© 2024, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Amazon Bedrock Re:Cap

Tanimiya Yusuke

Partner Sales Solutions Architect
Amazon Web Services Japan

© 2024, Amazon Web Services, Inc. or its affiliates.

自己紹介

谷宮　悠介

アマゾンウェブサービスジャパン
パートナーセールスソリューションアーキテクト

製造分野においてパートナー様とAWSの協業拡大のための技術支援を行っています。

生成 AI 活用アプリケーションにおける技術支援も行なっています。

好きな AWS サービス
Amazon Bedrock

© 2024, Amazon Web Services, Inc. or its affiliates.

アジェンダ

1. Amazon Bedrockとは
2. 新しい基盤モデルの追加・カスタムモデルへの対応
3. Agents for Amazon Bedrock
   Knowledge base for Amazon Bedrock
4. Model evaluation on Amazon Bedrock
   Guardrails for Amazon Bedrock
5. その他のアップデート

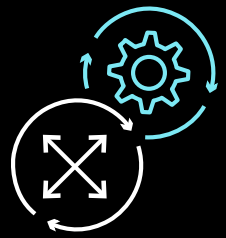
© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Bedrock とは

© 2024, Amazon Web Services, Inc. or its affiliates.

生成 AI を利用するお客様が重視すること

多くの基盤モデルから最適なものを選びたい



アプリケーションを素早く立ち上げたい

入出力データを安全に扱いたい

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Bedrock

基盤モデルによる 生成 AI アプリケーションを構築してスケーリングする最も簡単な方法

- 単一の API で主要な基盤モデルにアクセス
- 従量課金で使った分だけ支払い
- 様々なモデルを利用可能
- 東京リージョンでも利用可能
- セキュリティ・プライバシー・安全性

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Bedrock は主要な基盤モデルをサポートしています

AI21labs

Jurassic-2

文脈に応じた回答、要約、言い換え

ANTHROP\C

Claude & Claude Instant

要約、複雑な推論、ライティング、コーディング

cohere

Command & Embed

テキスト生成、検索、分類

Meta

Llama 2

対話のユースケースと言語タスク

stability.ai

Stable Diffusion XL

高品質の画像とアート

amazon

Amazon Titan

要約、画像とテキストの生成と検索、Q&A

aws

© 2024, Amazon Web Services, Inc. or its affiliates.

新しい基盤モデルの追加・カスタムモデルへの対応

© 2024, Amazon Web Services, Inc. or its affiliates.

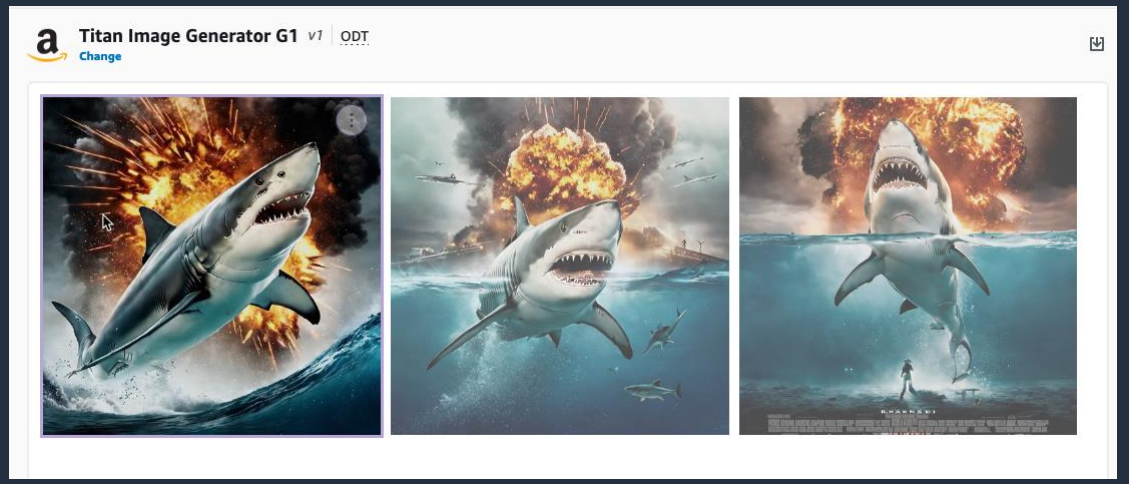
新しい基盤モデルが利用可能に

- Amazon Titan Image Generator (Preview)
- Amazon Titan Multimodal Embeddings
- Amazon Titan Text Lite and Express (GA)
- Added Llama2 70B
- Stable Diffusion XL 1.0 (GA)
- Claude 2.1 from Anthropic

aws
© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Titan Image Generator

- スタジオ品質のリアルな画像を大量かつ低コストで作成可能
  - 独自に用意した画像でモデルをファインチューニングすることで、ユースケースに適した画像生成を行うこともできる
  - Titan Image Generator で生成した画像には、目に見えない透かしが含まれており、AI が生成した画像を識別することができる
  - Txt2img だけでなく以下の機能も利用可能
    - Image-to-image
      - inpainting、masking、outpainting



“Movie poster featuring a shark with an explosion in the background, intense action, high-stakes drama, cinematic lighting”

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Titan Multimodal Embeddings

- 画像とテキストの組み合わせを入力しベクトル表現に変換。複数の情報に基づくベクトル検索を容易に
  - 以下の例のように、画像をテキストで検索するようなことも可能

    * 画像は 5MB 以内、テキストは 128 トークン

```
{
"inputText": "from Abyss",
"inputImage": "poster.jpg"
}
```

Titan Multimodal Embeddings

OpenSearch

Titan Multimodal Embeddings

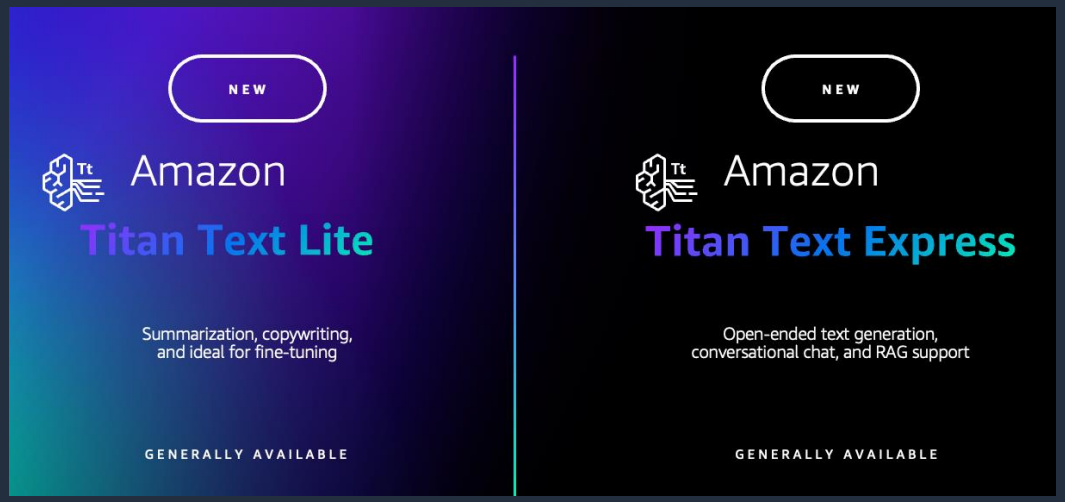
Exploding background, giant shark
(背景が爆発してて、巨大なサメがいる)

ユーザー

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Titan Text Lite and Express

- Titan Text Lite / Express が GA
  - Amazon Titan Text Express: 8,000トークンに対応
    - 英語に最適化されているが100+言語対応がプレビュー中
  - Amazon Titan Text Lite: 4,000トークンに対応
    - Titan Textで最も高速で、英語のタスクに適する

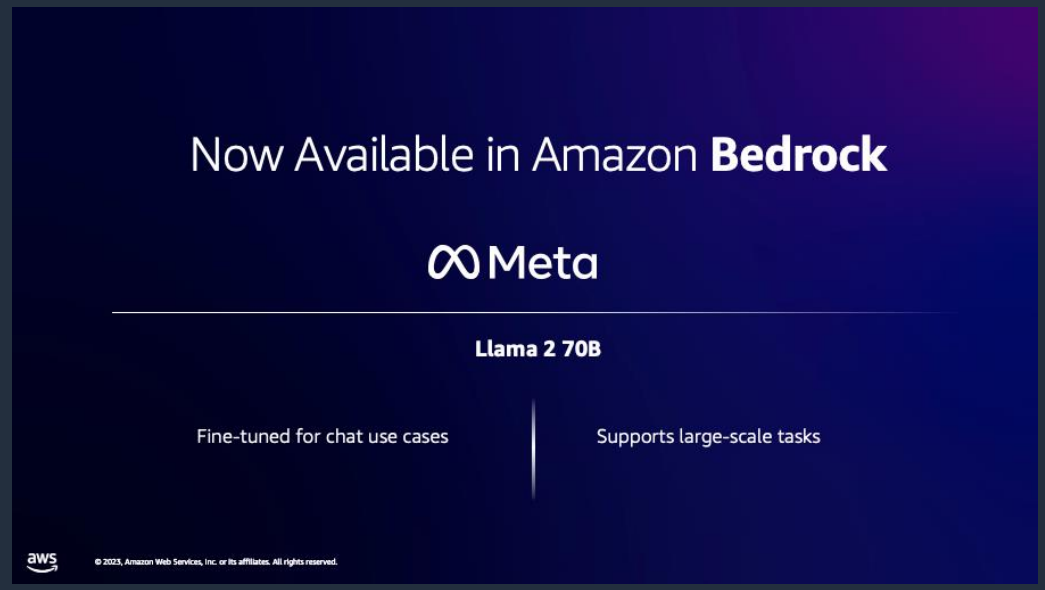


© 2024, Amazon Web Services, Inc. or its affiliates.

Meta Llama 2 70B

- Amazon Bedrock で Llama2 70B が利用可能に
  - 従量課金で Llama2 70B を利用できる
  - 13B と比べてより複雑なタスクを解決できる
  - Custom Models での Fine-tuning にも対応

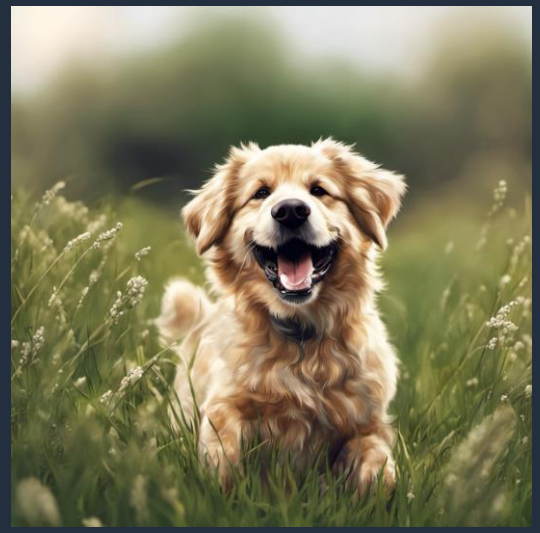
Fine-tuning したモデルの利用には
Provisioned Throughput の購入が必要
（Llama2 70B は 23.50 USD / Hour ）



© 2024, Amazon Web Services, Inc. or its affiliates.

Stability AI Stable Diffusion XL 1.0

- Stability AI 社の Stable Diffusion XL 1.0(SDXL1.0)が一般利用開始
  - SDXL 0.8 に比べて優れたコントラスト、照明、陰影を、ネイティブの 1024x1024 の解像度で表現
  - 従来の Stable Diffusion モデルと異なり、masterpiece（傑作）などの高品質化プロンプトを利用しなくても、高品質な画像を生成可能
  - API から以下の機能も利用できる
    - Image-to-image
    - Image-to-image/upscale
    - Image-to-image/masking



“a realistic happy dog playing in the grass”

© 2024, Amazon Web Services, Inc. or its affiliates.

Anthropic Claude 2.1

- Amazon Bedrock で Claude 2.1 が利用可能に
  - 200k の context windows
  - ハルシネーションの割合を2分の1に削減
  - 長いドキュメントの精度の向上
  - システムプロンプトが利用可能に
  - 東京リージョンでも利用可能

© 2024, Amazon Web Services, Inc. or its affiliates.

補足: Anthropic Claude 2.1 のシステムプロンプトとは

- "Human:" より手前に記載したプロンプトがシステムプロンプトとして扱われる
  - システムプロンプトを利用することで、Claude がルールや指示に従う能力を高めることができる

あなたはファッション業界の専門家です。あなたの仕事は、ファッション業界に関する質問に答えたり、依頼されたタスクを実行したりして、ユーザーを支援することです。

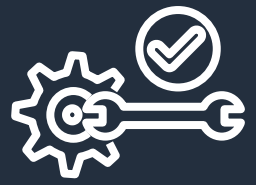
Human:ラグジュアリー商品をターゲットにするハンドバッグブランドの候補名を3つ挙げ、それぞれの候補について説明してください。

Assistant:

© 2024, Amazon Web Services, Inc. or its affiliates.

Custom Models

- Custom Models でモデルのファインチューニング / Continued pre-training が可能に（ Continued pre-training は Preview）



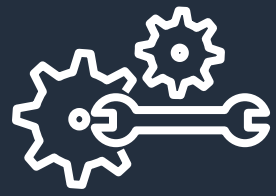
Fine tuning

PURPOSE

特定のタスクの精度を
最大限に高める

DATA NEED

少量の
ラベル付きデータ



Continued pre-training

PURPOSE

モデルに
ドメイン知識を与える

DATA NEED

大量の
ラベル無しデータ

aws
© 2024, Amazon Web Services, Inc. or its affiliates.

ファインチューニングと Continued pre-training の違い

{“prompt”: “日本の首都は？”, “completion”: “東京です"}
{“prompt”: “アメリカの首都は？”completion“: ”ワシントンです"}
{“prompt”: ”日本の次の元号は？“, ”completion“: “すみません、わかりません"}

{"input": "<raw text>"}
{"input": "<raw text>"}
{"input": "<raw text>"}

© 2024, Amazon Web Services, Inc. or its affiliates.

Agent for Amazon Bedrock Knowledgebase for Amazon Bedrock

© 2024, Amazon Web Services, Inc. or its affiliates.

生成 AI はなんでもできるのか？

- 生成 AI にできること
  - 文章やメディアの生成・要約・変更

- 生成 AI だけではできないこと
  - 学習データにない知識を使った回答作成
  - プログラムの実行

© 2024, Amazon Web Services, Inc. or its affiliates.

生成 AI はなんでもできるのか？

ユーザー

実家の両親にプレゼントを贈りたいので、選んで送ってください

© 2024, Amazon Web Services, Inc. or its affiliates.

Amazon Bedrock 単体での実行結果

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents for Amazon Bedrock

生成系 AI アプリケーションが社内のデータやシステムと連携しつつマルチステップでのタスクを完了できるようにします

1. 基盤モデルを選択
2. 基本的な指示を伝える
3. 関連するデータソースを選択
4. 実行可能なアクションを指定

| タスクを分解して調整 |

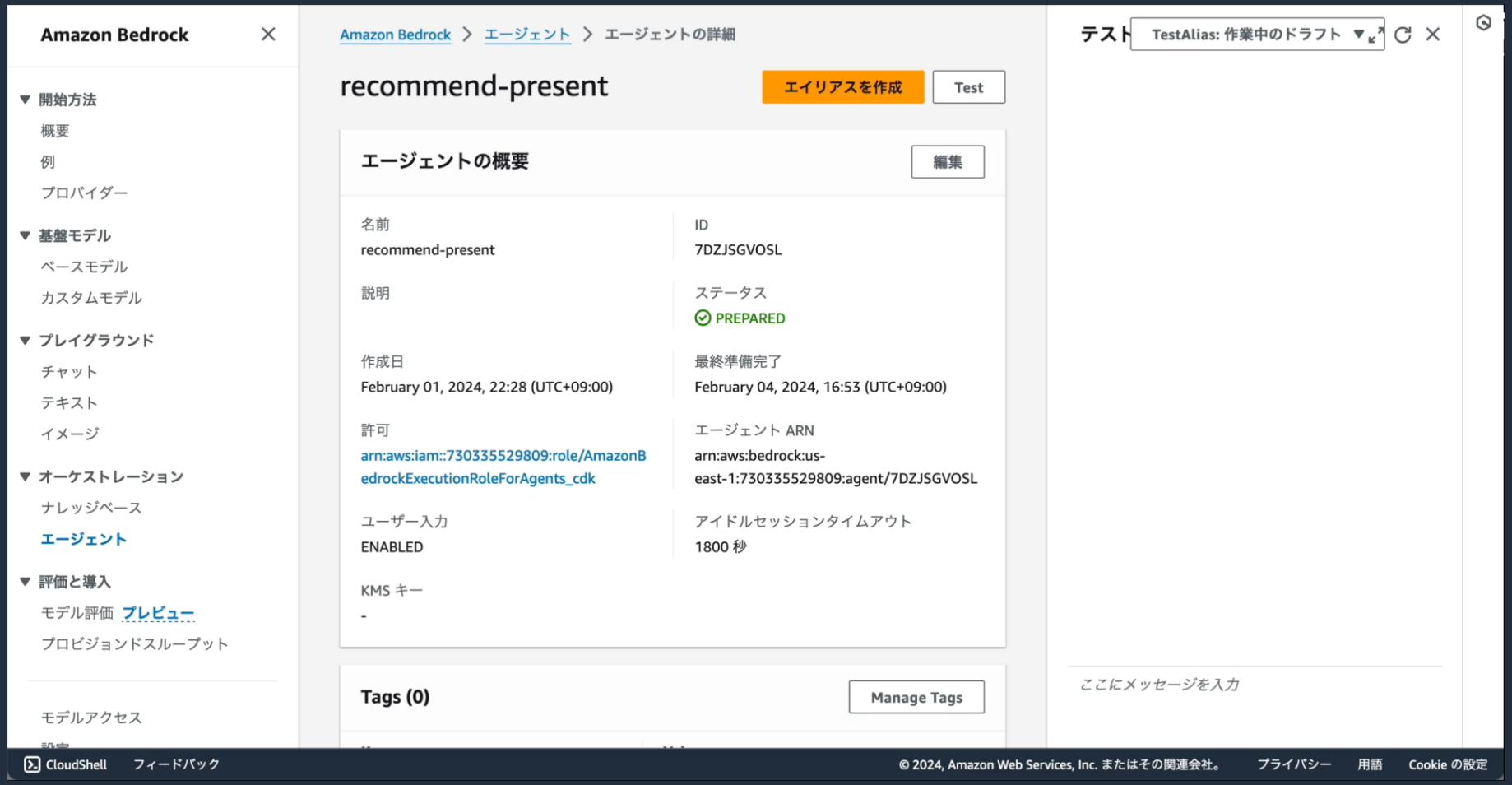
| RAG のための企業データへの安全なアクセスと取得 |

| ユーザーに代わって API 呼び出しを呼び出してアクションを実行する |

| FM の思考の流れを追跡し、エージェントプロンプトを変更する機能 |

© 2024, Amazon Web Services, Inc. or its affiliates.

Agent for Amazon Bedrockでの実行結果



© 2024, Amazon Web Services, Inc. or its affiliates.

Agent ができること– タスク分解

- 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する

ユーザー

実家の両親にプレゼントを贈りたいので、選んで送ってください

Agent (LLM)

実家の両親と質問者の関係性を確認しよう。
その後、なんのためにプレゼントを贈るのかを聞いてみよう。
プレゼントの内容は Amazon.co.jp から検索すればいいかな。
選んだものを提示して、OKなら購入してしまおう。

© 2024, Amazon Web Services, Inc. or its affiliates.

Agent ができること– 回答

- 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する

ユーザー

実家の両親にプレゼントを贈りたいので、選んで送ってください

Agent (LLM)

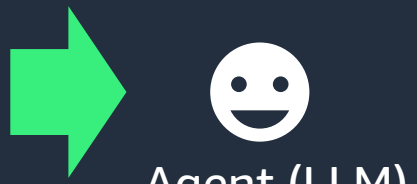
Amazon ギフトカードにしましょう。
50,000 円分購入して送信しておきました。

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents の使い方 - 指示を与える

- Agents に対して、どのように振る舞う必要があるか指示を与える

あなたはファイナンシャルアドバイザーです。質問文に入力された企業に関して、銘柄コード、最近の株価、財務諸表、その企業に関連したニュースを調べて、投資すべきか分析してください。回答は日本語でお願いします。



Agent (LLM)

aws
© 2024, Amazon Web Services, Inc. or its affiliates.

Agents の使い方 - 指示を与える

- Agents に対して、どのように振る舞う必要があるか指示を与える

あなたはファイナンシャルアドバイザーです。質問文に入力された企業に関して、銘柄コード、最近の株価、財務諸表、その企業に関連したニュースを調べて、投資すべきか分析してください。回答は日本語でお願いします。

Agent (LLM)

むむむ、
最新の株価の情報だったり、
企業ニュースは学習してないぞ

タスクをこなすための
知識や道具が必要

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents の挙動 – Action Group によるタスク処理

- ユーザーから入力されたインプットを元に、Action Group に定義された Lambda 関数を利用して回答を生成する

ユーザー

こんにちは！XXX の現在の株価を教えてください

Agent (LLM)

ユーザーからの問い合わせは、悪意があるものではないな。
文脈によると、XXX は企業の名前だな。
getStockTicker 関数で企業名から銘柄コードを取得して、
getStockPrice 関数で銘柄コードから株価を取得すれば回答できるな。

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents の使い方 - Action Groups を定義する

- Agents for Bedrock がタスク実行時に使うことができる道具として、Lambda 関数を定義することができる
  - 作成した関数のスキーマは OpenAPI フォーマットで定義する

getStockPrice 関数を使えば最新の株価情報を取得できるな。
銘柄コードを Ticker として引数に渡せばいいんだな。

Agent (LLM)

Action Group に定義した Lambda 関数

getStockPrice 関数

description: ”銘柄コードから株価を取得”
Required: ”ticker”
Ticker:
type: “string”
description: “銘柄コード”

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents の挙動 – Knowledge Base によるタスク処理

- ユーザーから入力されたインプットを元に、Knowledge Base から RAG（検索拡張生成）を行う

ユーザー

こんにちは！無料プランから有料プランに移行したいのですが
どうすればいいでしょうか？

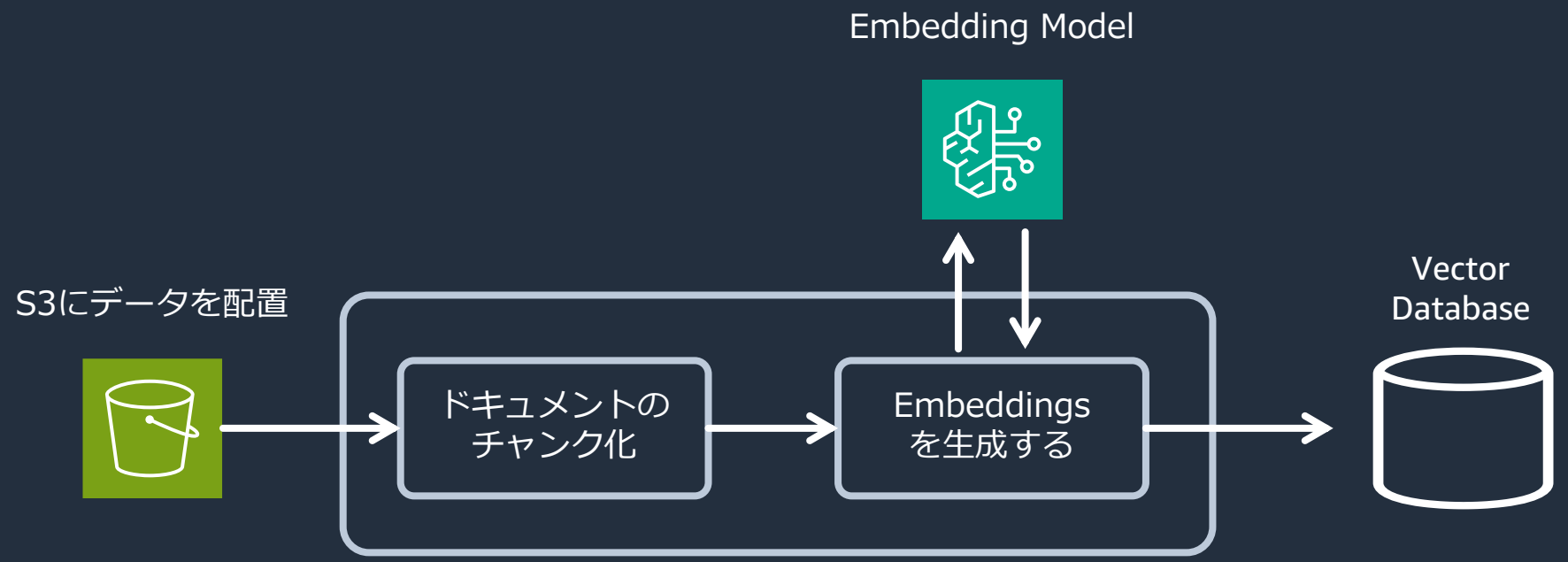
Agent (LLM)

ユーザーからの問い合わせは、悪意があるものではないな。
文脈によると、プランへのアップグレード方法について尋ねているな。
Knowledge Base の FAQ を検索すれば回答できるかもしれないな。

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents の使い方 - Knowledge Base を接続する

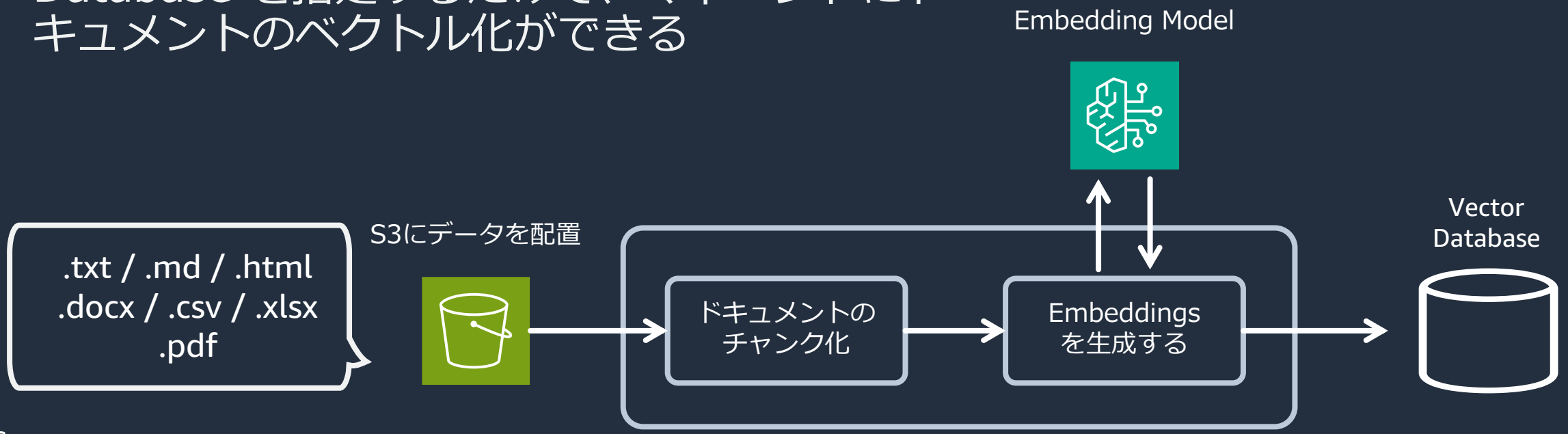
- Knowledge Base を Agent に参照させることで、拡張検索生成(RAG)をフルマネージドに実現可能
  - Agent は RAG が必要だと判断した際に、Knowledge Base から情報を参照し、回答を生成する



© 2024, Amazon Web Services, Inc. or its affiliates.

Knowledge base for Amazon Bedrock

- 基盤モデルとデータソースを組み合わせた拡張検索生成(RAG)をフルマネージドに実現可能に
  - データソースとなる S3 と格納先の Vector Database を指定するだけで、マネージドにドキュメントのベクトル化ができる



© 2024, Amazon Web Services, Inc. or its affiliates.

Knowledge base for Amazon Bedrock

- Vector database は、OpenSearch Serverless、Redis Enterprise Cloud、Pinecone から選択できる
  - Amazon Aurora と MongoDB のサポートも近日中に追加予定

© 2024, Amazon Web Services, Inc. or its affiliates.

Agents はどのように回答を生成するのか

- ユーザーの入力を受け、Agents は 3 つのステップで処理を行う
  - それぞれのステップは事前に定義されたプロンプトによって制御される

ユーザーから入力された内容を評価し、後続の Orchestration に渡すかを判定する

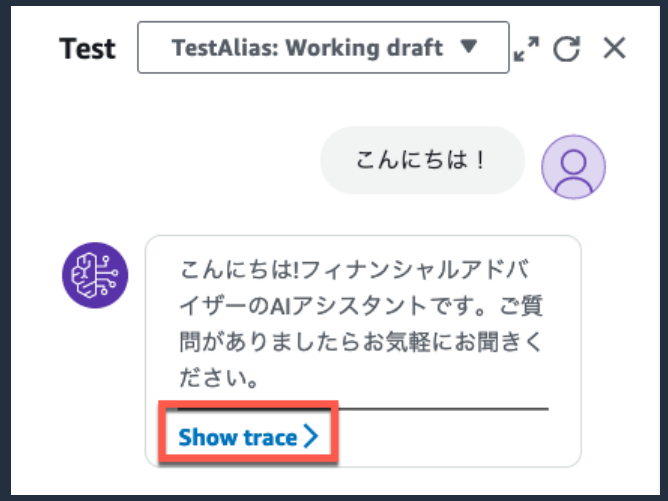
事前に定義された Action Group から、ユーザーの質問に回答するために必要な Lambda 関数を選択し、実行する

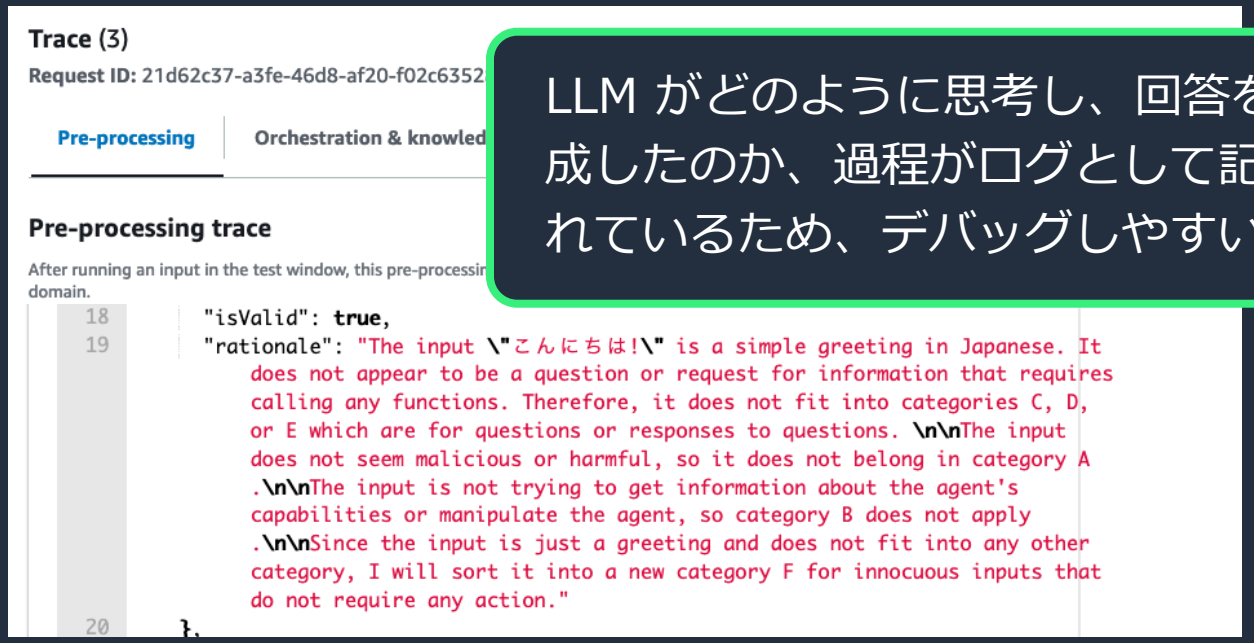
ユーザーに対する回答を生成する際に、Post-processing テンプレートに記載のプロンプトによって、最終的な調整を行う

© 2024, Amazon Web Services, Inc. or its affiliates.

Trace Event 機能によりデバッグが容易

- マネジメントコンソール上で Agents のテストが可能
  - Trace Event 機能によって、Agents がどのように思考したのか確認できる





LLM がどのように思考し、回答を生成したのか、過程がログとして記録されているため、デバッグしやすい

© 2024, Amazon Web Services, Inc. or its affiliates.

Model Evaluation on Amazon Bedrock Guardrail for Amazon Bedrock

© 2024, Amazon Web Services, Inc. or its affiliates.

責任ある AI （Responsible AI）

- AI システムの誤用・悪用・設計不良、その他 AI システムが直面したり引き起こしたりする意図しない悪影響を最小限に抑えられるように、AI システムを設計・開発・デプロイ・運用する枠組みおよび一連の原則
  - AI は、ビジネスにかつてない機会をもたらすと同時に、大きな責任をもたらす
  - AI システムからの出力は、人々の生活に実際に影響を与えるため、AI の倫理・データガバナンス・信頼・および合法性に関する考慮すべき観点が存在
  - AI を「信頼されるシステム」にするための考え方

© 2024, Amazon Web Services, Inc. or its affiliates.

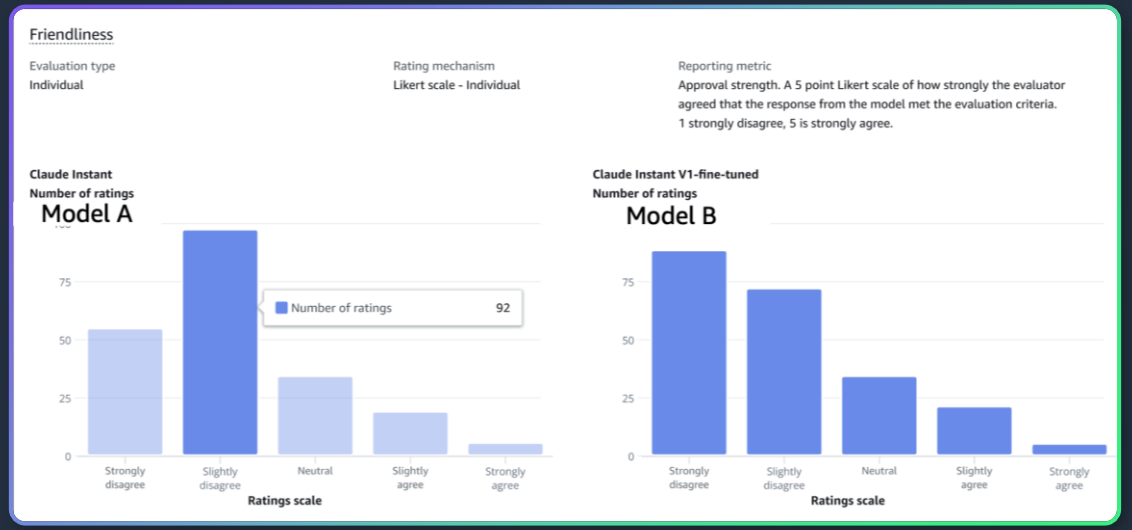
Model Evaluation on Amazon Bedrock(Preview)

自動評価方法または人間による評価
方法を選択

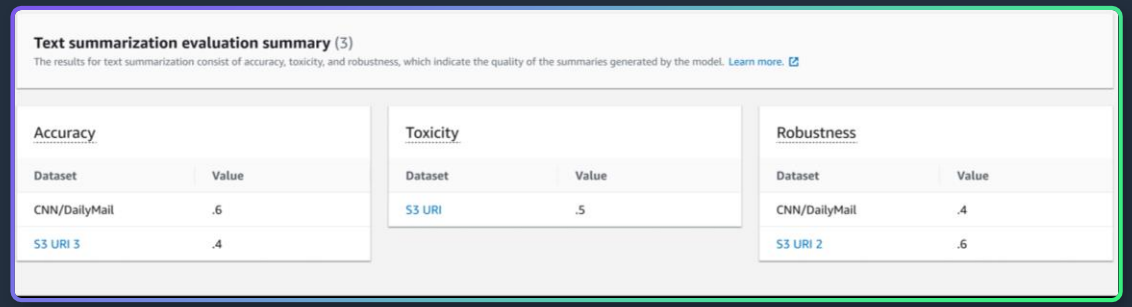
厳選されたデータセット、
または独自のデータセット

事前定義済みの指標とカスタム指標

人間による評価についてのレポート



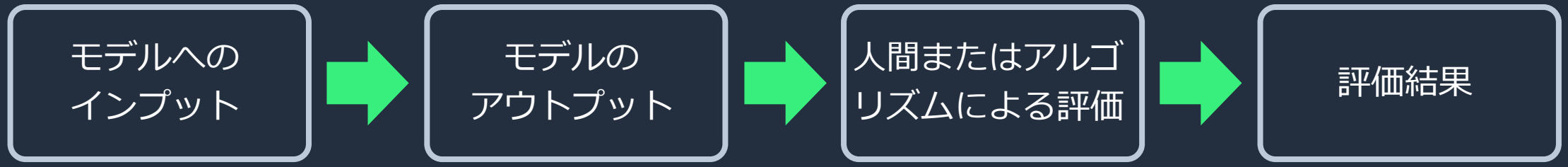
自動評価レポート



© 2024, Amazon Web Services, Inc. or its affiliates.

モデルを評価するとは？

- モデルにタスクを与え、そのアウトプットの内容を人間、またはアルゴリズムによって評価することで、選択したモデルがビジネスケースに適しているかを判断することができる



© 2024, Amazon Web Services, Inc. or its affiliates.

Automatic Model Evaluation

- 事前に用意された一般的なタスクについて、評価メトリクス、データセットを指定することで、自動的にモデルの評価を行うことができる
- 評価メトリクス
  - Accuracy:
    - モデルがどれだけ正確にタスクを実行できるか
  - Toxicity:
    - 生成するコンテンツに有害・不適切な言葉が含まれていないか
  - Robustness:
    - モデルがテキストの小さな変更にも動じず、同じような意味の文をしっかりと理解して適切な反応ができるか

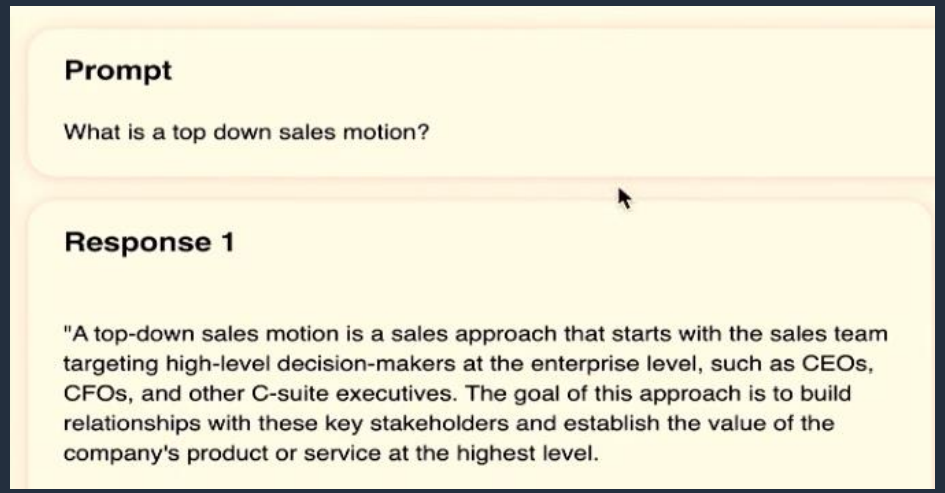
対応しているタスク

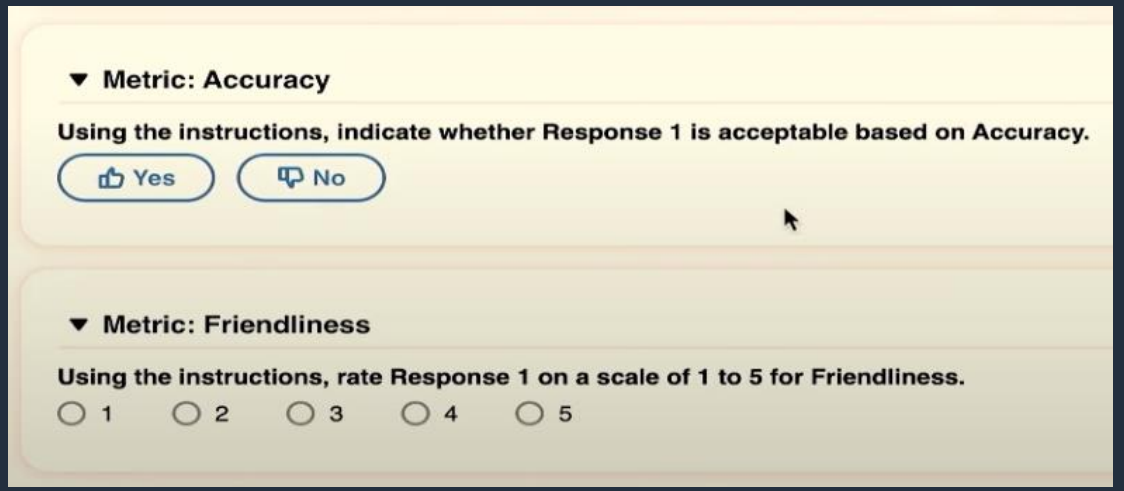
- Genetal txt generation
- Text summarization
- Question and answer
- Text classification
- Custom ← Coming Soon

aws
© 2024, Amazon Web Services, Inc. or its affiliates.

Human Model Evaluation

- モデルのプロンプトと対になる回答について、事前に定義した Metric に基づいて人間が評価を行うことができる
  - 以下の例では、「親しみやすさ」など機械的に判定しづらい項目を評価している
  - Bring your own team または AWS Managed team を利用することが可能





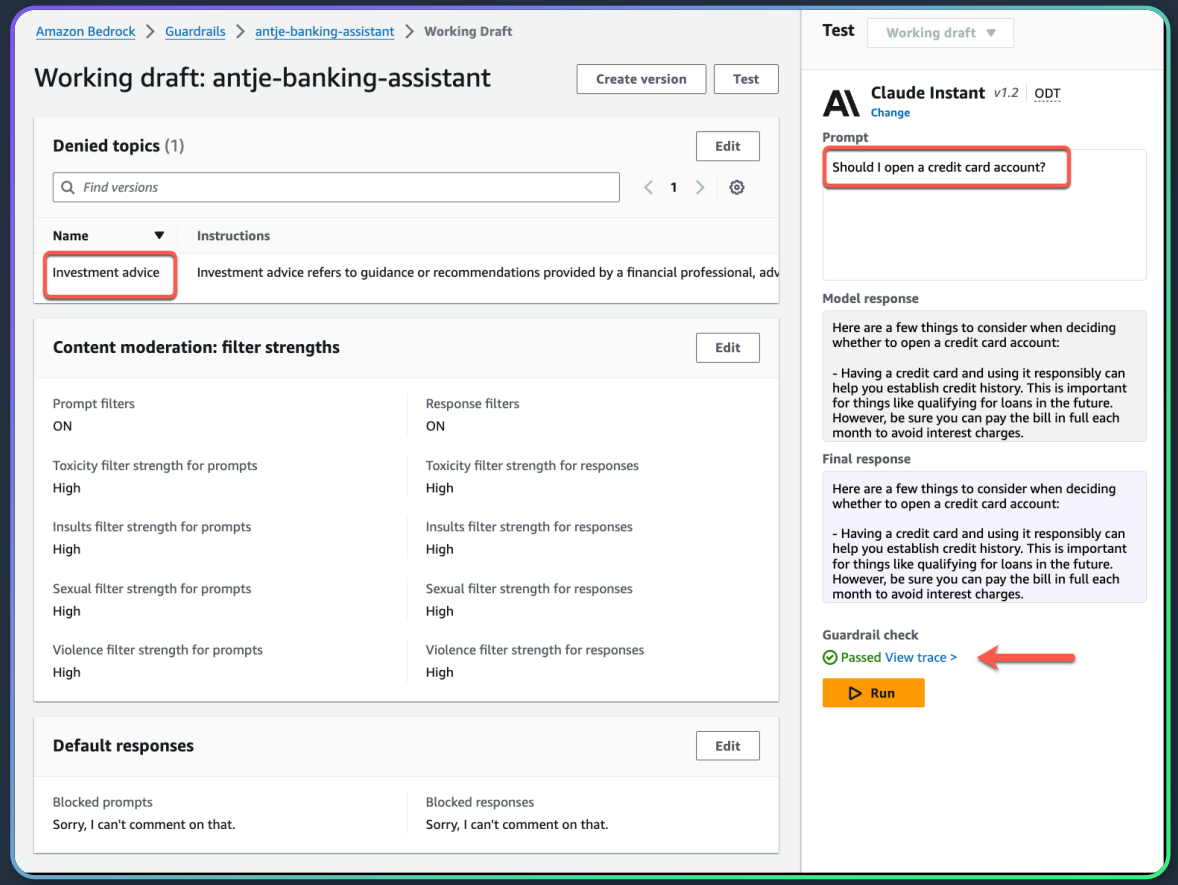
© 2024, Amazon Web Services, Inc. or its affiliates.

Guardrails for Amazon Bedrock

ファインチューニングモデルやエージェントを含め、すべてのFMにガードレールを適用

責任ある AI ポリシーに基づいて、有害なコンテンツやトピックをフィルタリングして回避するように設定します。

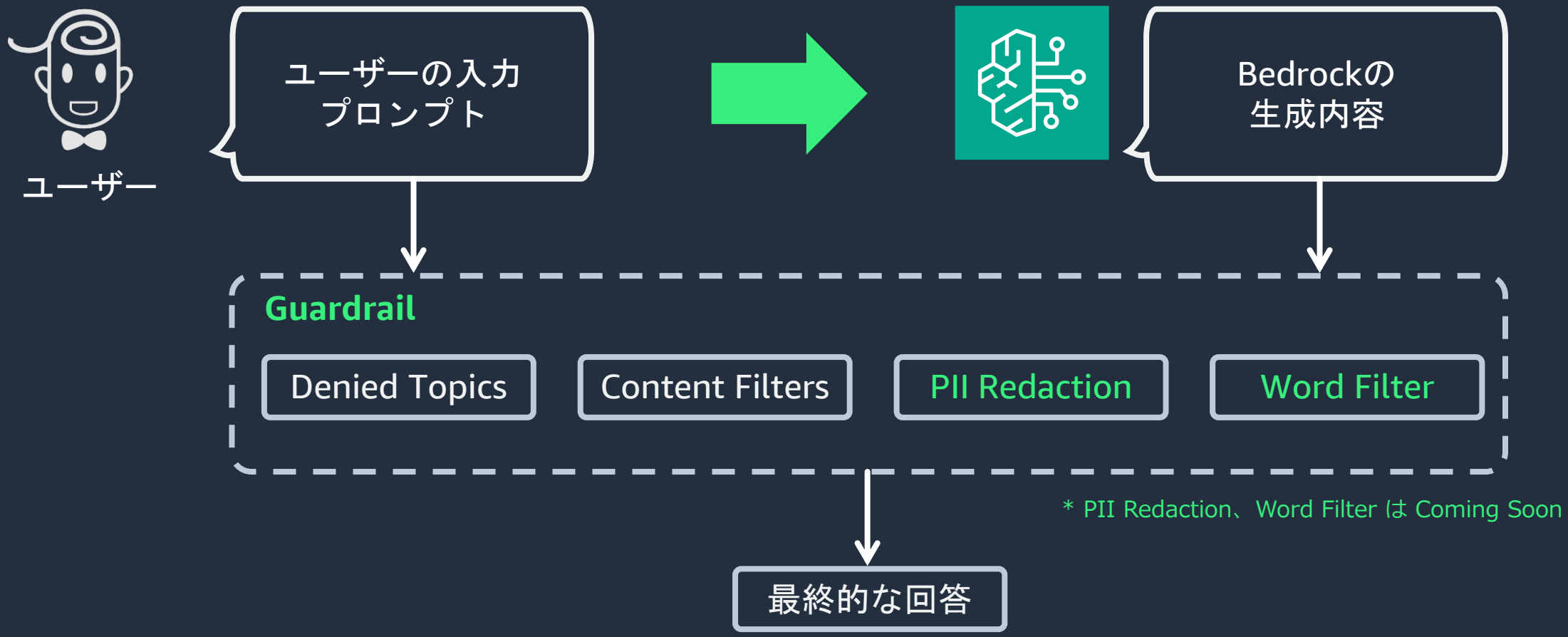
個人を特定できる情報を編集する
(近日公開予定)



© 2024, Amazon Web Services, Inc. or its affiliates.

プレビュー

Guardrails for Amazon Bedrock のしくみ



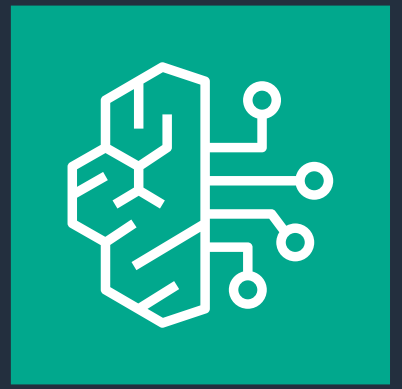
© 2024, Amazon Web Services, Inc. or its affiliates.

その他のアップデート

© 2024, Amazon Web Services, Inc. or its affiliates.

バッチ推論をサポート（Preview）

- Bedrock が複数のプロンプトをバッチで受け付けて推論できるようになった
  - 非同期実行により、モデルの評価やオフラインでの一括処理を効率化
  - CreateModelInvocationJob APIからインプットデータと出力先のS3ロケーションを指定して利用
- 料金はオンデマンドモードと同一
- クォータはバッチ推論用のクォータを利用（参考）

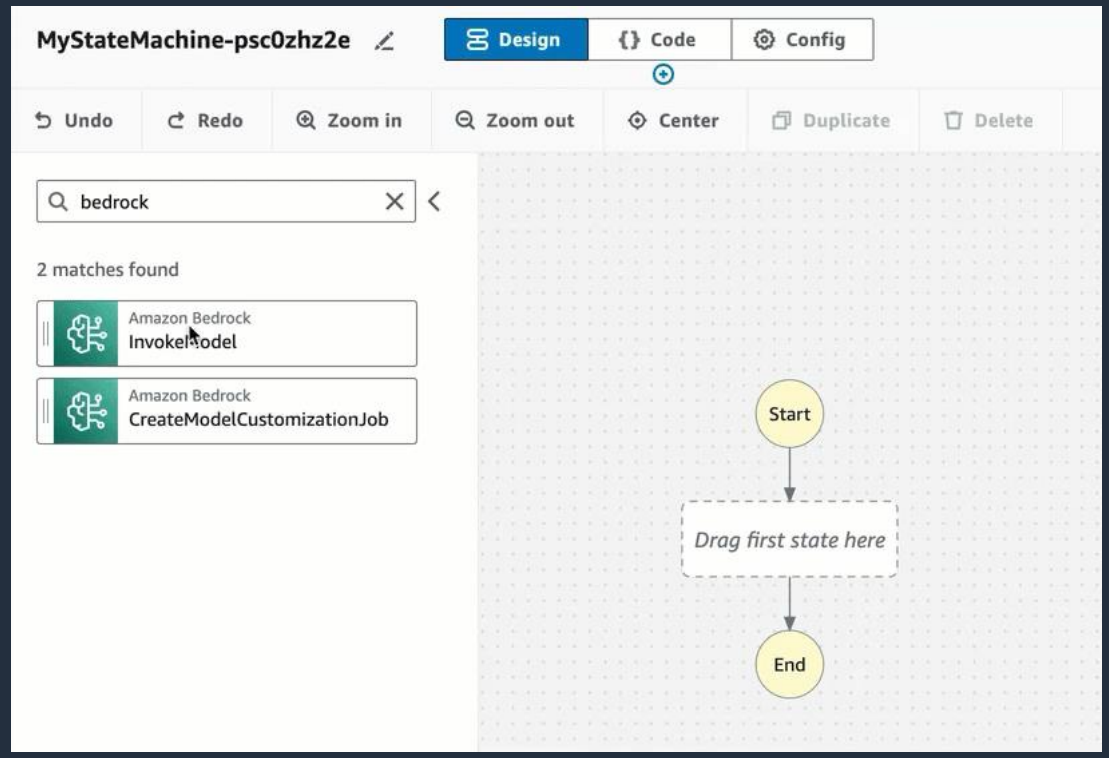


Amazon Bedrock

© 2024, Amazon Web Services, Inc. or its affiliates.

Step Functions との統合をサポート

- AWS Step FunctionsのワークフローにAmazon Bedrockによる生成系AIを組み込む機能を提供。カスタムコードの実装が不要になった
- 2つのBedrock APIとの統合が可能
  - InvokeModel API統合: S3からの情報を基盤モデルに渡したり、出力をS3に書き込むといったインタラクションを可能にする
  - CreateModelCustomizationJob API統合: モデルのファインチューニングが必要なワークフローで、その完了をまって次のステップの処理に移ることが可能



© 2024, Amazon Web Services, Inc. or its affiliates.

まとめ

- Amazon Bedrock は生成 AI の基盤モデルをオンデマンド・従量課金で利用できるサービスです。
- Amazon Bedrockで利用できる基盤モデルが追加されました。
- 外部情報と連携してユーザーからの命令を実行していくAgent for Amazon Bedrockが利用可能になりました。
- 「責任ある AI」を担保するためのコンプライアンスサービスとしてModel Evaluation・Guardrailsが新たに登場しました。
- Amazon Bedrockを使って、生成 AI を素早くビジネスに取り入れましょう！

© 2024, Amazon Web Services, Inc. or its affiliates.

Thank you!

© 2024, Amazon Web Services, Inc. or its affiliates.