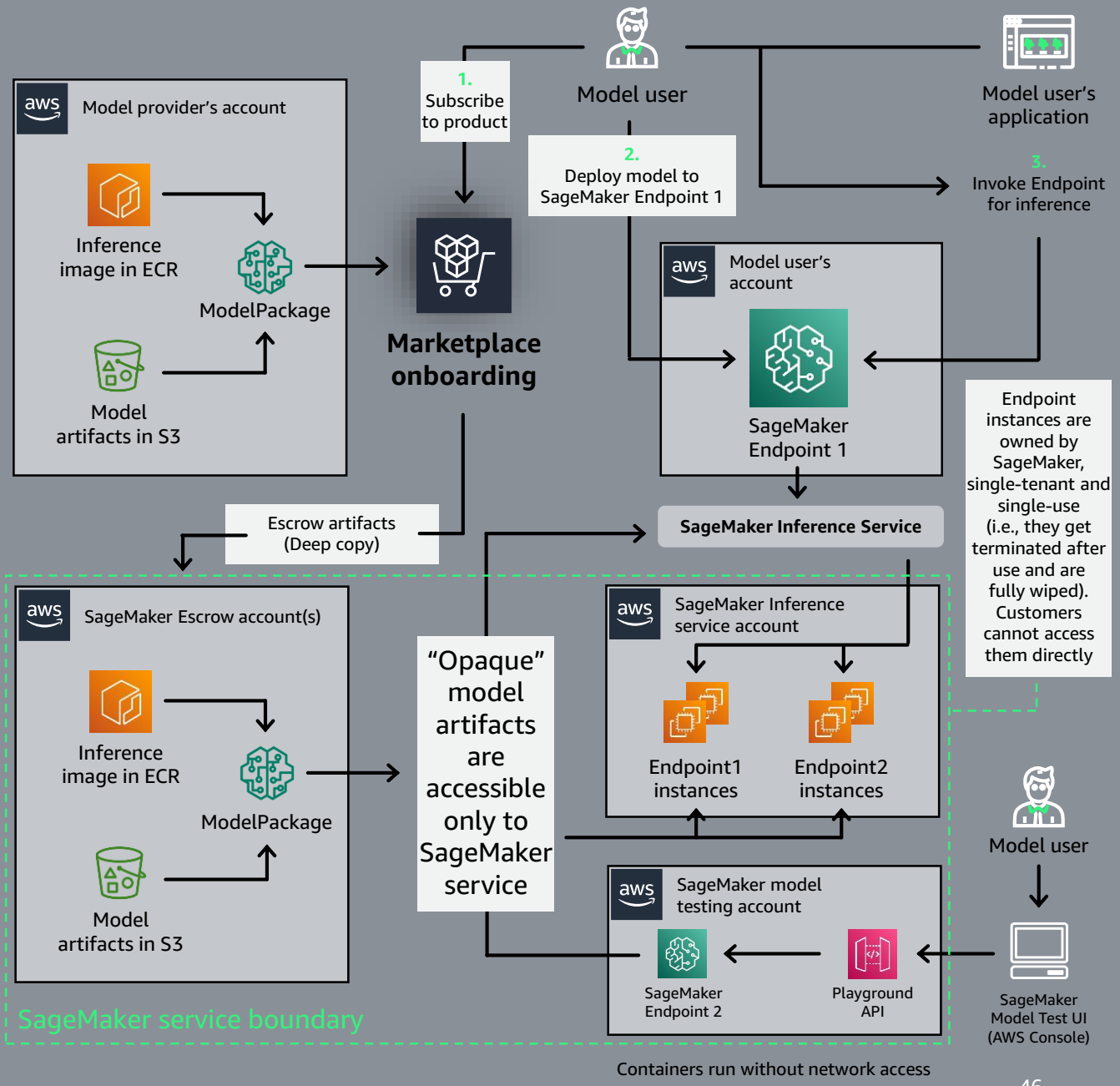
- まずは何ができるかの検証から
  - Stable Diffusion Web UI on AWS
- システムに組み込みスケールさせる
  - Stable Diffusion on SageMaker JumpStart
    - たった数枚の画像で Stable Diffusion をファインチューニングできる効率的な Amazon SageMaker JumpStart の使い方
    - Stable Diffusion で画像の部分的な差し替えを行う環境を、 Amazon SageMaker JumpStart で簡単に構築する
  - Extension for Stable Diffusion Web UI on AWS

学習データのデータ保護について

SageMaker でのデータ保護

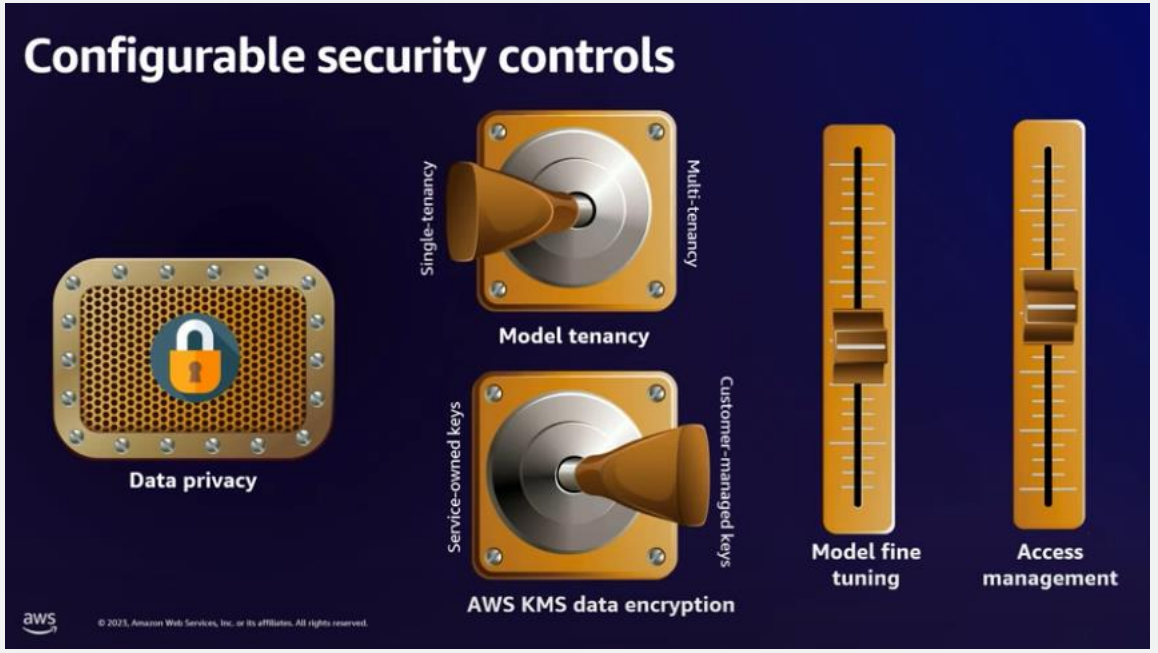
お客様のデータも
お客様専用にチューニングしたモデルも
お客様AWS環境内でセキュアに管理可能
(大元の基盤モデルへフィードバックされることはない)

- プロプライエタリ基盤モデルのパッケージとエンドポイントは SageMaker 仲介アカウント （Escrow Account）にホストされます。
- コンテナによるアウトバウンドのNWアクセスはございません。お客様のデータとモデルプロバイダーの知財は保護されています。
- JumpStart が提供する基盤モデルの更新・学習にお客様データが使用されることはありません



Amazon Bedrock のセキュリティ

- お客様のデータ（プロンプト、レスポンス、Fine-tune されたモデル）はお客様ごとに隔離され、リージョンから出ることはありません。また、AWS KMS 及び TLS 1.2 以上により暗号化され保護されます。
- 基盤モデルの更新・学習にお客様データが使用されることはありません



re:Inforce 2023 – Securely build generative AI apps & control data with Amazon Bedrock (APS208)

早速始めるために

AWS で生成系 AI を動かすには

再掲

- Amazon Bedrock（生成AI開発/限定プレビュー中）
  - 基盤モデルを選択し生成系 AI アプリをサーバレスで開発
- Amazon SageMaker（機械学習特化のPaaS）
  - SageMaker JumpStart ではモデルを SageMaker 上にコピーし Fine Tuning 可能
- Amazon EC2（IaaS/4月に新インスタンス発表）
  - EC2上で独自開発する場合。学習/推論特化のインスタンスを4月に発表。

ユースケースに応じてアプリ開発しやすいクラウド環境を用意

再掲

基盤モデルが SageMaker JumpStart でご利用いただけます

公開済みモデル

stability.ai

モデル

テキスト→画像
高精細化

タスク

テキスト入力から
フォト・リアルな
画像生成
生成画像の品質改善

特徴

Stable Diffusion 2.1
ファイン・チューン対応

alexa

モデル

AlexaTM
20B

タスク

機械翻訳
質問回答
テキスト要約
注釈付与
データ生成

Hugging Face

モデル

Flan T-5 models
(8 種類)
DistilGPT2, GPT2
Bloom models
(3 種類)

タスク

機械翻訳
質問回答
テキスト要約
注釈付与
データ生成

プロプライエタリ・モデル

co:here

モデル

Cohere
generate-med

タスク

テキスト生成
情報抽出
質問回答
テキスト要約

Lighton

モデル

Lyra-Fr
10B

タスク

テキスト生成
キーワード抽出
情報抽出
質問回答
テキスト要約
意味分析
(Sentiment analysis)
テキスト分類

AI21labs

モデル

Jurassic-1
Grande 17B

Tasks

テキスト生成
長文生成
テキスト要約
言い換え
チャット
情報抽出
質問回答
テキスト分類

今日から試せるドキュメントからの回答文生成

再掲

- まずは RAG で社内ドキュメント等を検索できるソリューションをデプロイ

  高精度な生成系 AI アプリケーションを Amazon Kendra、LangChain、大規模言語モデルを使って作る

- 軽量なモデルをデプロイする/望まれる回答例でモデルをチューニングする
  - Instruction Tuning サンプルコード

今日から試せる音声要約

再掲

- 音声要約

  Amazon言語系AIサービスによるコンタクトセンターのライブ通話分析とエージェントアシスト

  - ソリューションからの変更点についての解説

今日から試せる画像生成/編集

再掲

- まずは何ができるかの検証から
  - Stable Diffusion Web UI on AWS
- システムに組み込みスケールさせる
  - Stable Diffusion on SageMaker JumpStart

    たった数枚の画像で Stable Diffusion をファインチューニングできる効率的な Amazon SageMaker JumpStart の使い方

    Stable Diffusion で画像の部分的な差し替えを行う環境を、 Amazon SageMaker JumpStart で簡単に構築する

  Extension for Stable Diffusion Web UI on AWS

まとめ

まとめ

- ✓**AWS 上で生成系 AI を活用できる環境が既に整っている**
  - ✓SageMaker JumpStart がおすすめ！AWS 上にデータがあればそれも活かせます。
- ✓**複数の基盤モデルを取り揃えており様々なニーズに対応可能**
  - ✓大規模言語モデルの Claude、画像生成モデルの Stable Diffusion など実績あるモデルを AWS 上ですぐ使えます！
- ✓**大事なコンテンツをしっかりと保護可能**
  - ✓お客様のデータもお客様専用にチューニングしたモデルもお客様 AWS 環境内でセキュアに管理可能です！

Thank you!