生成 AI アプリケーションでデータ活用

Ren Kurosawa

SOLUTIONS ARCHITECT

2024/05/16

Ren Kurosawa

黒澤　蓮

AWS Japan Solutions Architect

Analytics/ML が Speciality で、普段はWebネイティブ業界のお客様を担当しております！

好きなAWSサービス

- AWS Glue
- Amazon Chime SDK
- Amazon Bedrock

将来の夢

宇宙でポエムを詠むこと

本日のアジェンダ

1 自社データを使う意義とユースケース

2 プロンプトエンジニアリングとRAGについて

3 自前データを使った、モデルの学習

4 データの可視化

本日のセッションとゴールの確認

汎用的な
生成 AI

自社のビジネスと顧客を
理解している生成 AI

本日のセッションとゴールの確認

汎用的な
生成 AI

自社のビジネスと顧客を
理解している生成 AI

生成 AI
を活用
したい

アプリケーション
にどう
落とし込むか
（ユースケース）
（事例）

本日のセッションとゴールの確認

汎用的な
生成 AI

自社のビジネスと顧客を
理解している生成 AI

生成 AI
を活用
したい

アプリケーション
にどう
落とし込むか
（ユースケース）
（事例）

差別化/最適化の
ために1st-Party
データを使いたい

データ
貯める

AI に
フィード
バック
/ 分析

本日のセッションとゴールの確認

今回のイベントの対象

汎用的な
生成 AI

自社のビジネスと顧客を
理解している生成 AI

生成 AI を活用したい

アプリケーションにどう落とし込むか（ユースケース）（事例）

差別化/最適化のために1st-Partyデータを使いたい

データ貯める

セッション 1,2

AI にフィードバック / 分析

本日のセッションとゴールの確認

今回のイベントの対象

汎用的な生成 AI

自社のビジネスと顧客を理解している生成 AI

生成 AI を活用したい

アプリケーションにどう落とし込むか（ユースケース）（事例）

本セッション：軽く

差別化/最適化のために1st-Partyデータを使いたい

本セッション

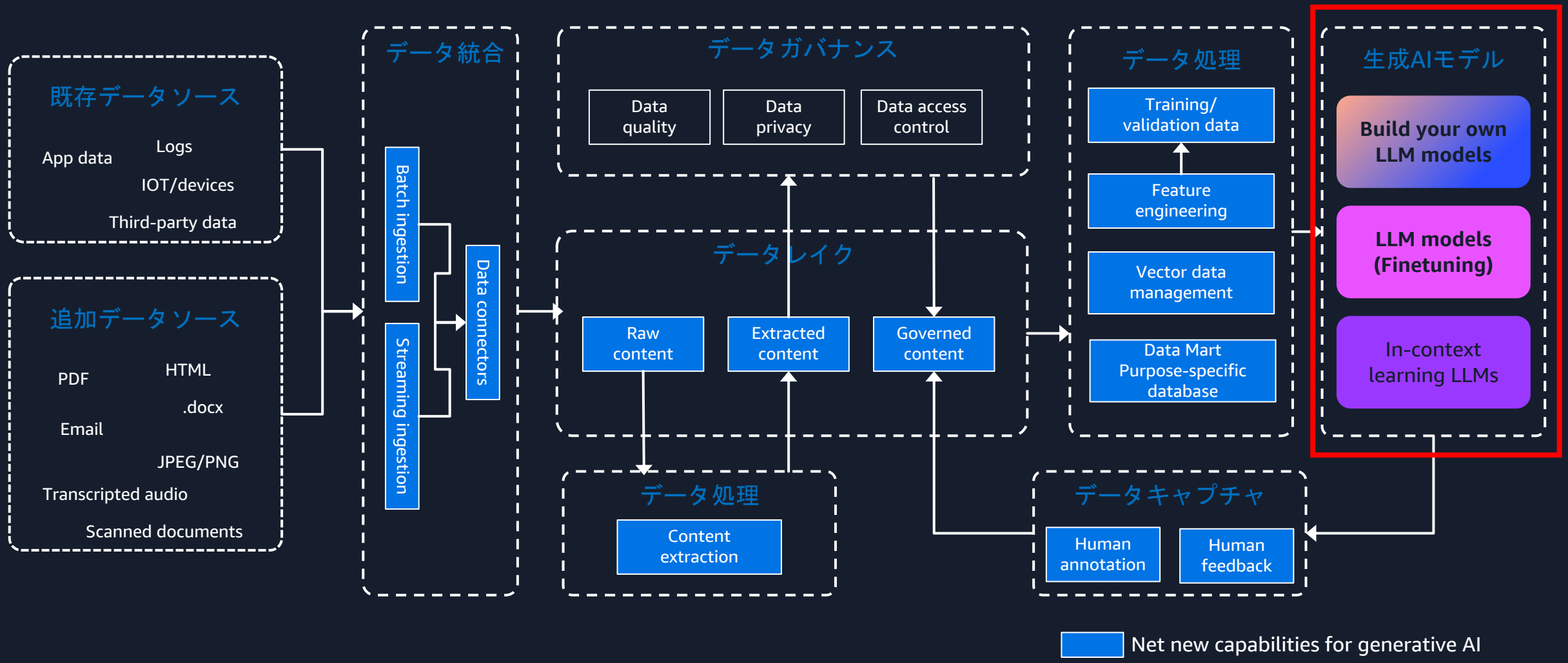
データ貯める

セッション 1,2

AI にフィードバック / 分析

本セッション

本セッションの範囲



自社データを利用した生成 AI の活用先

本日のセッションとゴールの確認

汎用的な
生成 AI

自社のビジネスと顧客を
理解している生成 AI

ここ

生成 AI
を活用
したい

アプリケーション
にどう
落とし込むか
（ユースケース）
（事例）

差別化/最適化の
ために1st-Party
データを使いたい

データ
貯める

セッション
1,2

AI に
フィード
バック
/ 分析

データを集めることでできるようになること

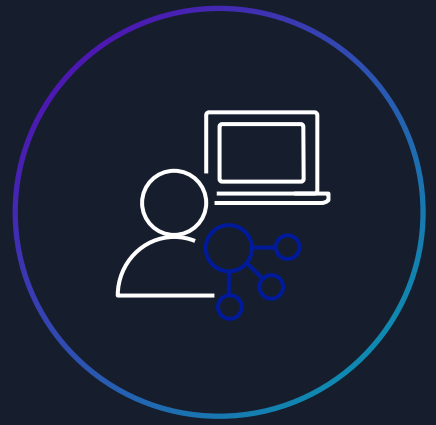
データを集めることは準備、データ活用は手段、その先の（ビジネス）ゴールは何か

業務効率化

在庫管理

異常検知

社内ナレッジの共有



顧客分析

施策が成功したか

潜在的なニーズの発掘

売上の予測

資産

機械学習データセット

データの売買

社会貢献

データを集めることでできるようになること

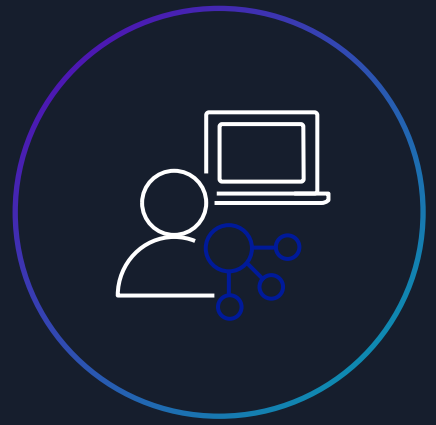
データを集めることは準備、データ活用は手段、その先の（ビジネス）ゴールは何か

業務効率化

在庫管理

異常検知

社内ナレッジの共有



顧客分析

施策が成功したか

潜在的なニーズの発掘

売上の予測

資産

機械学習データセット

データの売買

社会貢献

生成 AI を/へ活用

生成 AI (Generative 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できる AI の手法

- 画像生成

  cute cat

- 会話生成

Human

どうすれば人に好かれますか

Assistant

人に好かれるためには、以下のような点を心がけると良いでしょう。

- 相手の話をじっくり聞く。質問したり相づちを打ったりして関心を示す。
- 相手の話題に合わせて会話を作る。自分の話に偏らない。
- 褒める機会があれば積極的に褒める。人は褒められると嬉しくなる。
- 相手の話を否定せず、共感を示す態度を心がける。
- 相手の話のポイントを汲み取り、要点をまとめて伝える。

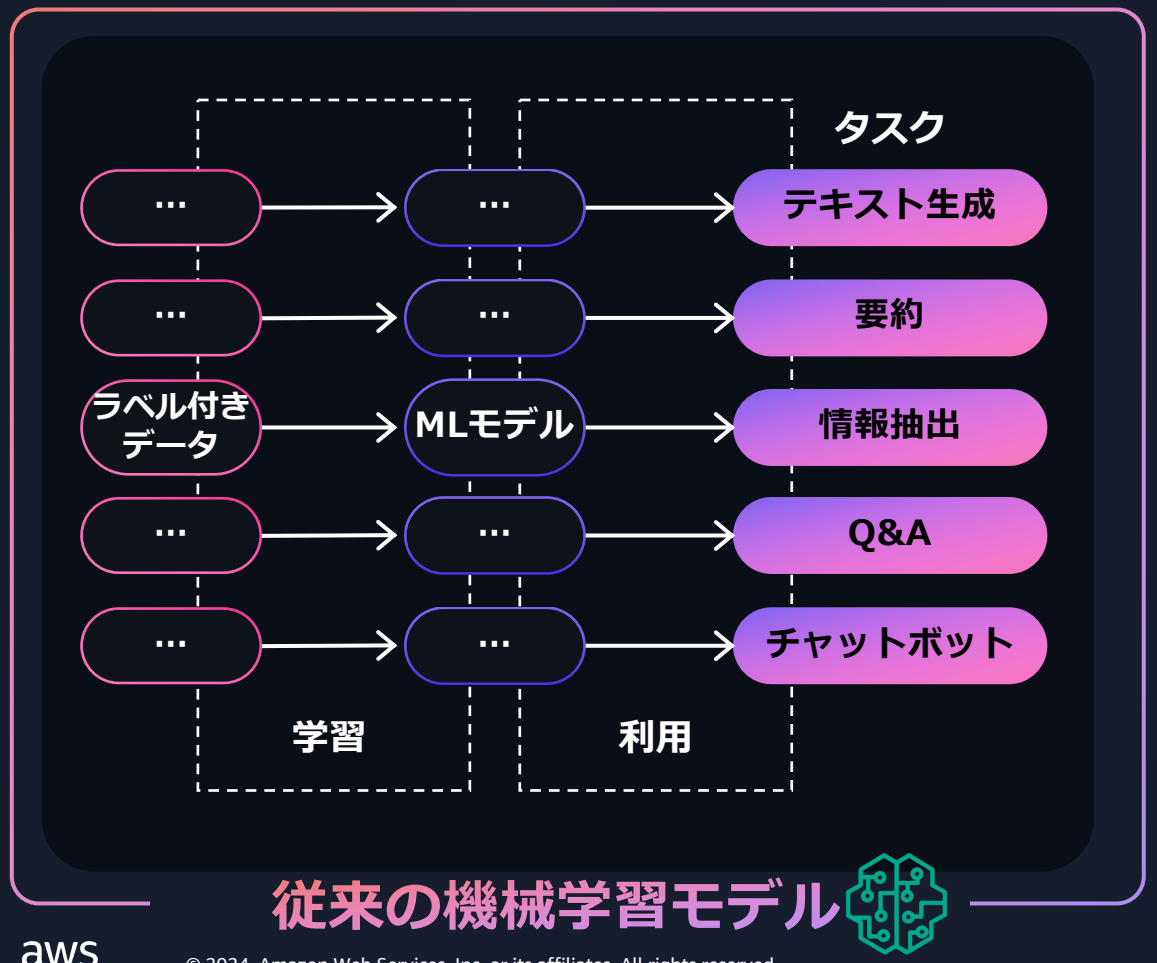
- コード生成

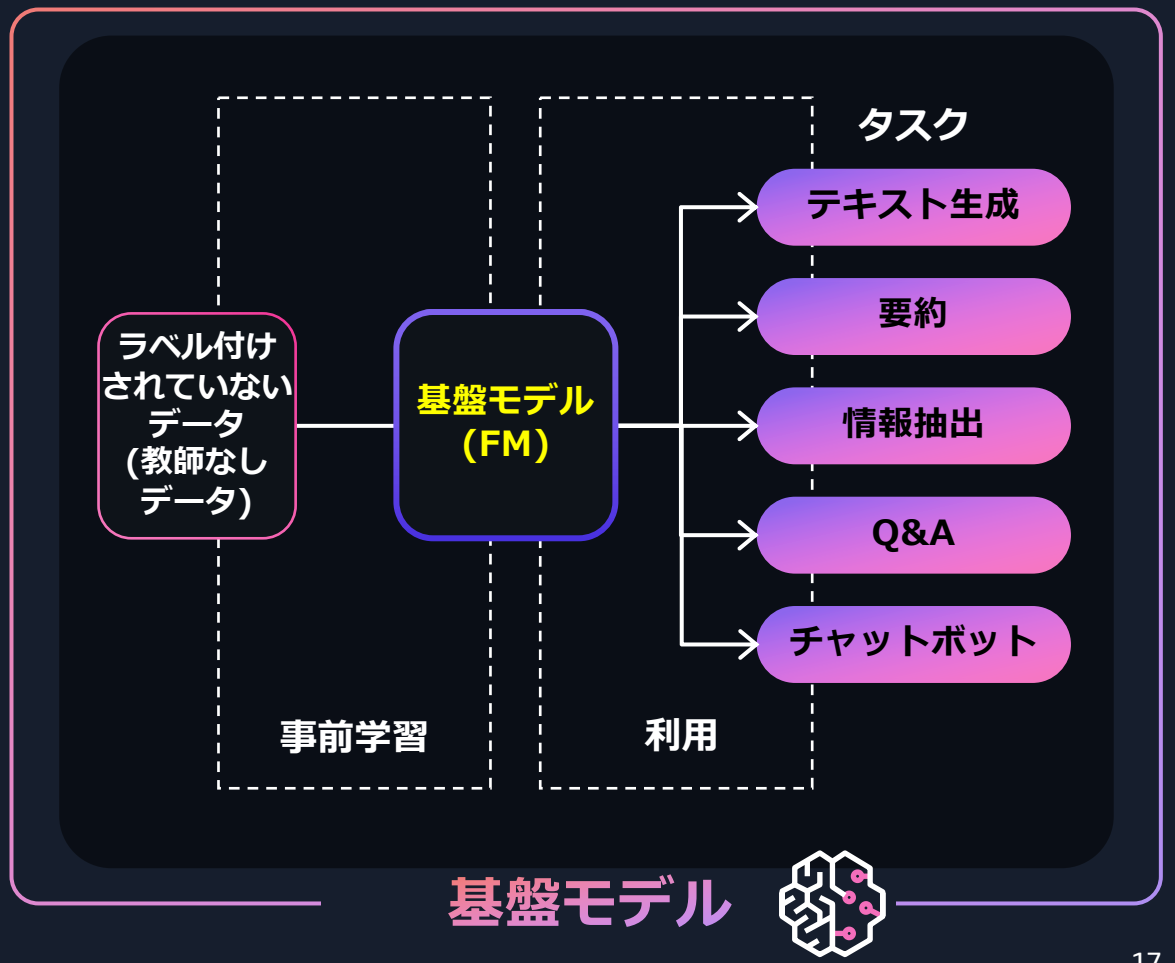
入力:
1から100までの数字を順番に
出力してください。
ただし、2で割り切れるならFizz,
3で割り切れるならBuzz,
2でも3でも割り切れるならFizzBuzz
と出力してください。

```python
for i in range(1, 101):
  output = ""
  if i % 2 == 0:
    output += "Fizz"
  if i % 3 == 0:
    output += "Buzz"
  print(output if output != "" else i)
```

生成系 AI を実現する仕組み - 基盤モデル

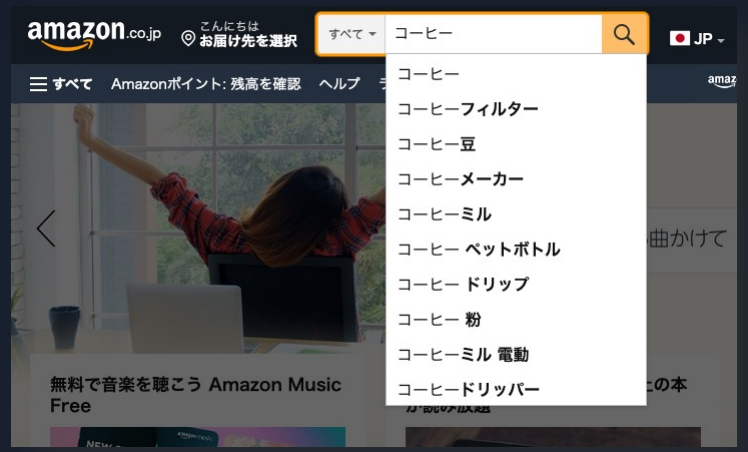
生成系 AI は「**基盤モデル**」と呼ばれる
「**膨大なデータに基づき事前にトレーニングされた大規模な機械学習モデル**」によって実現





基盤モデルで何ができるか？

Amazonにおける 生成AI 活用例

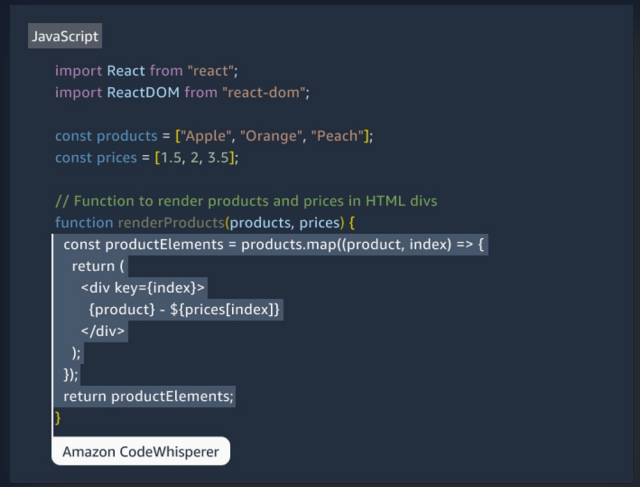


Amazon の検索機能

基盤モデルを導入し
検索結果を改善

Alexa 教師モデル

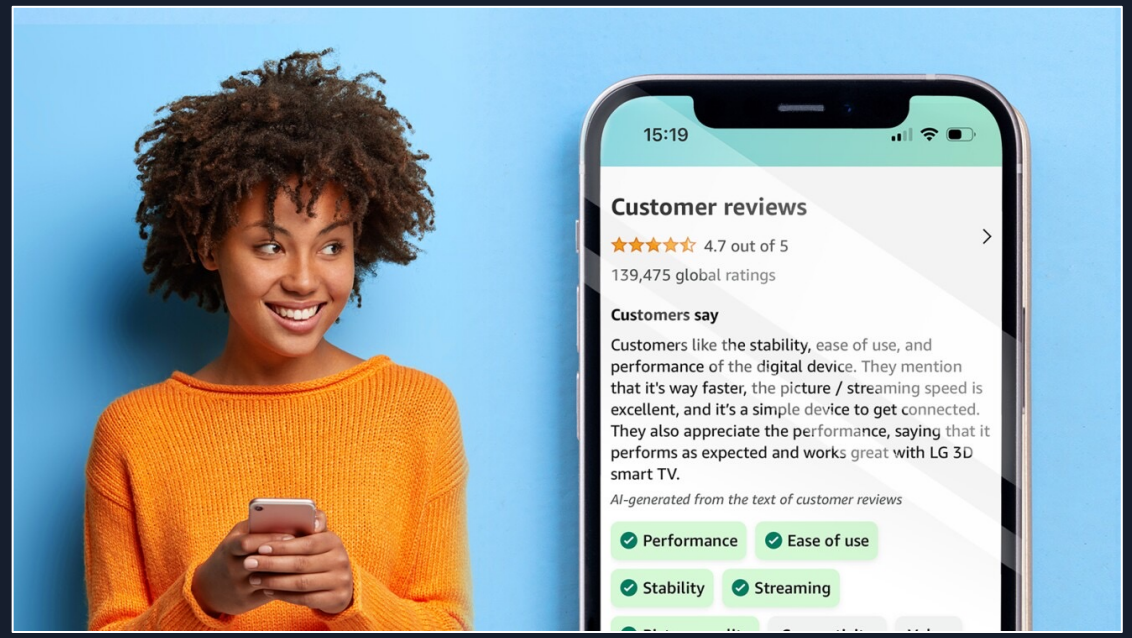
最小限の入力情報から
多言語による応答を実現



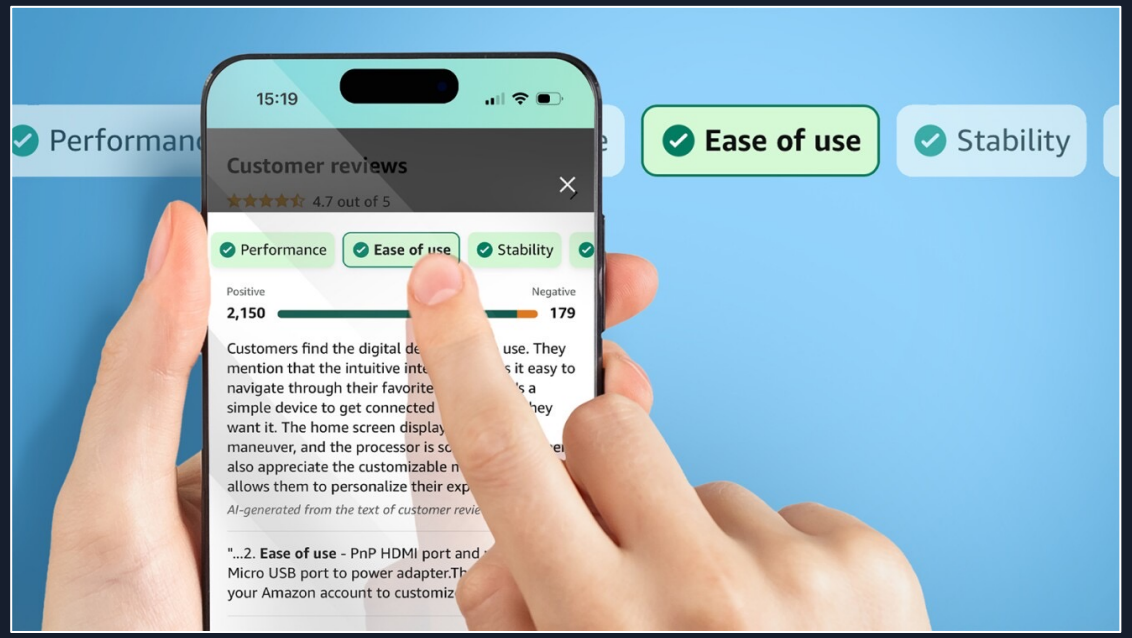
Amazon CodeWhisperer

コメントとコードから
推奨コードを自動生成

Amazonにおける 生成AI 活用例



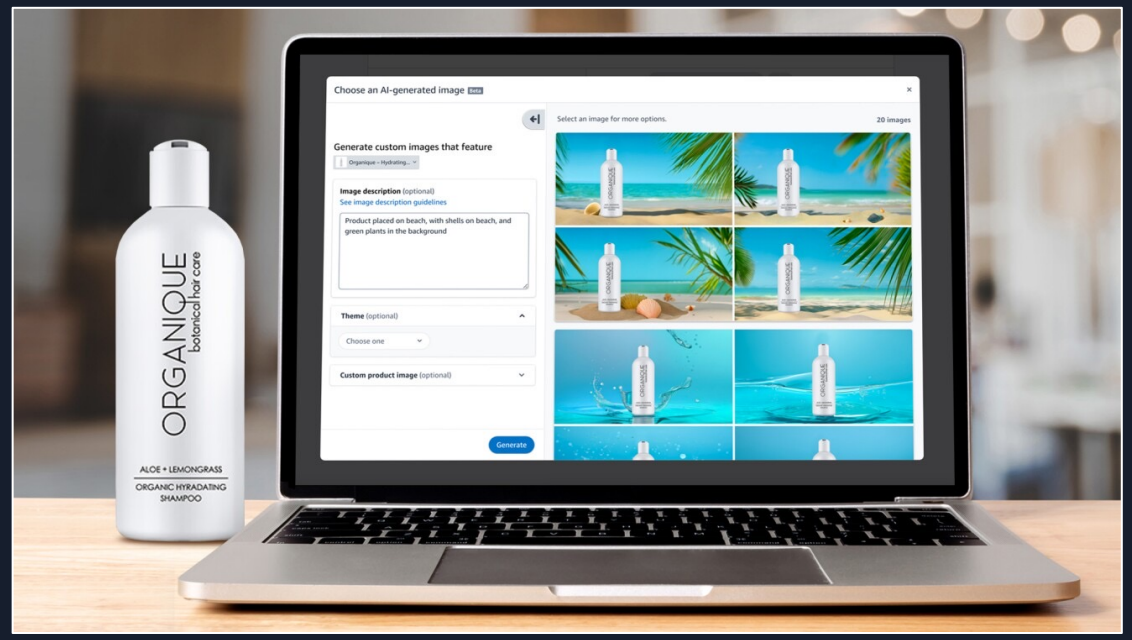
- 信頼できるユーザーのレビューに基づき AI が製品レビューのハイライトを生成
- 製品の特徴を示すアイコンを AI が生成（性能、使い易さ、安定性など）



- アイコンを選択すると、関連するレビューテキストの詳細をまとめて読める
- 出品者の作業負担を軽減するAI のサービスも開始し、商品の効率的な販売を支援

Source: Amazon.com / “How Amazon continues to improve the customer reviews experience with generative AI” https://www.aboutamazon.com/news/amazon-ai/amazon-improves-customer-reviews-with-generative-ai

Amazonにおける 生成AI 活用例



- 商品の宣伝に失敗した出品者の 75% がクリエイティブの選択を課題に挙げている
- 出品者の労力なしに魅力的な画像を生成しつつ、お客様体験の向上を助ける策を検討

- 出品者がワンクリックで商品に適した背景を生成できるツールのベータ版をリリース
- プロンプトによる画像調整にも対応
- リアルな背景のクリックスルー率は 40% 増

Source: Amazon.com / “Amazon rolls out AI-powered image generation to help advertisers deliver a better ad experience for customers” https://www.aboutamazon.com/news/innovation-at-amazon/amazon-ads-ai-powered-image-generator

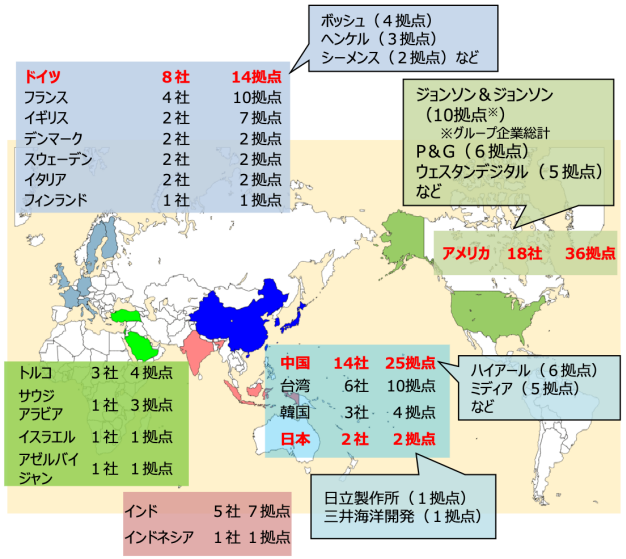
資料理解

製造業をめぐる新しい国際的な潮流

第５章　環境の変化

- 世界経済フォーラムは、2020年より世界最先端工場「Global Lighthouse」を選出しており、**132のうち日本企業は２拠点**。
- 選出に当たり、経済的効率性だけでなく、デジタル技術を活用した**サプライチェーン全体での最適化を通じ**、生産性の向上・**市場ニーズに応じた柔軟な生産の実現**や、**環境負荷の低減**等が図られているかどうかが重要視されている。
- **DXやGXによる全体最適化の達成**が、製造事業者の先進性の評価軸となる、国際的な潮流が生まれている。

図　Global　Lighthouse選出状況※（本社所在国別）



選出に当たって重視される審査ポイント

インパクト	製造プロセスや製品等の高度化に大きなインパクトを達成しているか
ユースケース	複数の4IRのテクノロジーを活用・統合し、ユースケースを創出しているか
イネーブラー	4IRの活用戦略、IoTのアーキテクチャ、労働力の巻き込み・活用や能力開発などの複数要素をうまく統合・活用しているか
テクノロジープラットフォーム	革新的でスケーラブルなテクノロジープラットフォームであるか

※4IRのテクノロジーは、人間の生活・仕事に根本的な変化をもたらす技術革新を可能とする先端的な技術を差し、例えば、ビッグデータ、IoT、AI、ロボットの活用等である。（4IR＝第四次産業革命）
※Sustainability Lighthouseの選出に当たっては、「目的：明確なサステナビリティ目標があるか」、「インパクト：複数の環境カテゴリーで改善ができているか」、「スケール：複数の4IRのテクノロジーを活用し、サステナビリティ面での効果・インパクトを残しているか」の3つの審査ポイントが追加される。

選出工場の取組と効果

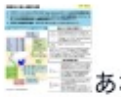
ハイアール【中国・家電メーカー】	• 冷蔵庫の生産工場において、注文内容をリアルタイムで製造工程に反映させる高度なデジタルプラットフォーム「COSMOPlat」を開発。 • 柔軟に製造工程を組み換えるマス・カスタマイゼーションを実現した。
	• ビッグデータの活用により、市場調査に要する時間を85%削減 • AIによる工程最適化により、エネルギー消費量を37%削減 • ロボットの連結と自動化により、組立効率を52%向上
ボッシュ【ドイツ・自動車部品メーカー】	• 全世界に点在する工場の機械やデバイスをネットワークで連結することで、生産や物流計画の管理、機械の状態の把握等を一本化。 • 全社的な全体最適化に成功した。
	• 物流システムの統合により、生産リードタイムを36%削減 • 設備に生じる異常の事前検知により、メンテナンスコストを25%削減 • エネルギーマネジメントの実現により、消費電力を18%削減

（資料）世界経済フォーラム公表資料より経済産業省作成

22

https://www.meti.go.jp/report/whitepaper/mono/2023/index.html

Claude 3 Sonnet v1 | ドット
モデルを変更



あなたは製造業のコンサルタントです、この画像を丁寧に確認した上で、内容を要約してください。また、そのようやく結果から得られる示唆と日本の製造業が取り組むべき課題について教えて下さい

この画像は、デジタル技術の活用によりグローバルな製造業の変革が進展していることを示しています。主な内容は以下の通りです。

- 世界的なイニシアチブ「Global Lighthouse」では、132社の先進企業が製造現場のデジタル化を推進している。
- AIやIoTなどの先端技術を活用することで、生産性の向上や製品・サービスの高度化が図られている。
- DX(デジタルトランスフォーメーション)の進展により、製造業の在り方や付加価値が変化している。

この動きから得られる示唆は以下の通りです。

1. デジタル技術の積極的な活用が製造業の競争力強化に不可欠である。
2. 製品・サービスの高付加価値化や新ビジネスモデルの創出が重要になる。
3. データ駆動型のスマート製造への転換が求められる。

日本の製造業が取り組むべき課題としては、以下が考えられます。

- AIやIoTなどの先端技術の一層の導入・活用
- DXを見据えた業務プロセスの見直しと人材育成
- サプライチェーン全体でのデータ連携と共有
- 製品ライフサイクル全体を視野に入れた高付加価値化

グローバルな潮流に遅れを取らないよう、デジタル化を積極的に推進していくことが欠かせません。

観光画像へのタグ付けと魅力的な説明文生成

建物, 夜景, 和風 [1]

この画像は、日本の伝統的な温泉旅館の夜景を捉えています。木造の建物と赤い灯籠が並ぶ情緒あふれる風景は、日本の古き良き時代の雰囲気を感じさせます。夜の静けさの中で、温泉の湯けむりに包まれながら過ごす時間は、現代の喧騒から離れてゆっくりと心を休める贅沢な体験となるでしょう。日本を象徴する景色の中で、伝統的な温泉文化に触れることができる魅力的な場所です。

通り, 紅葉, 車 [1]

この美しい街路樹が黄金色に染まった風景は、秋の訪れを華やかに告げています。落ち着いた住宅街の一角にあり、ゆったりとした雰囲気が漂う通りです。太陽の光が木々の葉を透かし、幻想的な光景を創り出しています。歩いて散策するのがおすすめです。秋の穏やかな空気に包まれながら、この美しい景色を堪能できるでしょう。地元の人々の生活の一コマを垣間見ることもできる、魅力的な場所です。

[1] 候補タグ50個からふさわしいものを最大3つ選定させた

SNSマーケティング支援

Instagram Marketing Assistant

Choose an image...

Drag and drop file here
Limit 200MB per file • JPG, JPEG, PNG
Browse files

2024-03-21_18-36-36.jpg 87.7KB

Focus item

パンツ

画像をアップロードして、
紹介したいアイテムを指定

Analyze Image

Generate Instagram posts and hashtags

画像の分析と解説

画像全体の説明: この画像は、都会的な街路を背景に、全身黒の服装でポーズをとる男性を捉えています。彼の表情は真剣で、キャップとバックパックを身に付けています。前景には車道と仕切りの壁があり、外国の街並みが窺えます。全体的にストリートスタイルのファッションが際立っています。

パンツの説明: 男性が着用しているパンツは、黒色のワイドシルエットで裾広がりのデザインです。生地は厚手で落ち着いた雰囲気を醸し出しています。ヒップからウエストにかけてゆとりがあり、ルーズなスタイリングがされています。足元のスニーカーと相まって、カジュアルでありながらも洗練された印象を与えるアイテムです。

提案

Sonnet で読み込んだ画像を元に、インスタへの
投稿用原稿とハッシュタグを出力

インスタグラムの投稿の原稿提案:

1. 都会的なストリートファッションに映える、ワイドシルエットのブラックパンツ。 ルーズなシルエットとドレープ感のある生地で、カジュアルさとこなれ感を両立。 旬のスタイリングに取り入れて、あなたもストリートの逸材に。

2. 街角のスナップショットから生まれた、最新ブラックパンツ。 落ち着いた色味と上質な生地で、あなたのスタイルに洗練された雰囲気をプラス。 きちんと感とカジュアル感のバランスが絶妙なこのパンツは、毎日のコーデに欠かせない。

3. ストリートカルチャーから着想を得た、ブラックワイドパンツ。 ゆとりのあるシルエットとドレープ感が、さり気ないこなれ感を演出。 黒の重厚感とスポーティーなディテールの絶妙なブレンドは、センスの良さを物語る。

適切なハッシュタグ提案:

#ストリートファッション #ワイドパンツ #ブラックパンツ #ルーズシルエット #シックスタイル #カジュアルスタイル #ストリートスナップ #スタイリングアイデア
#モノトーンコーデ #メンズファッション

文字起こし結果の要約

架空の打ち合わせの文字起こし結果から、要約とネクストアクションを抽出しJSONで出力

{

"summary": "会議では、春らしい具材を使ったおにぎりの提案がなされました。桜の葉、新玉ねぎ、筍、山菜、桜の花、菜の花などの具材が挙げられ、それぞれの味付けや見た目の工夫について議論されました。",

"next_action": "出されたアイデアを元に、それぞれの具材を使ったおにぎりの試作品を作り、味や見た目を確認して商品化の可否を検討することになりました。”

}

論文の要約 (自然なキャラクターロールプレイも)

太郎「博士、今日はどんな論文を教えてくれるの?」

博士「ほっほっほ、今日はAWSのサーバレスコンピューティングサービス、Lambdaにおける高速コンテナ起動の仕組みについての論文を紹介するぞい」

太郎「Lambdaってクラウドの関数実行サービスですよね。コンテナの起動が速いって、どういうことなんですか?」

博士「おお、鋭い質問じゃ。Lambdaではユーザーの関数をコンテナ内で実行するのじゃが、そのコンテナイメージを起動する際の待ち時間、いわゆるコールドスタート時間を最小限に抑える必要があるのう」

太郎「なるほど。ユーザーにとってコールドスタート時間が長いと使いにくいですからね」

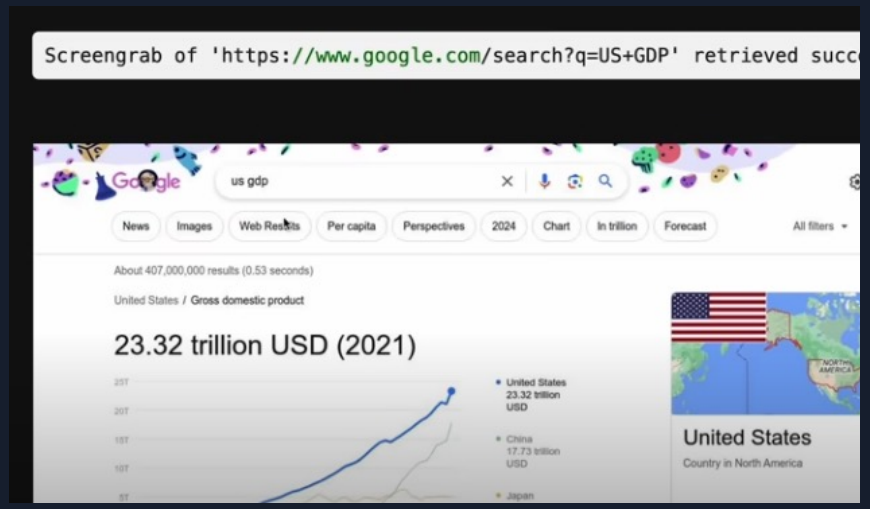
博士「そうじゃ。従来はコンテナイメージ全体を転送・展開する必要があり、イメージサイズが大きくなるとコールドスタート時間が長くなってしまうのじゃ。そこで本論文では、コンテナイメージを小さなチャンクに分割し、必要なチャンクのみを転送・ロードする方式を提案しているのう」

太郎「へえ、スパースローディングって手法ですね。でも大きなイメージを分割するとチャンク数が膨大になりそうですし、セキュリティの問題もありそうです ...

世界のGDPトレンドの分析

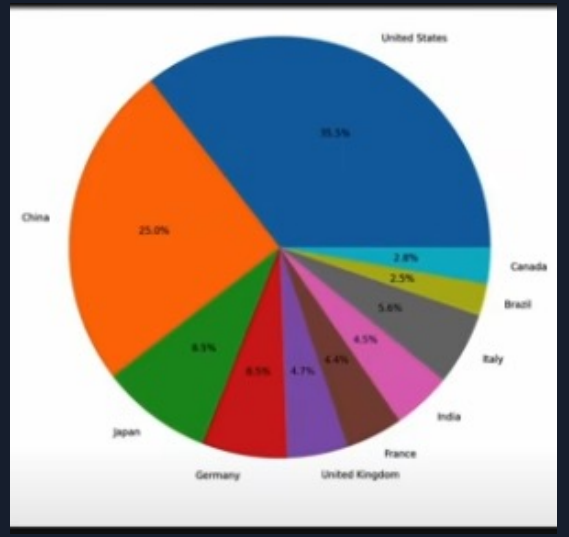
Amazon Bedrockで利用可能なClaude 3 Opusの公式デモ

世界のGDPトレンド分析という複雑なタスクに対して、Claude 3が「Webサイト閲覧」「Python実行」「サブエージェントの実行」などの複数のツールを駆使して自律的に分析を実行している



Webサイトを閲覧しグラフを読み取る

国ごとに並列で分析を進める
サブエージェントを実行



Pythonで結果を可視化

https://www.youtube.com/watch?v=sjL6Gl6ZIqs&t=1s

動画のハイライト生成 with Bedrock

README MIT license

Soccer Highlight Generator

Automate the creation of soccer match highlights with the power of Generative AI and AWS. This solution leverages AWS Bedrock (Anthropic's Claude 3 Sonnet model), AWS MediaConvert, Lambda, Step Functions and other AWS services to identify and compile exciting game moments without manual editing.

Author

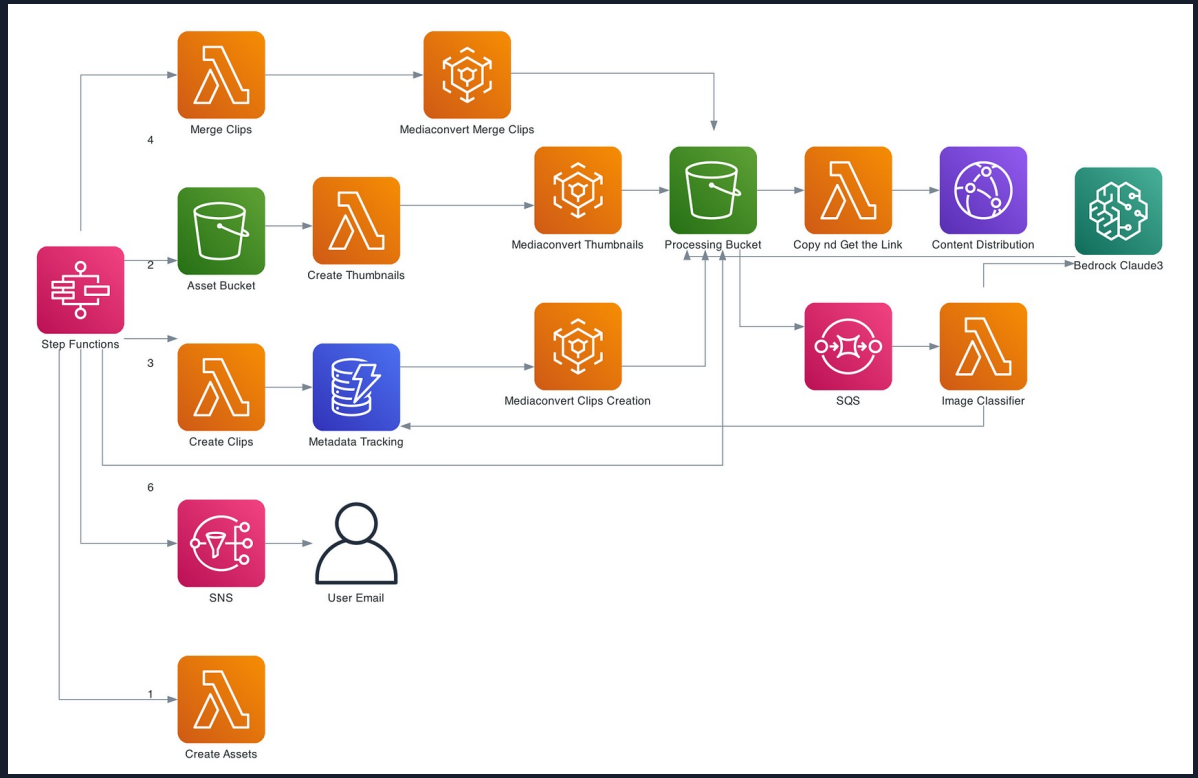
Pedram Jahangiri

Getting Started

For a detailed explanation of what this solution does and the benefits it offers, please refer to my blog

Prerequisites

- AWS CLI installed and configured with the necessary permissions
- Node.js and npm
- Python 3.11 and pip
- An AWS account with the required services enabled
- Access to Amazon Bedrock foundation models (Before you can use a foundation model in Amazon Bedrock, you must request access to it. Use this Link for detail https://docs.aws.amazon.com/bedrock/latest/userguide/model-access.html)

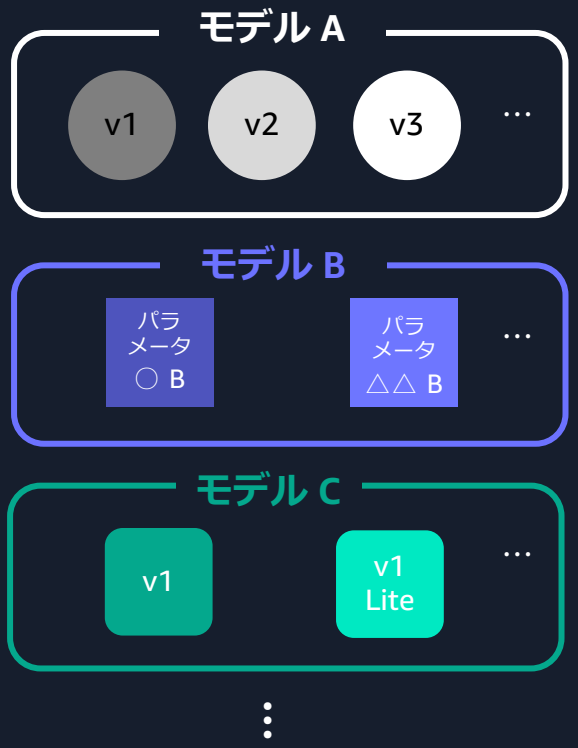


https://github.com/iut62elec/Soccer-Highlight-Generator-with-GenAI

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



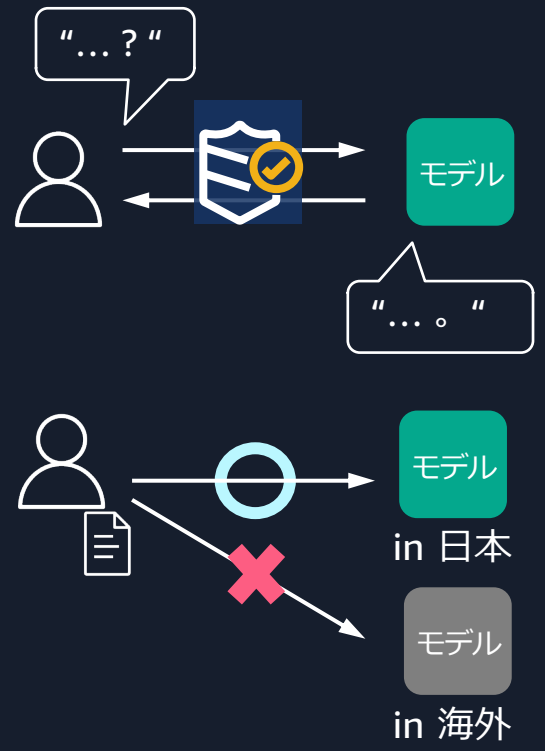
インフラストラクチャ

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
- …

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



Generative AI tools and services

APPLICATIONS THAT LEVERAGE LLMs AND FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR MODEL TRAINING & INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

Generative AI tools and services

APPLICATIONS THAT LEVERAGE LLMs AND FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR MODEL TRAINING & INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

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

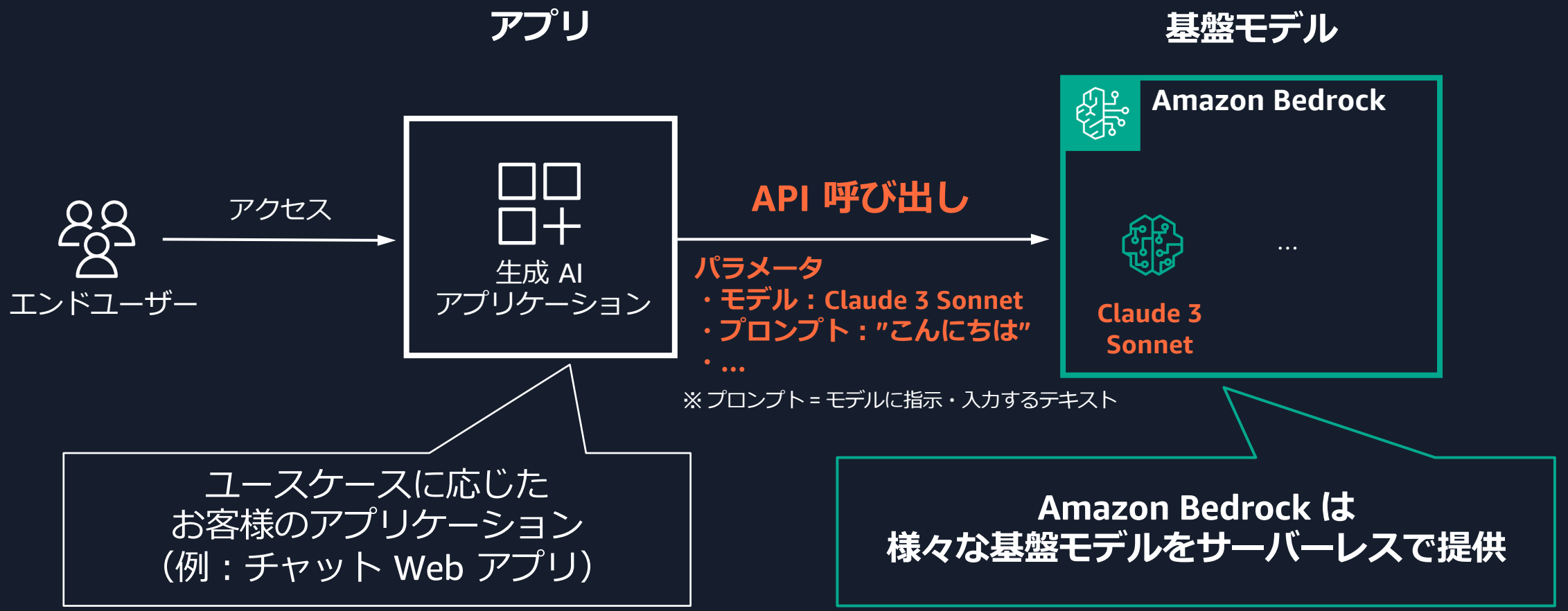
お客様の業務用途に適した基盤モデルを選択
Amazon、AI21 Labs、
Anthropic、Cohere、Meta、Stability AI

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

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



Amazon Bedrock

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

Jurassic-2 Ultra
Jurassic-2 Mid

CLAUDE

テキスト & ビジョン

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 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 2

テキスト

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

SDXL

画像

SDXL 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

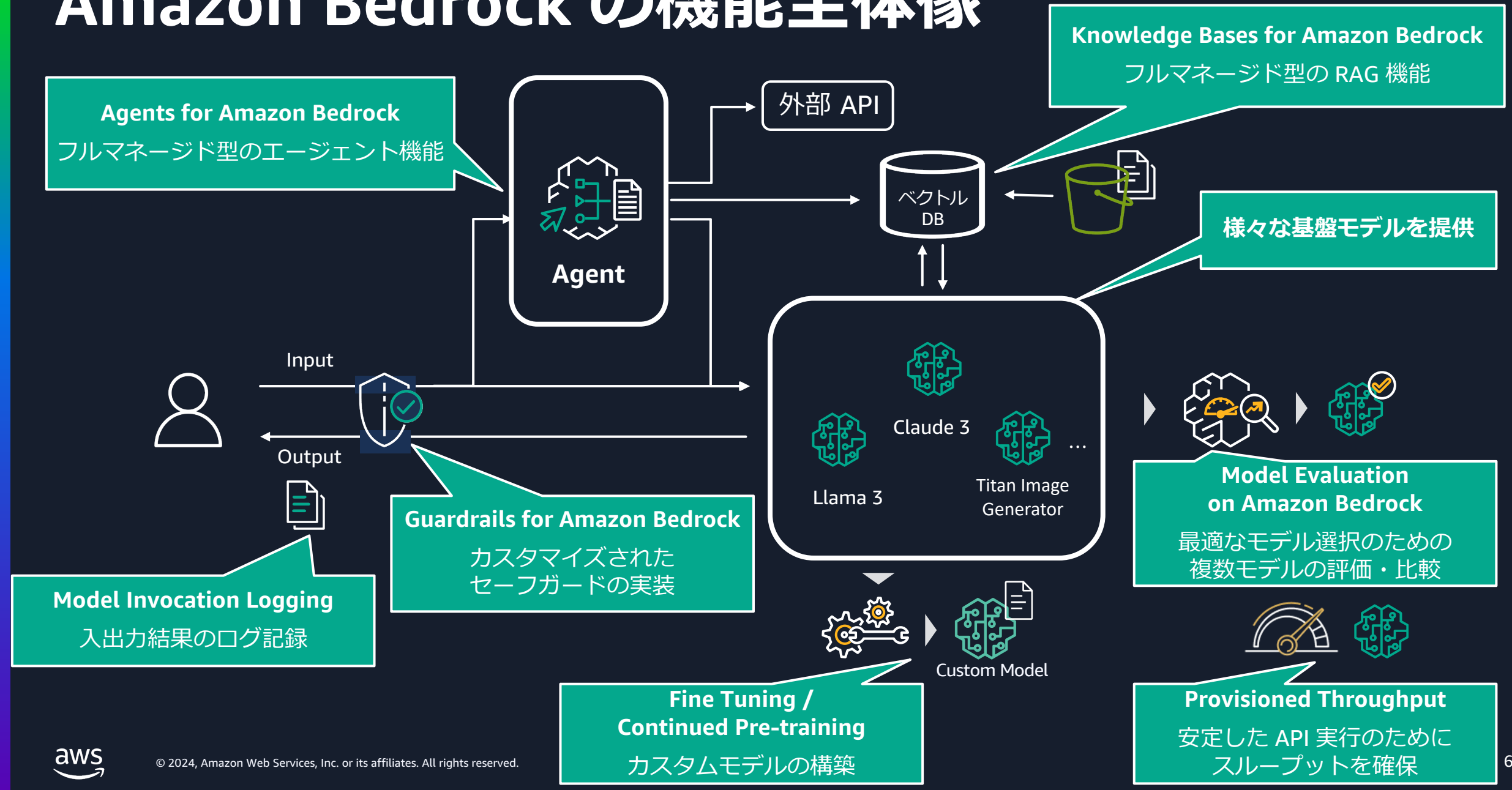
Titan Multimodal Embeddings
Titan Text Embeddings V2
Titan Text Embeddings

※ 2024 年 4月以降に発表あるいは一般提供開始

Amazon Bedrock は
様々な基盤モデルを
API 経由で利用できる サービス

基盤モデルだけでなく
生成 AI アプリの構築・運用をサポートする
幅広い機能を提供するサービス

Amazon Bedrock の機能全体像



本日のセッションとゴールの確認

汎用的な
生成 AI

自社のビジネスと顧客を
理解している生成 AI

ここ

生成 AI
を活用
したい

アプリケーション
にどう
落とし込むか
（ユースケース）
（事例）

CLEAR

差別化/最適化の
ために1st-Party
データを使いたい

データ
貯める

セッション
1,2

AI に
フィード
バック
/ 分析

1st party データを利用した生成 AI

新たな 生成 AI のデータパターン

1

検索拡張生成（RAG)

事前にトレーニングされたモデルに、自社のドメイン固有のコンテキストデータを加えた結果を生成

例: ドメイン固有の、要件が限定された仮想エージェント

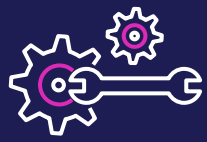
2

既存モデルのファインチューニング

既存のモデルをベースに、ドメイン固有のデータによる追加学習を実施

例: ドメイン特化型の知識エージェント

3



独自モデルの作成

ドメイン固有データによるモデルの関するモデルのトレーニングを実施

例: 特定のデータで学習を行った、高度にドメインに特化したアプリケーション

新たな 生成 AI のデータパターン

0

プロンプトエンジニアリング

1

検索拡張生成（RAG)

事前にトレーニングされたモデルに、自社のドメイン固有のコンテキストデータを加えた結果を生成

例: ドメイン固有の、要件が限定された仮想エージェント

2

既存モデルのファインチューニング

既存のモデルをベースに、ドメイン固有のデータによる追加学習を実施

例: ドメイン特化型の知識エージェント

3

独自モデルの作成

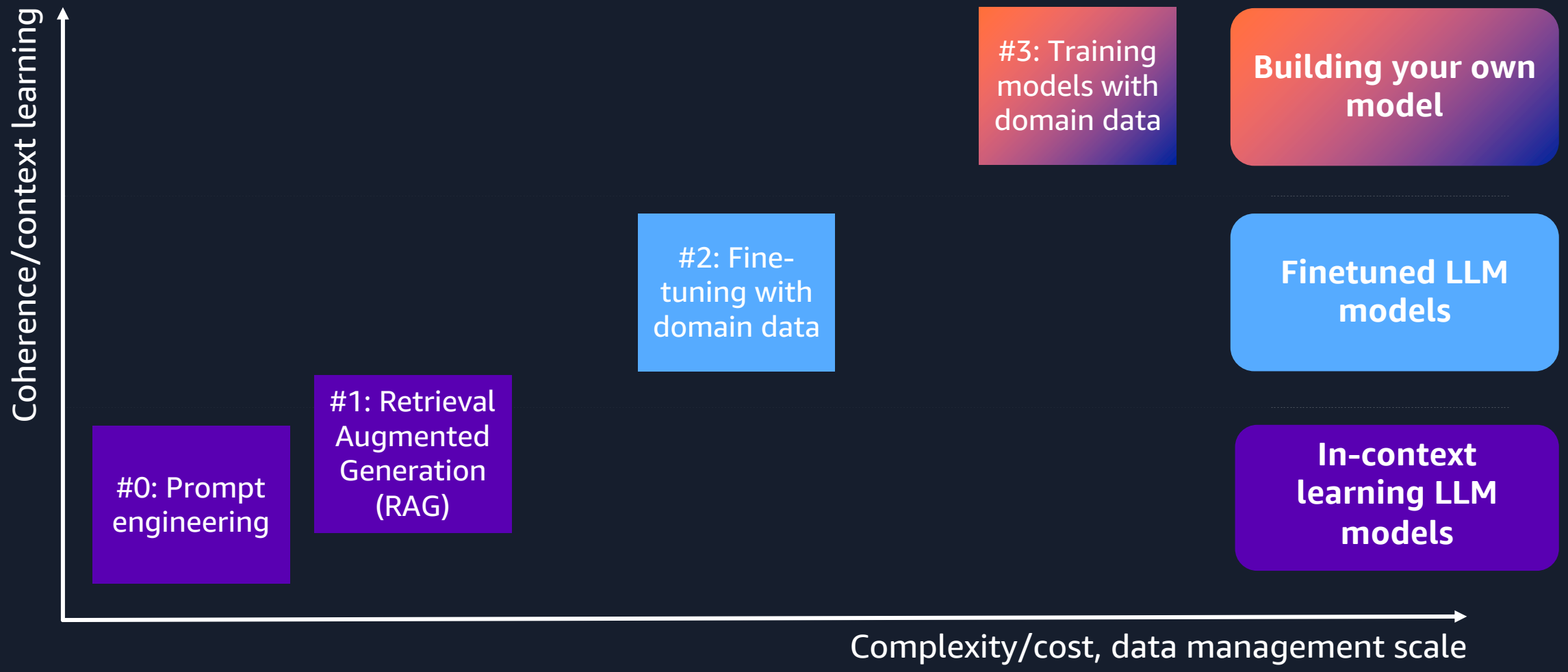
ドメイン固有データによるモデルの関するモデルのトレーニングを実施

例: 特定のデータで学習を行った、高度にドメインに特化したアプリケーション

一般のモデルを利用し、
入力するデータをカスタマイズ

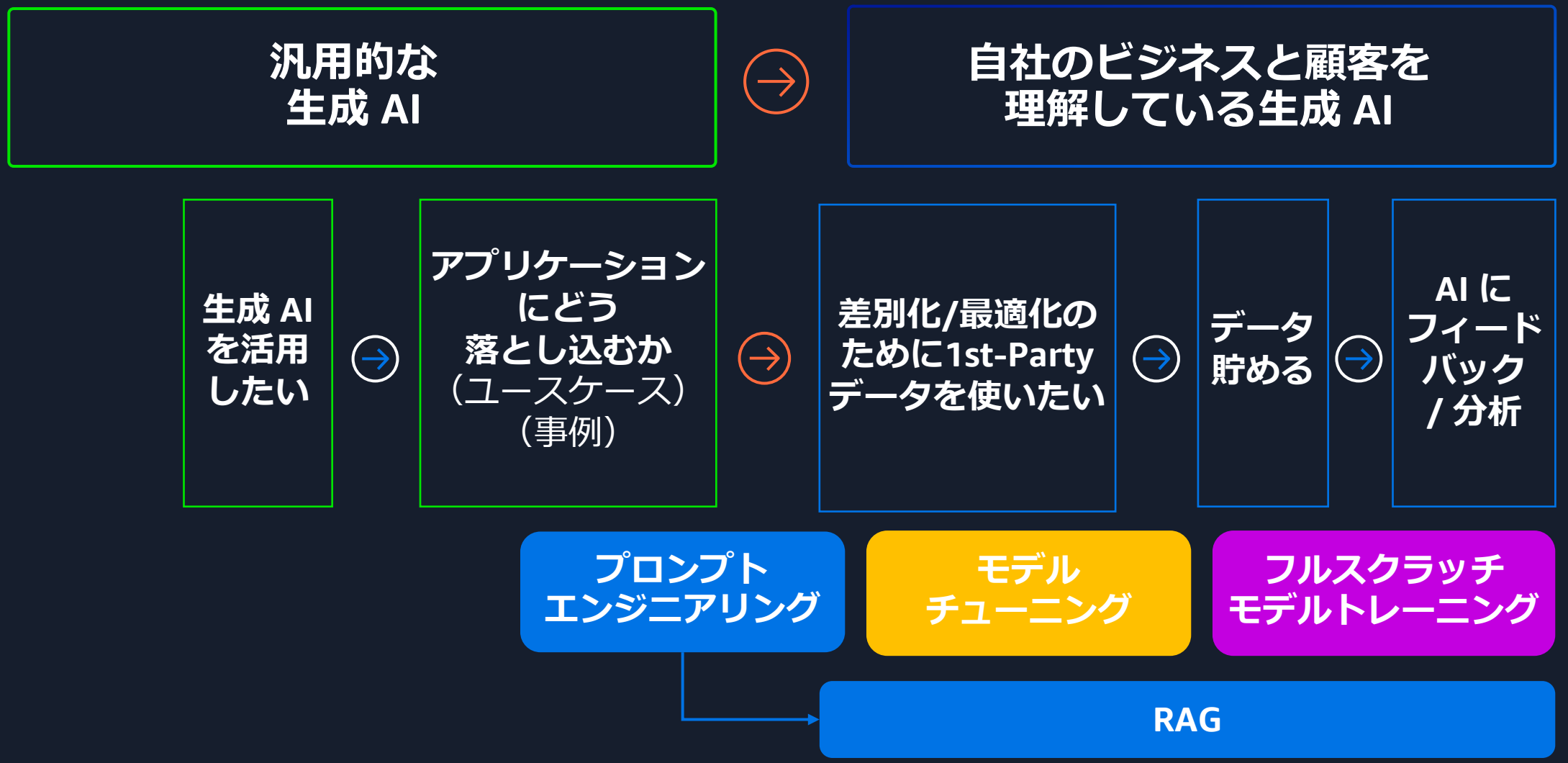
モデルそのものをカスタマイズする

新たな 生成 AI のデータパターンと難易度



** Qualitative comparison and not drawn to scale*

本日のセッションとゴールの確認



プロンプトエンジニアリング

プロンプトエンジニアリングとは

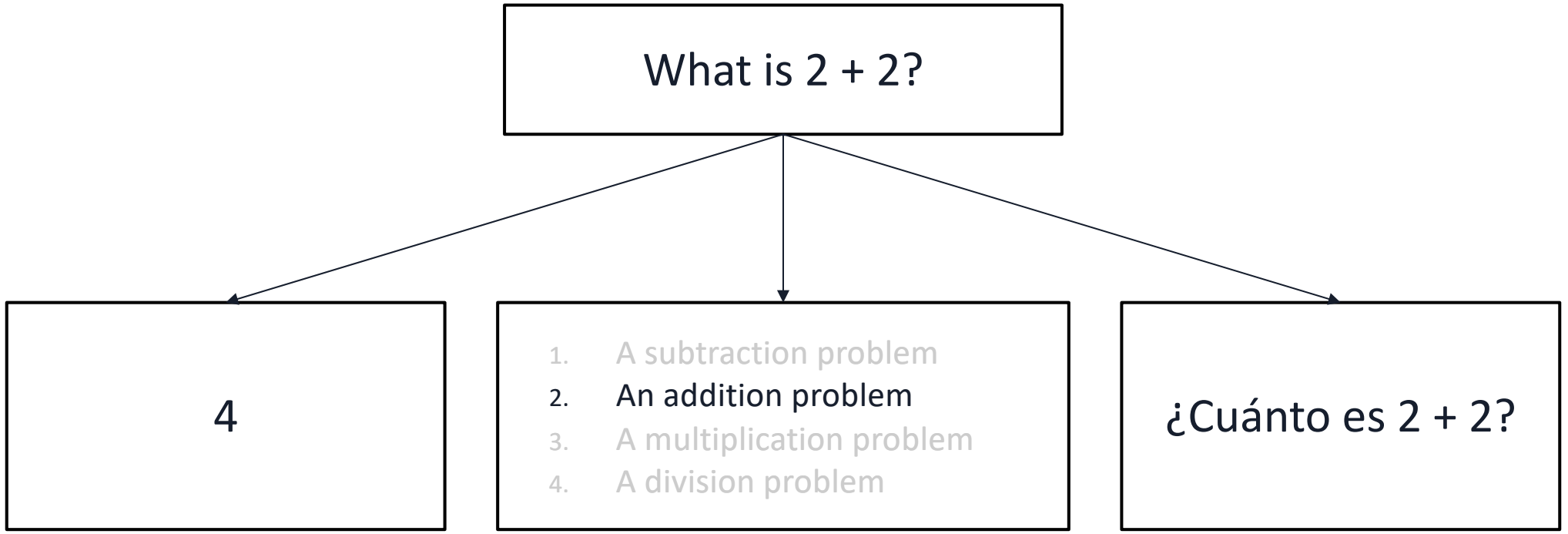
生成系 AI で適切な出力を得るためには、入力の指示文章(プロンプト)をうまく設計する必要があります

以下のようなプロンプトが有効とされています

- タスクを具体的に指示する
- 否定系を使わない
- 望ましい出力の例を含める

など

What is prompt engineering?



Prompt engineering is the process of **controlling model behavior** by **optimizing your prompt to elicit high performing LLM responses** (as assessed by rigorous evaluations tailored to your use case).

https://docs.anthropic.com/en/docs/intro-to-claude#claude-3-technical-presentation

Parts of a prompt

1. Task context
2. Tone context
3. Background data, documents, and images
4. Detailed task description & rules
5. Examples
6. Conversation history
7. Immediate task description or request
8. Thinking step by step / take a deep breath
9. Output formatting
10. Prefilled response (if any)

Example:

User	You will be acting as an AI career coach named Joe created by the company AdAstra Careers. Your goal is to give career advice to users. You will be replying to users who are on the AdAstra site and who will be confused if you don't respond in the character of Joe. You should maintain a friendly customer service tone. Here is the career guidance document you should reference when answering the user: <guide>{{DOCUMENT}}</guide> Here are some important rules for the interaction: - Always stay in character, as Joe, an AI from AdAstra careers - If you are unsure how to respond, say "Sorry, I didn't understand that. Could you repeat the question?" - If someone asks something irrelevant, say, "Sorry, I am Joe and I give career advice. Do you have a career question today I can help you with?" Here is an example of how to respond in a standard interaction: <example> User: Hi, how were you created and what do you do? Joe: Hello! My name is Joe, and I was created by AdAstra Careers to give career advice. What can I help you with today? </example> Here is the conversation history (between the user and you) prior to the question. It could be empty if there is no history: <history> {{HISTORY}} </history> Here is the user's question: <question> {{QUESTION}} </question> How do you respond to the user's question? Think about your answer first before you respond. Put your response in <response></response> tags.
Assistant (prefill)	<response>

https://docs.anthropic.com/en/docs/intro-to-claude#claude-3-technical-presentation

Prompt Engineering: Few-shot Prompting

指示を出す際に、望ましい回答を例示することで、意図した出力を得ます

アンケートの回答結果を、
下記の例に沿って分類してください。
<example>
最寄り駅から10分は歩く必要がありました：施設
音響が良くなく、音がひび割れていました：施設
ドリンクの種類が豊富でした：料理
従業員の応対はとても親切でした：人
軽食はどれもおいしかったです：料理
</example>
<question>
館内の清掃はとても行き届いていました：
</question>

質問のContextを別段落で追加することも多い

RAG

新たな 生成 AI のデータパターン

0

プロンプトエンジニアリング

1

検索拡張生成（RAG)

事前にトレーニングされたモデルに、自社のドメイン固有のコンテキストデータを加えた結果を生成

例: ドメイン固有の、要件が限定された仮想エージェント

2

既存モデルのファインチューニング

既存のモデルをベースに、ドメイン固有のデータによる追加学習を実施

例: ドメイン特化型の知識エージェント

3

独自モデルの作成

ドメイン固有データによるモデルの関するモデルのトレーニングを実施

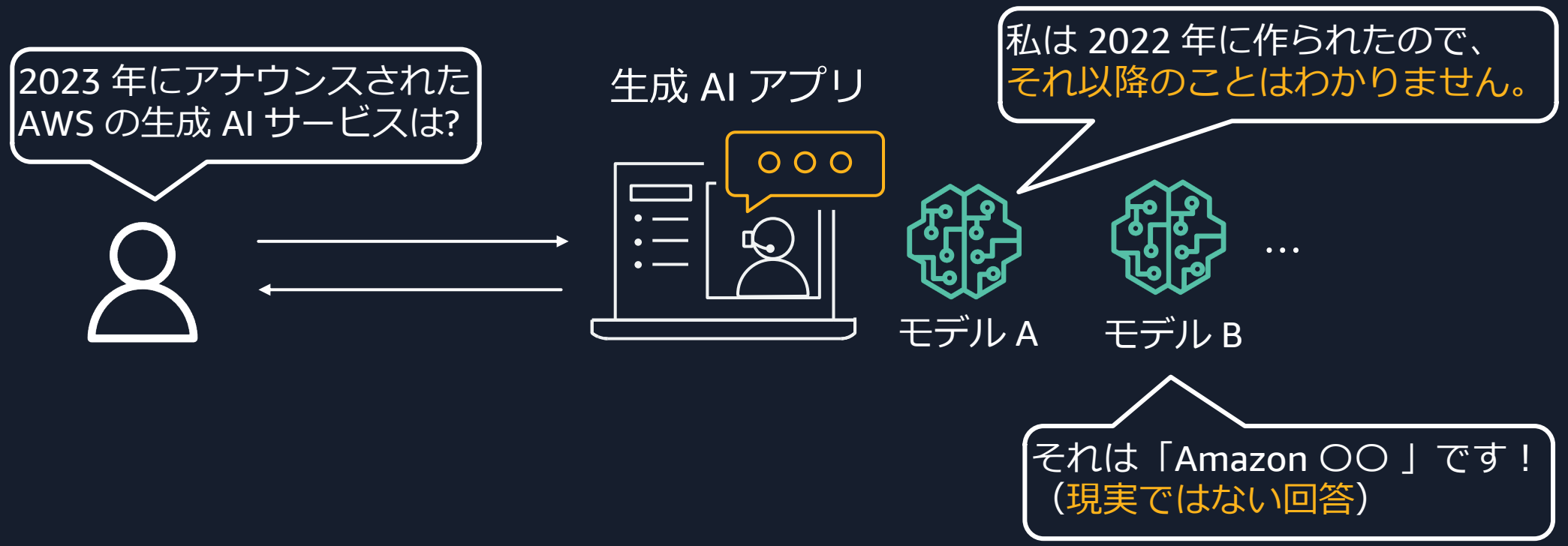
例: 特定のデータで学習を行った、高度にドメインに特化したアプリケーション

一般のモデルを利用し、
入力するデータをカスタマイズ

モデルそのものをカスタマイズする

よくある課題：お客様独自のデータの活用

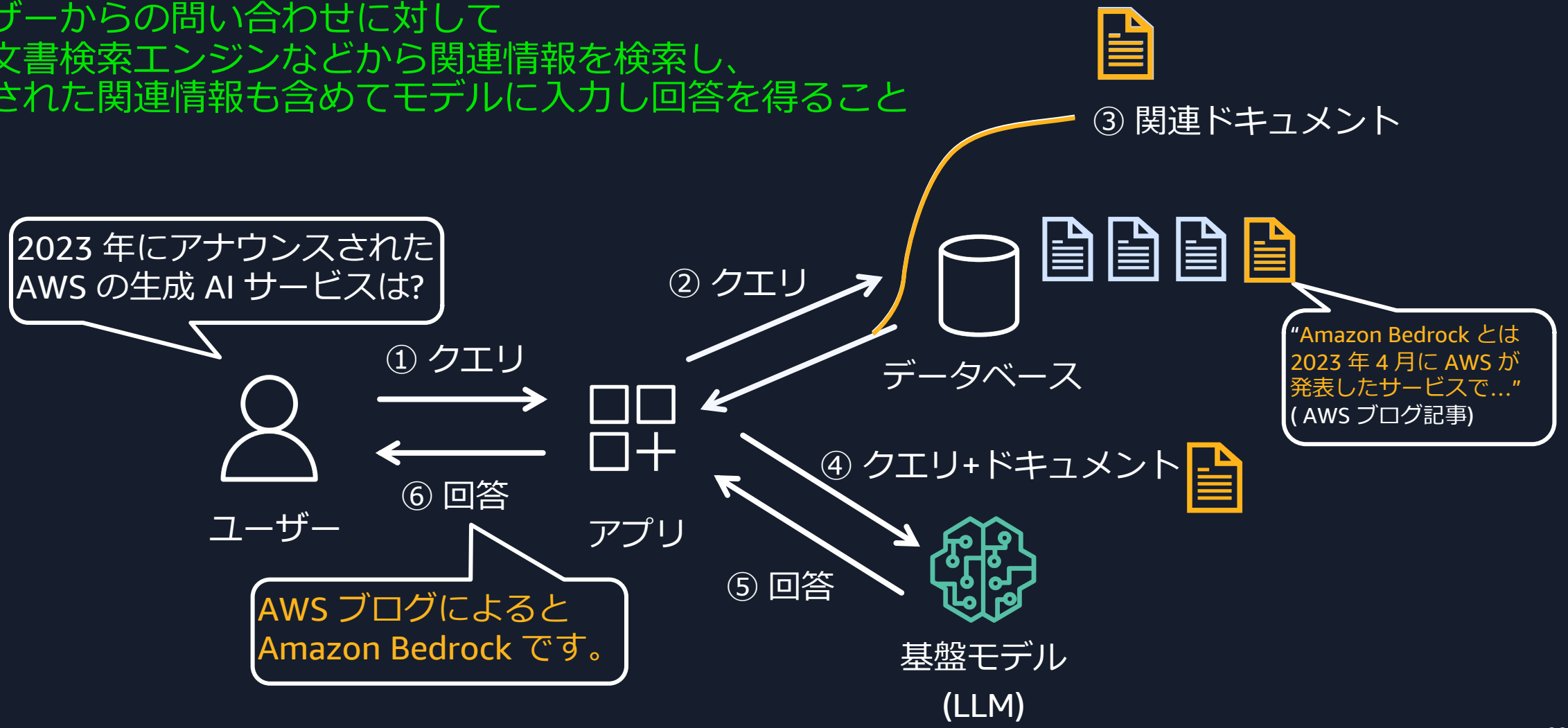
- 基盤モデルは学習データに含まれていないような情報は回答できない。場合によっては不正確な回答をしてしまう。（=ハルシネーション）
- 社内の情報やドキュメントなどを回答に含めたい場合に陥る課題



RAG とは

Retrieval-Augmented Generation (検索拡張生成)

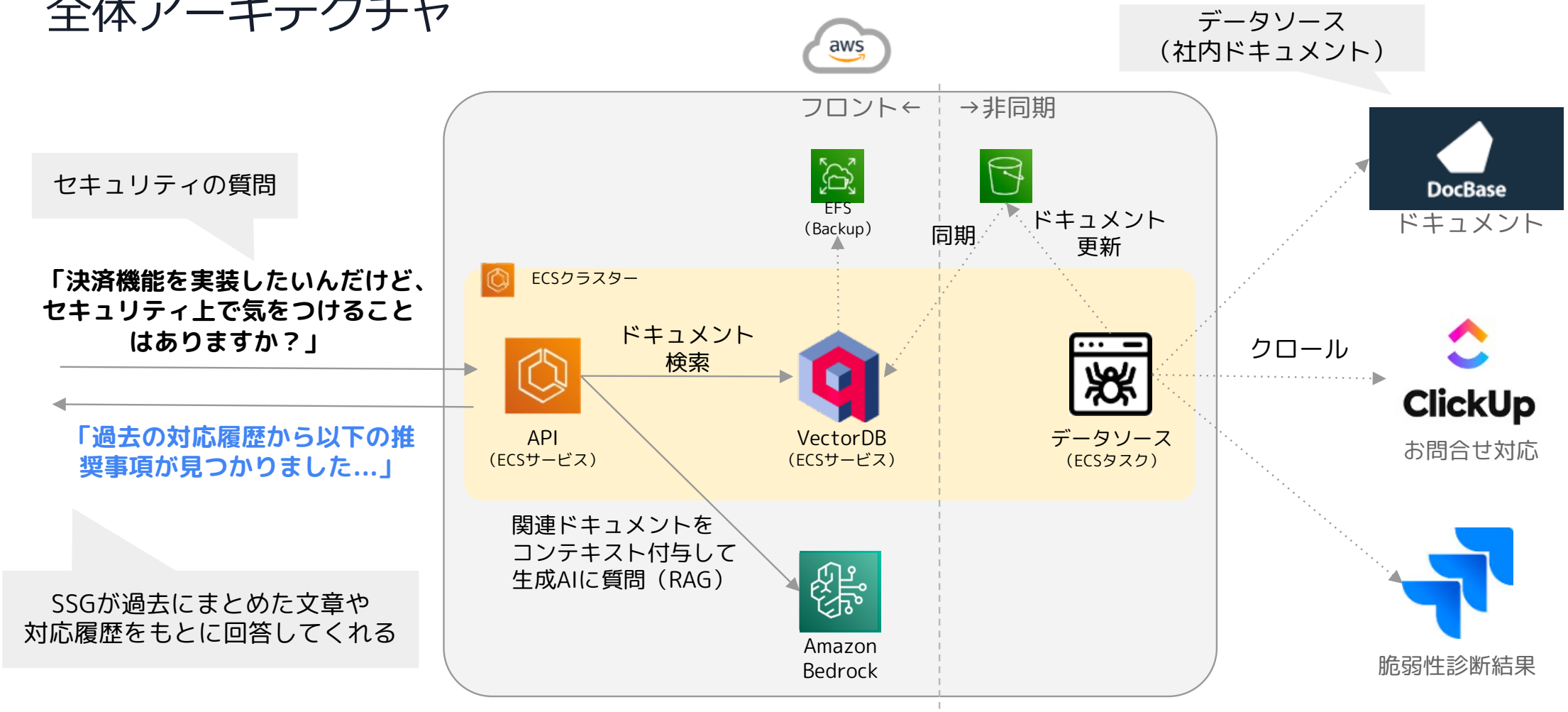
ユーザーからの問い合わせに対して
社内文書検索エンジンなどから関連情報を検索し、
取得された関連情報も含めてモデルに入力し回答を得ること



※ LLM = 大規模言語モデル

CyberAgent SSG 様のセキュリティデスク RAG 事例

全体アーキテクチャ



https://developers.cyberagent.co.jp/blog/archives/45548/

現状のAWSのRAG俯瞰図

完全フル
マネージド

手動作成

Amazon Q Business

[Retriever] ・Native ・Kendra ＋ [Data Source]

Knowledge Base for Amazon Bedrodk

[VectorDB] ・OpenSearch SL ・PineCorn ・Redis・Aurora (pgVector) ・MongoDB
＋ Bedrock (FM)

LangChain

[フルマネージド検索] ・Kendra

＋LLM

[VectorDB] ・OpenSearch SL ・Document DB ・Memory DB ・pgVector (Postgres DB)

[Embedding] ・Titan ・Cohere (Bedrock)

Knowledge bases for Amazon Bedrock

検索拡張生成 (RAG) の
フルマネージドネイティブサポート

テキスト文書を埋め込み形式に
自動的に変換

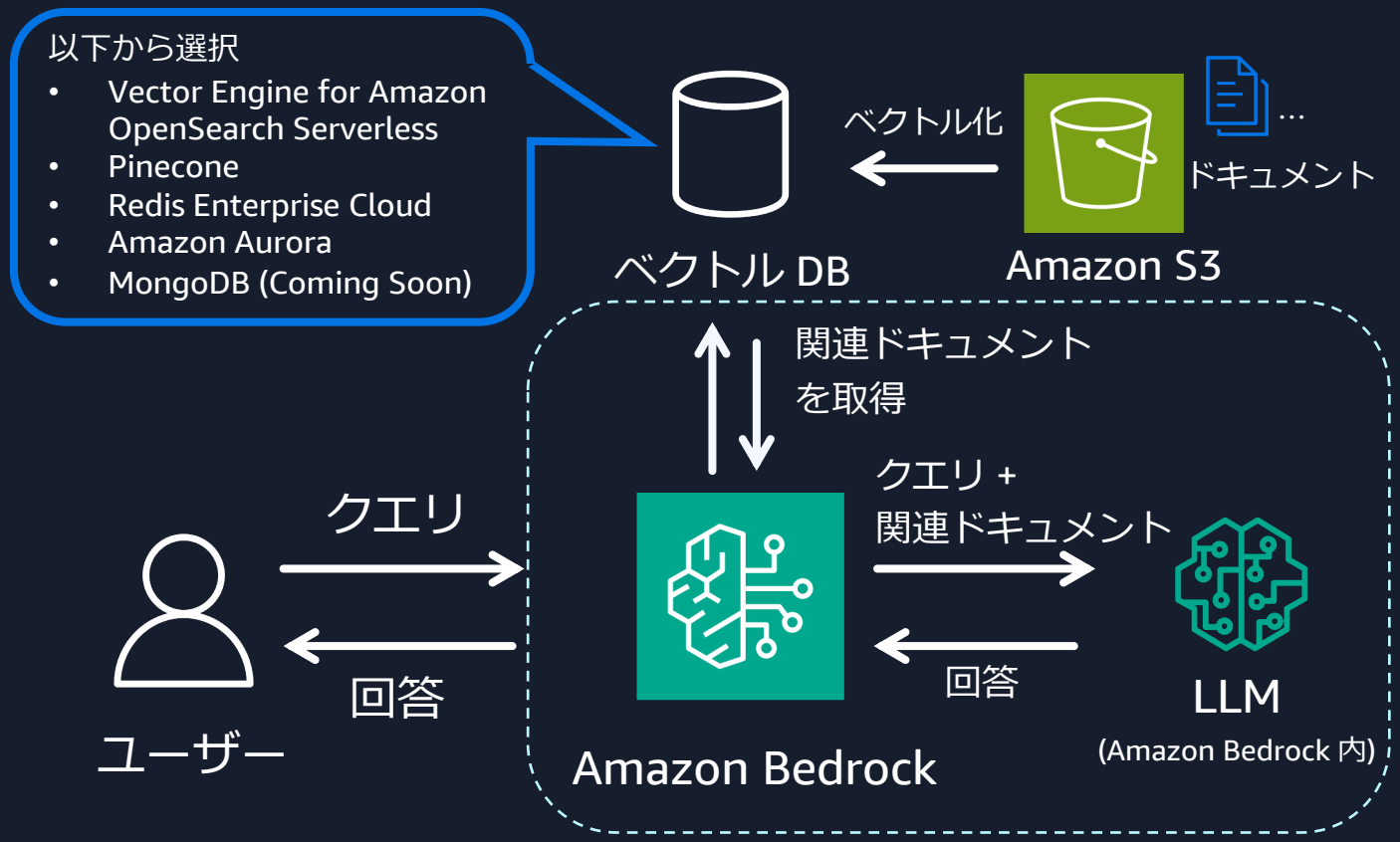
変換した埋め込み形式データを
ベクトルデータベースに保存

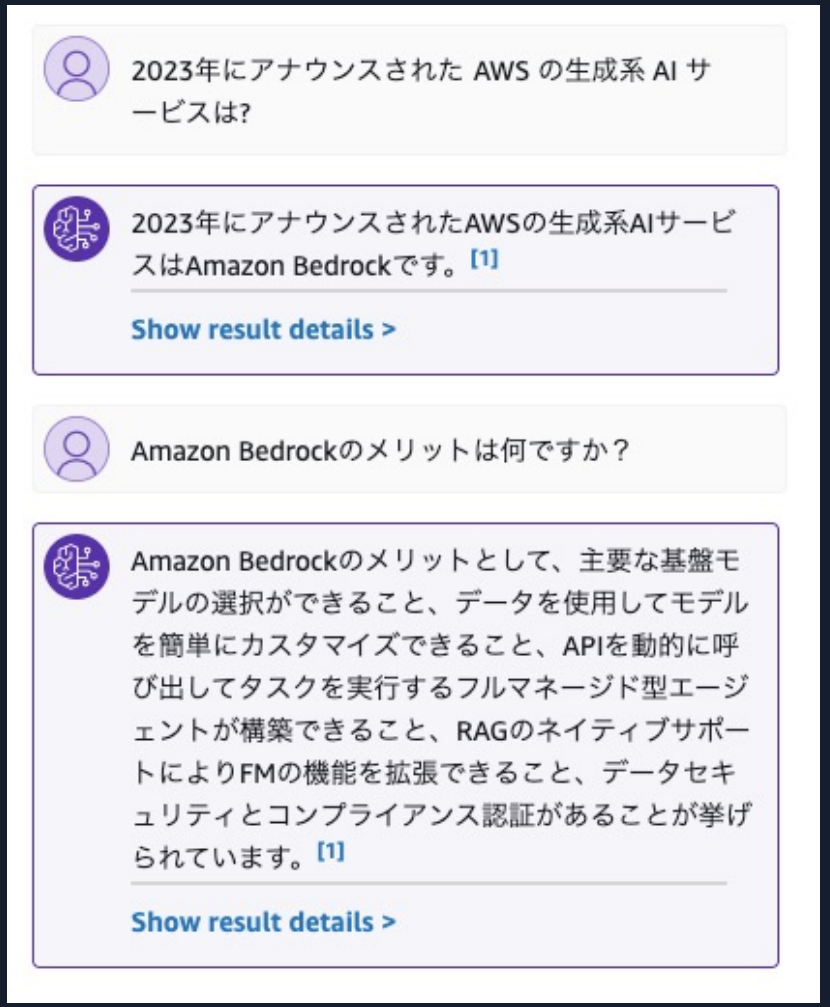
保存した埋め込み形式データを
取得してプロンプトを拡張

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

Knowledge Base for Amazon Bedrock

- Vector database は、OpenSearch Serverless、Redis Enterprise Cloud、Pinecone、Aurora、MongoDB から選択できる

現状のAWSのRAG俯瞰図

完全フル
マネージド

Amazon Q Business

[Retriever] ・Native ・Kendra ＋ [Data Source]

Knowledge Base for Amazon Bedrodk

[VectorDB] ・OpenSearch SL ・PineCorn ・Redis・Aurora (pgVector) ・MongoDB
＋ Bedrock (FM)

手動作成

LangChain

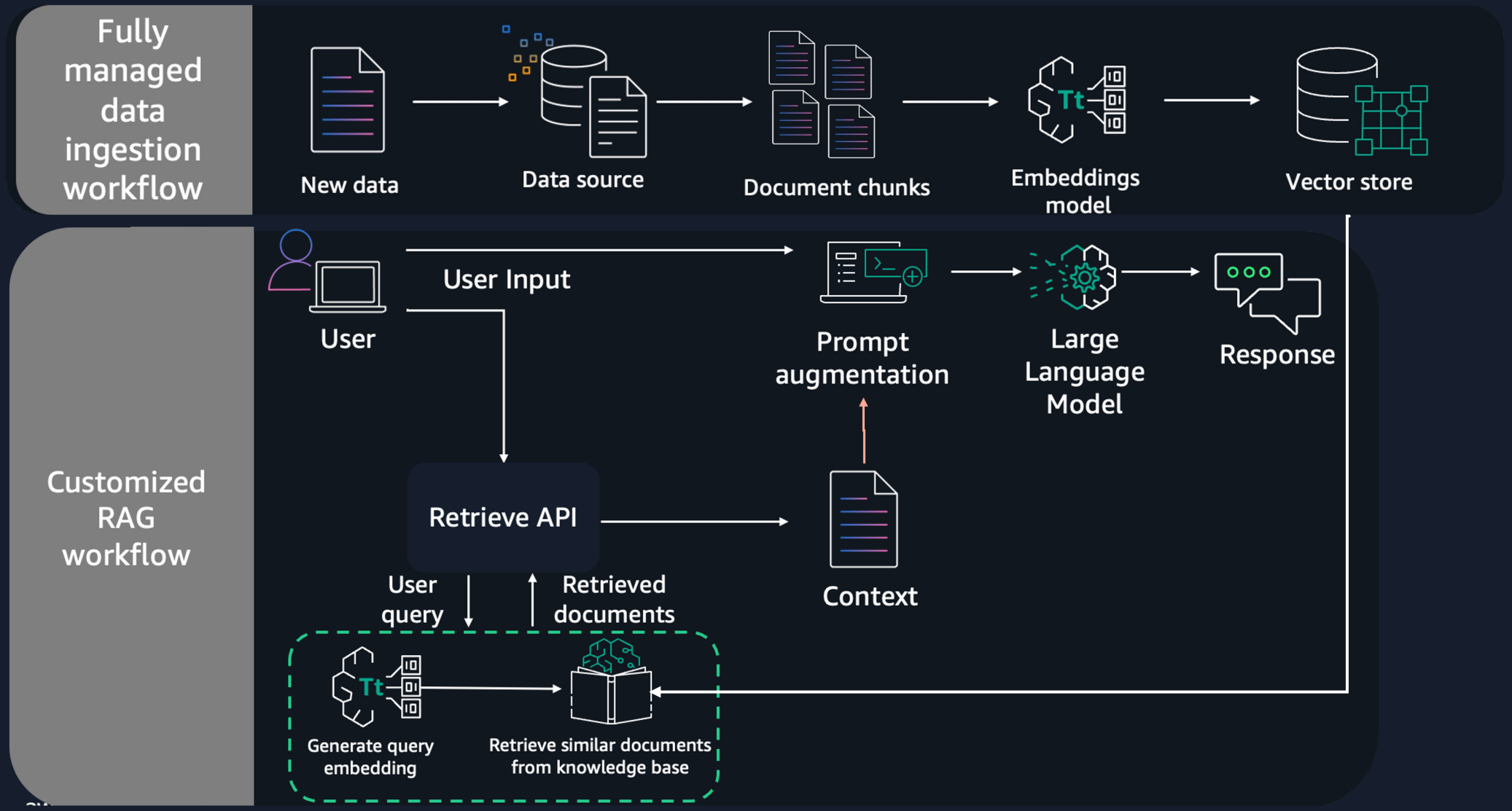
[フルマネージド検索] ・Kendra

[VectorDB] ・OpenSearch SL ・Document DB ・Memory DB ・pgVector (Postgres DB)

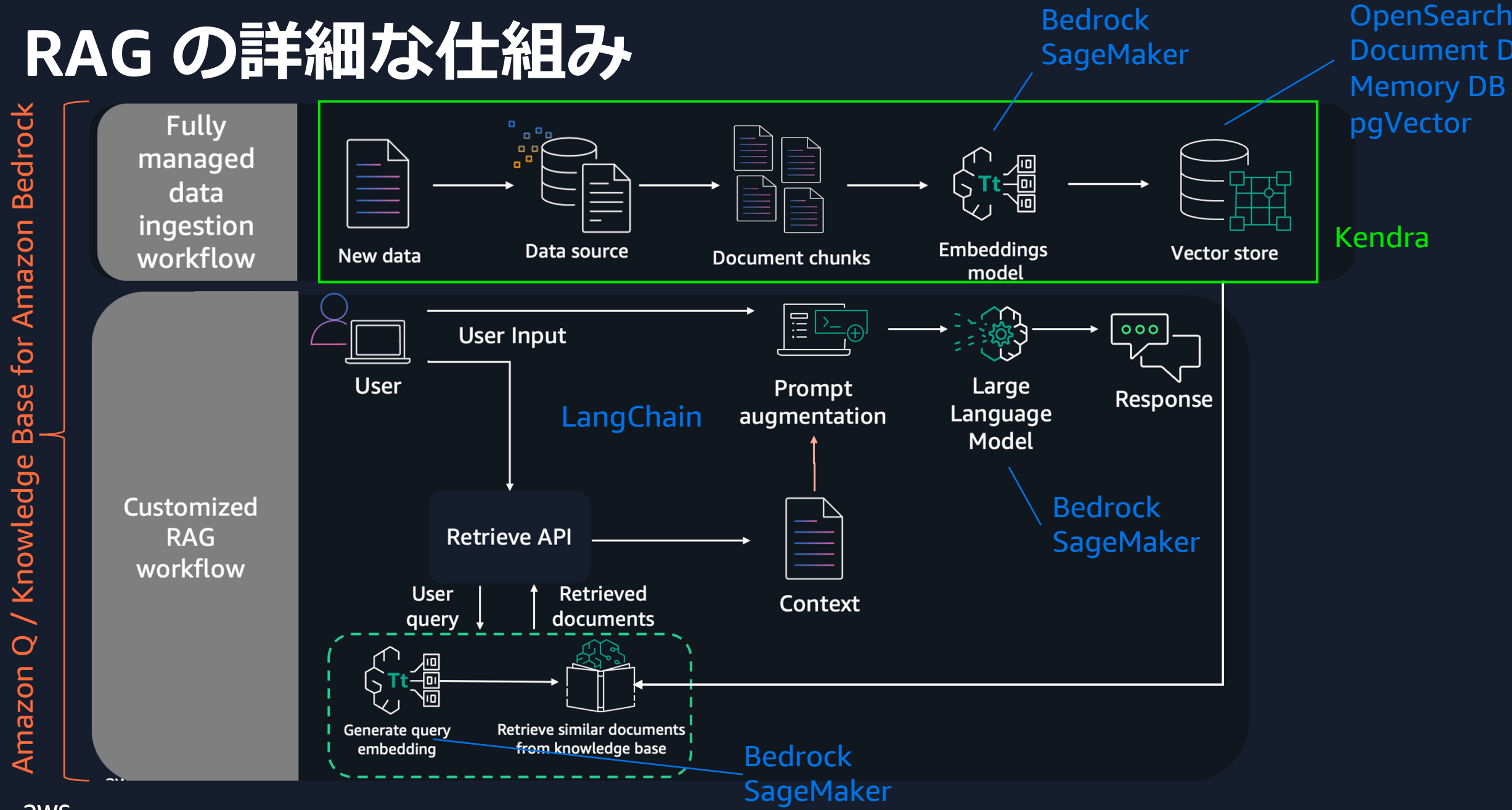
[Embedding] ・Titan ・Cohere (Bedrock)

＋LLM

RAG の詳細な仕組み



RAG の詳細な仕組み



基盤モデルのカスタマイズ

新たな 生成 AI のデータパターン

0

プロンプトエンジニアリング

1

検索拡張生成（RAG）

事前にトレーニングされたモデルに、自社のドメイン固有のコンテキストデータを加えた結果を生成

例: ドメイン固有の、要件が限定された仮想エージェント

2

既存モデルのファインチューニング

既存のモデルをベースに、ドメイン固有のデータによる追加学習を実施

例: ドメイン特化型の知識エージェント

3

独自モデルの作成

ドメイン固有データによるモデルの関するモデルのトレーニングを実施

例: 特定のデータで学習を行った、高度にドメインに特化したアプリケーション

一般のモデルを利用し、入力するデータをカスタマイズ

モデルそのものをカスタマイズする

ドメイン学習による基盤モデルの最適化

経済情報特化の生成AI、日経が開発　40年分の記事学習

生成AI　＋フォローする

2024年4月24日 11:00 (2024年4月26日 13:53更新)

保存

性能の指標となる「パラメーター数」が130億と700億の2種類を開発した

日本経済新聞社は24日、経済情報に特化した生成AI（人工知能）の基盤技術を開発したと発表した。大規模言語モデルと呼ばれるもので、約40年分の日経グループの新聞や雑誌の記事を学習させた。記事の要約機能などで活用を見込む。

パラメーター数が700億のモデルは米メタが2023年7月に公開した大規模言語モデル「Llama（ラマ）2」をベースに、「ファインチューニング（追加学習）」と呼ぶ手法を使って開発した。ラマ2は外部の企業や開発者による改良が広く認められている。

開発したモデルは最新ニュースに関する知識を備え、記事の要約などで性能を発揮することを確認した。現在はメタが4月に公開した最新版の「ラマ3」をベースに、パラメーター数が80億のモデルの学習にも着手している。

日経イノベーション・ラボは19年に言語モデルの研究を始め、著作権上の課題やAIがもっともらしく嘘をつくハルシネーション（幻覚）の問題に対処する方法を探ってきた。

https://www.nikkei.com/article/DGXZQOUC1941R0Z10C24A4000000/

Fine Tuning / Continued Pre-training

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

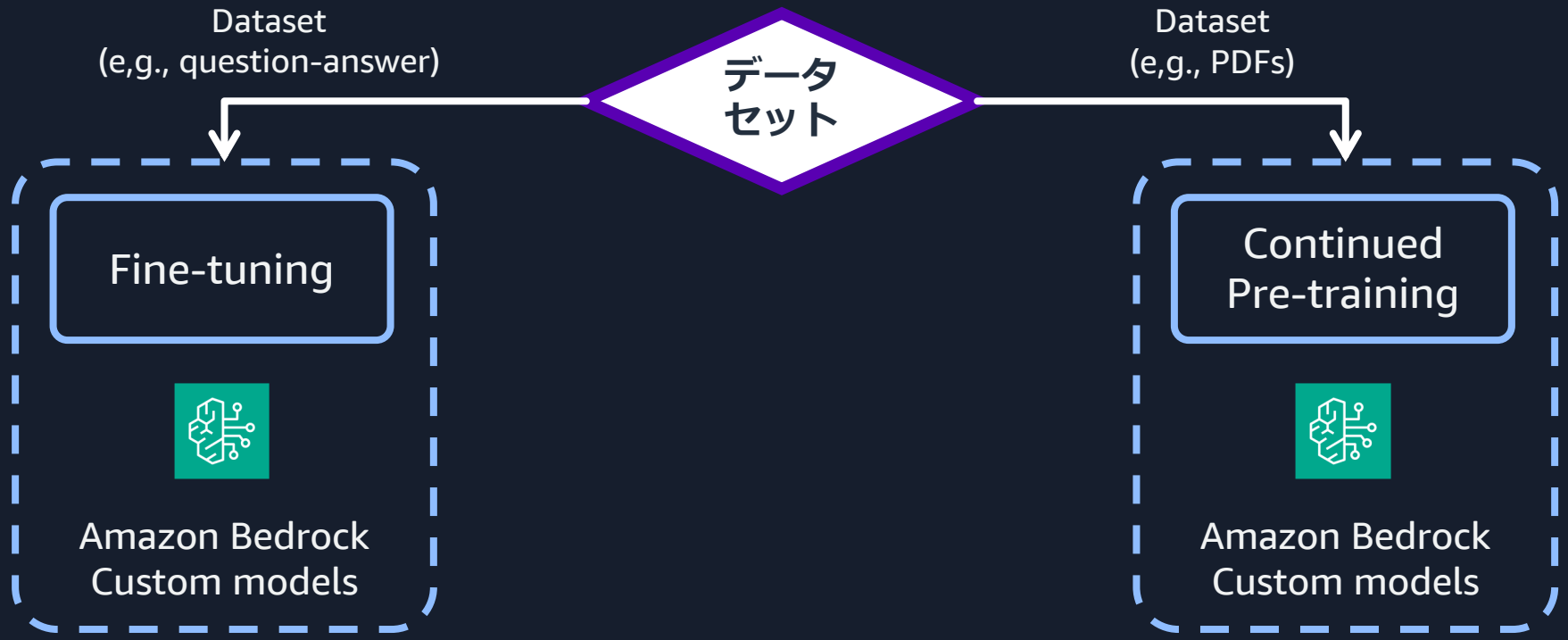
ラベル付きデータを新たに学習して
特定のタスクの精度を高める手法

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

Fine-tuning と Continued pre-training の違い



{“prompt”: “日本の首都は？”, “completion”: “東京なのだ"}
{“prompt”: “アメリカの首都は？”completion“: ”ワシントンなのだ"}
{“prompt”: ”日本の次の元号は？“, ”completion“: “うーん、わからないのだ"}

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

Generative AI tools and services

APPLICATIONS THAT LEVERAGE LLMs AND FMs

- Amazon Q Business
- Amazon Q Developer
- Amazon Q in QuickSight
- Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR MODEL TRAINING & INFERENCE

- GPUs
- Trainium
- Inferentia
- SageMaker
- UltraClusters
- EFA
- EC2 Capacity Blocks
- Nitro
- Neuron

Amazon Bedrock Custom Models

Bedrock に搭載されている一部のモデルを対象として
Custom Models 機能でモデルのファインチューニング / Continued pre-training が可能に

一般利用開始

Amazon Bedrockでモデルのファインチューニングが可能に

- ラベル付きの小規模なデータセットで、モデルを特定のタスクに特化させる手法がファインチューニング
- Amazon Bedrockで選択できる以下のモデルで、ファインチューニングを実行可能に
  - Cohere Command Lite, Meta Llama 2, Amazon Titan Text Lite/Express
- バージニアとオレゴンにて一般利用開始

Amazon Bedrock

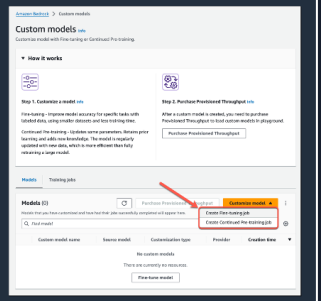
AWS RE:INVENT RE:CAP

9

プレビュー

Amazon BedrockのContinued pre-trainingを発表

- Amazon Titan Text Express/Liteについて、安全で管理された環境を利用して独自のラベルなしデータによるカスタマイズが可能に
  - ドメイン知識を必要とするアプリケーションを構築するには、一般的な公開データでトレーニングされた基盤モデルでは不十分な場合がある
  - 用途に応じた知識を基盤モデルに与えるために、ユーザが提供する独自データを加えて継続的に事前トレーニングを行うことが可能
- バージニアとオレゴンでプレビューを開始



AWS RE:INVENT RE:CAP

10

Amazon SageMaker JumpStart
数百のモデルが活用可能

AI21labs

モデル
Jurassic-2 Ultra, Mid

状況に応じた回答
テキスト要約
言い換え
文法エラー訂正

タスク
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

Meta AI

モデル
Llama 2 7B, 13B, 70B

タスク
質問回答
チャット
テキスト要約
言い換え
感情分析
テキスト生成

cohere

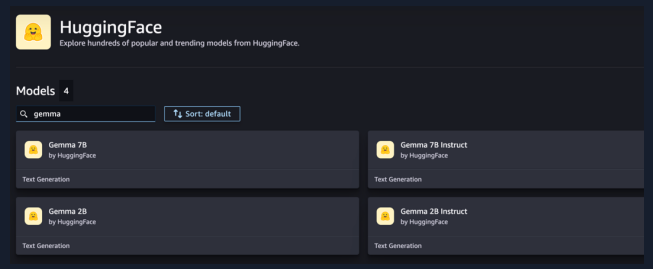
モデル
Cohere
Command XL

タスク
テキスト生成
情報抽出
質問回答
テキスト要約

Hugging Face

モデル
Falcon-7B, 40B
Open LlaMA
RedPajama
MPT-7B
BloomZ 176B
Flan T-5 models (8 種類)
DistilGPT2
GPT NeoXT
Bloom models
(3 種類)

タスク
機械翻訳
質問回答
テキスト要約



stability.ai

モデル
Stable Diffusion XL 1.0
2.1 base
高精細化
インペインティング

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

特長
Stable Diffusion 2.1
ファイン・チューン対応

LightOn

モデル
Lyra-Fr
10B, Mini

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
感情分析
テキスト分類

databricks

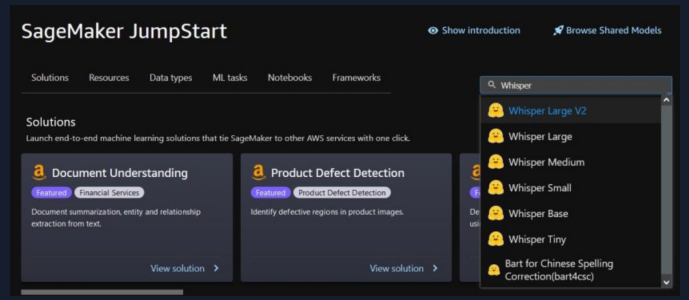
モデル
Dolly

タスク
質問回答
チャット
言い換え
感情分析
テキスト生成

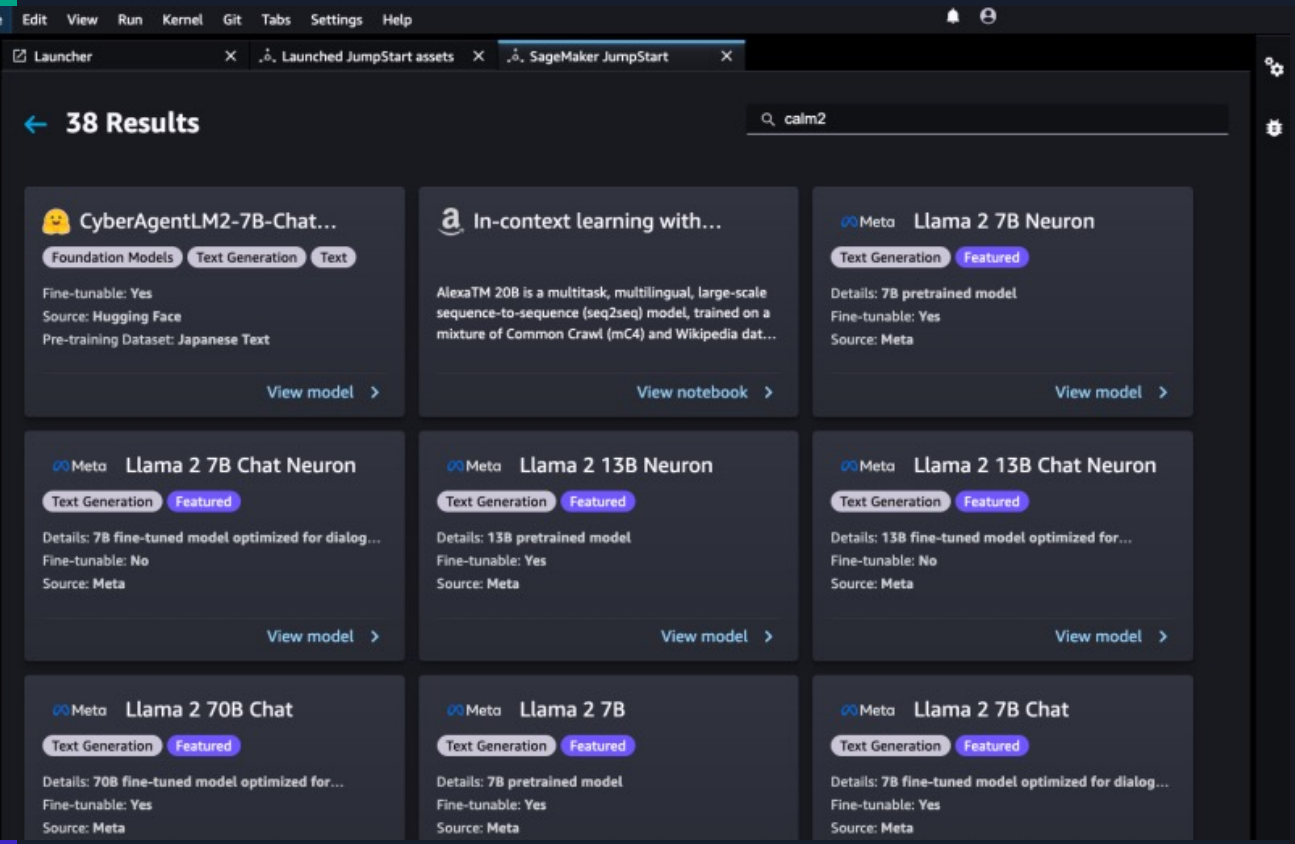
alexa

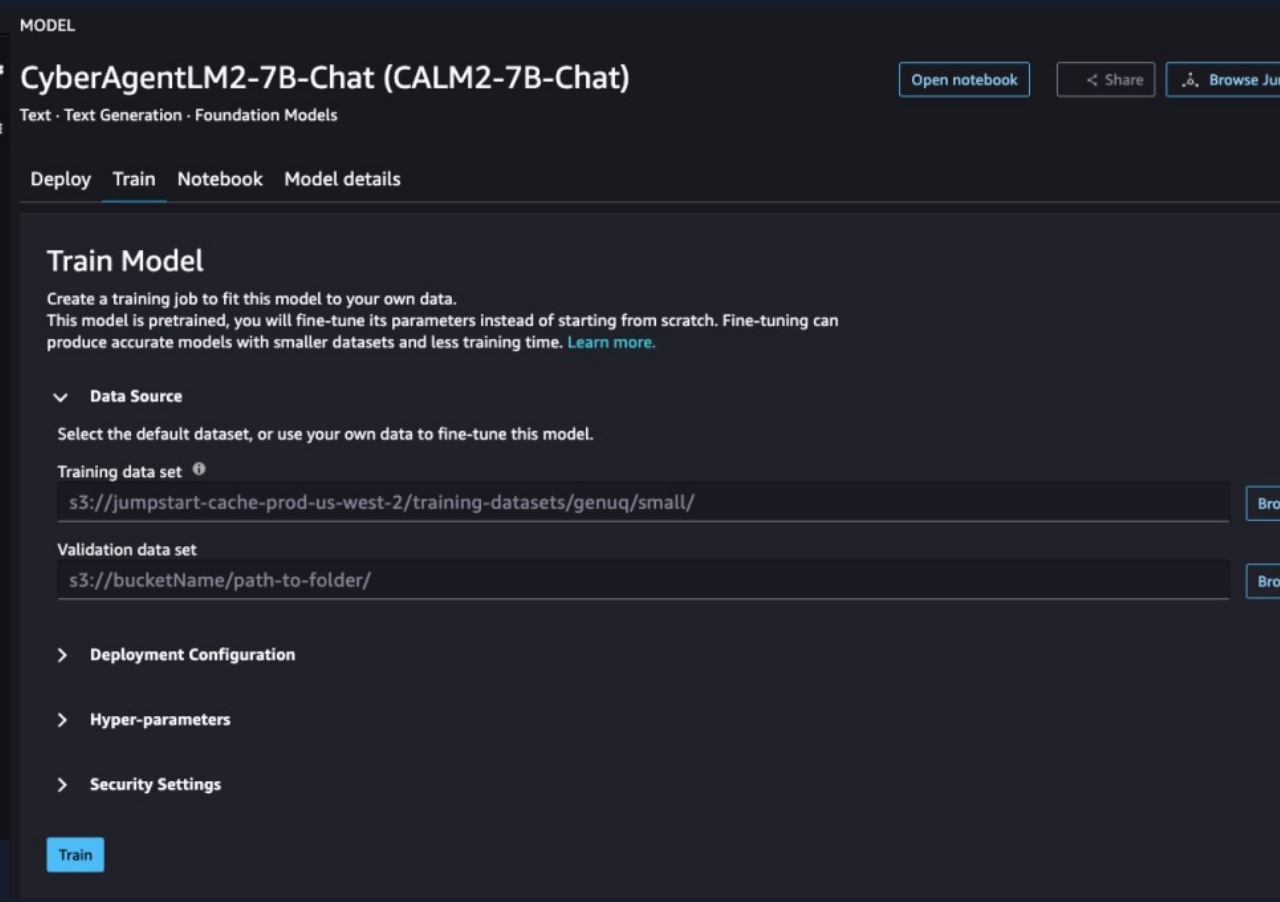
モデル
AlexaTM 20B

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



Amazon SageMaker JumpStart で Fine-tuning / Continued Pre-training





モデルを選んで、データを用意してクリックするだけ

独自モデルのフルスクラッチ開発

APPLICATIONS THAT LEVERAGE LLMs AND FMs

Amazon Q Business

Amazon Q Developer

Amazon Q in QuickSight

Amazon Q in Connect

TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock Guardrails | Agents | Customization Capabilities

INFRASTRUCTURE FOR MODEL TRAINING & INFERENCE

GPUs

Trainium

Inferentia

SageMaker

UltraClusters

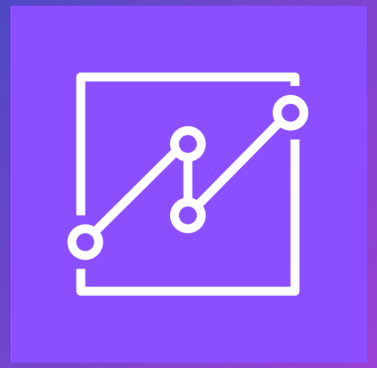
EFA

EC2 Capacity Blocks

Nitro

Neuron

データの可視化と生成 AI



Amazon QuickSight

UNIFIED BI SERVICE
AT HYPERSCALE

- あらゆるデータ活用を実現し BI環境を統合
- オートスケーリングで高パフォーマンスを実現
- 生成系AI機能で分析をさらに増強
- 従量課金制でコスト削減

Amazon Q in QuickSight

あらゆるユーザのデータ活用の意思決定過程を加速する

AIにより加速された ダッシュボード作成

ビジネスアナリストがより早くビルドを実現

AIがオンデマンドでデータの質問に答える

ビジネスユーザによるインサイト抽出を簡単に

AIアシスタントによる データストーリー

ビジネスユーザによる発見や所見を共有可能に

カスタムアプリにAIを拡張

開発者にも高度なAI機能をすばやく利用可能

Powered by
Amazon Bedrock

※英語のみのサポート
※2024年5月現在US East (N. Virginia), US West (Oregon), EU(Frankfult)のみでの提供,

AI-powered ダッシュボード作成におけるエクスペリエンス

A NEW DASHBOARD BUILDING EXPERIENCE POWERED BY GENERATIVE BI

ビジュアルのビルド

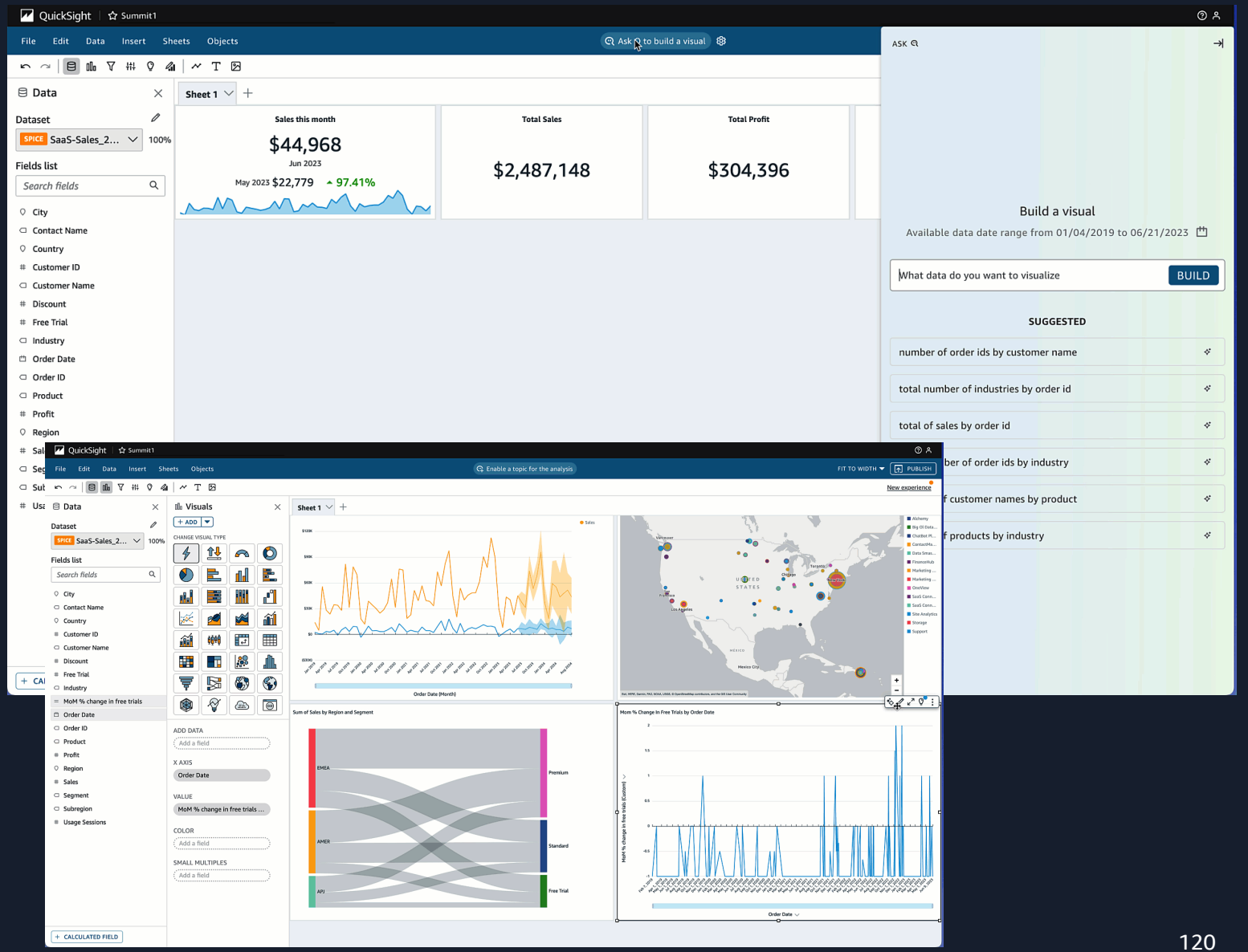
ダッシュボードやレポートのビジュアルのビルドを自然言語ですばやく実現

計算のビルド

文法を学んだり調べることなく、自然言語で計算式を作成

ビジュアルの設定

設定したいフォーマットを自然言語で表現することにより、ビジュアルの更新をすばやく実現



AIがオンデマンドでデータへの質問に答える

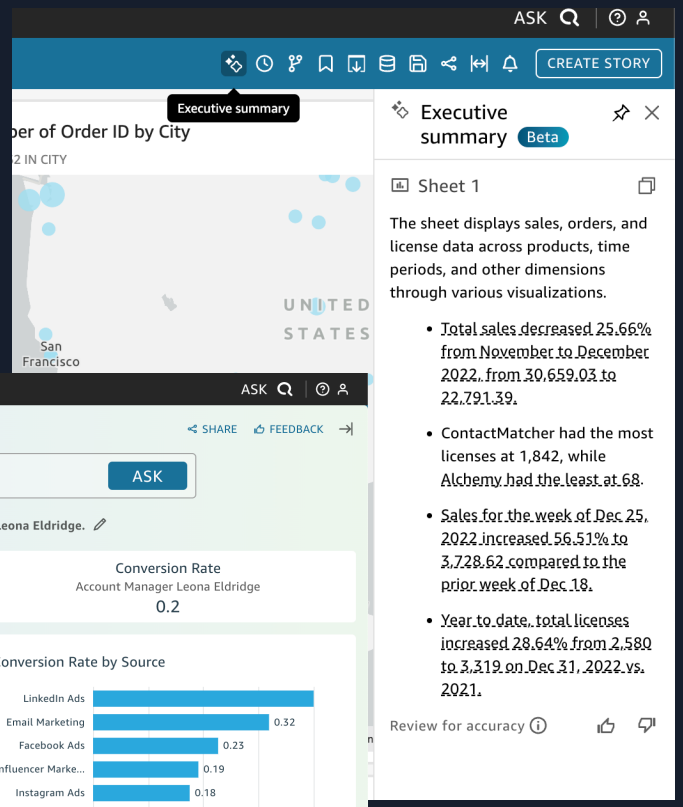
DEEP INSIGHTS AT YOUR FINGERTIPS

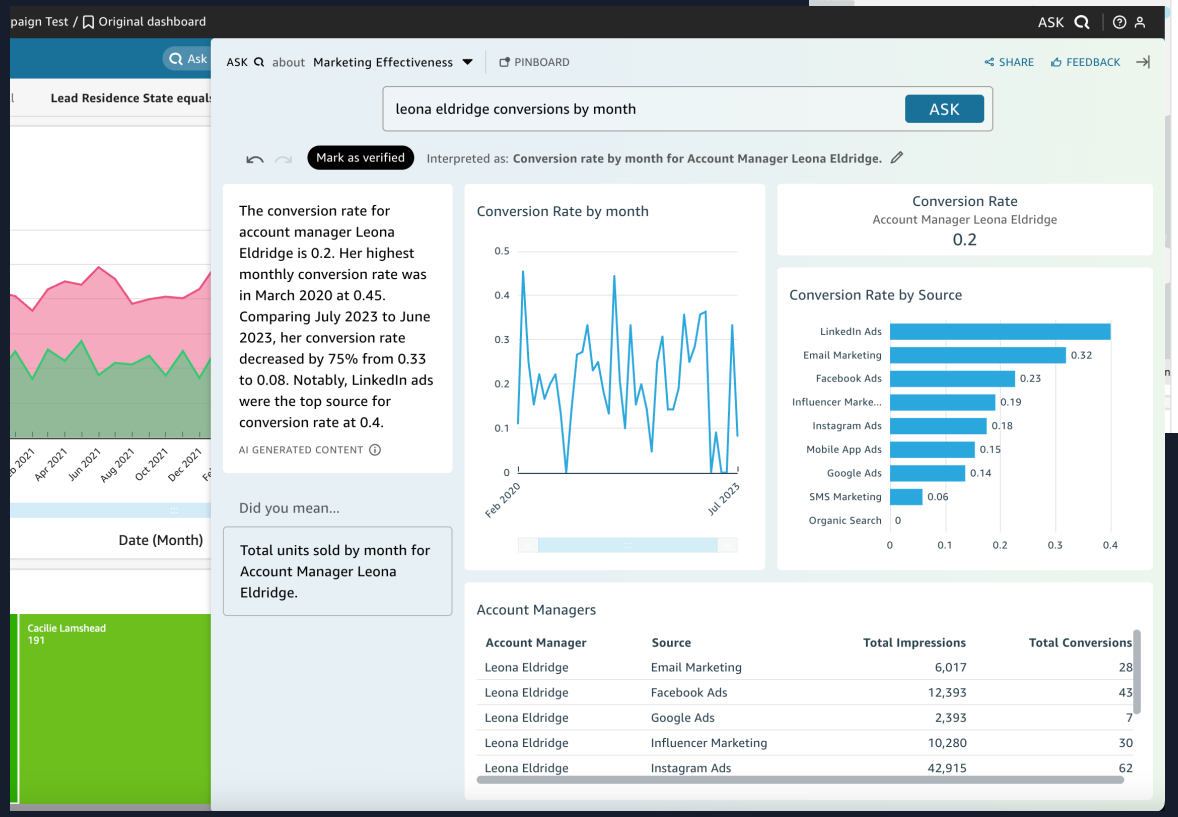
ダッシュボードのエグゼクティブサマリー
自然言語によるダッシュボードの主要なインサイト（上位ムーバー、外れ値など）を速攻で要約

エクスパートでなくても容易に、パワフルなQ&Aが可能
Suggested questionsと**what's in my data**にて質問例が提示

複数のビジュアルによる返答と**インサイトサマリーのナラティブ**にて答えのコンテキストが説明

曖昧な質問サポートや**did-you-mean による代替案** にて、ファクトを繰り返して発見できる





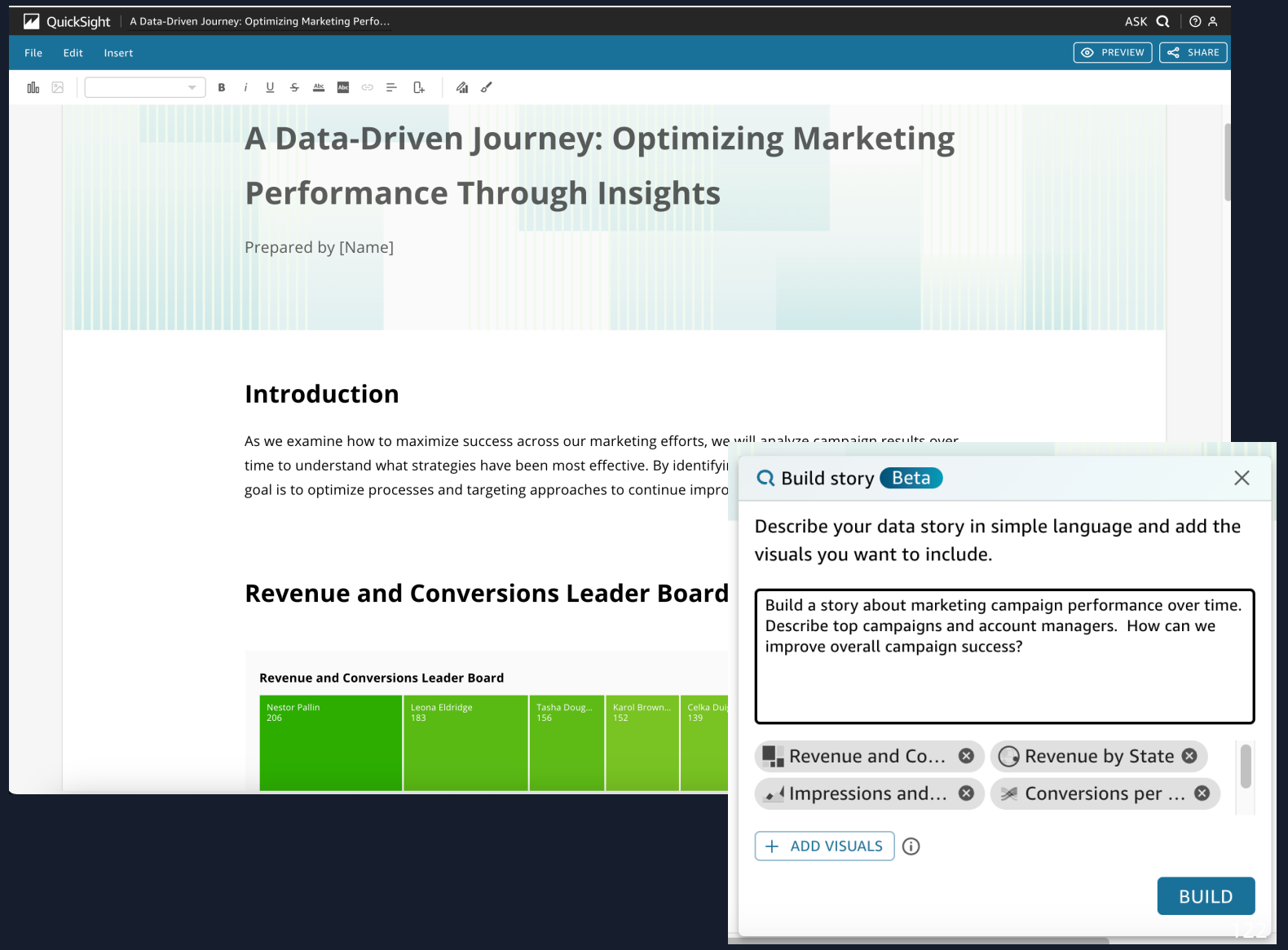
AIアシスタントによる データストーリー

データ解釈の共有
データの意味を説明し意思決定へ導くための結論を示す

AIを使用してストーリーを生成
数個の単語だけで、データを分析し、親密でパワフル、洞察力あるナラティブを生成

洗練されたコンテンツの作成
AIによる長い文章の制御、ナラティブのテキストのカスタマイズ、ビジュアルのテーマ適用により、コンテンツが洗練

統制され常に最新の状態
すばやく容易に共有でき、いつでもデータを更新可能



カスタムアプリにAIを拡張

QUICKLY ADD GENERATIVE BI CAPABILITIES TO WEB APPS

統合を簡単に実現

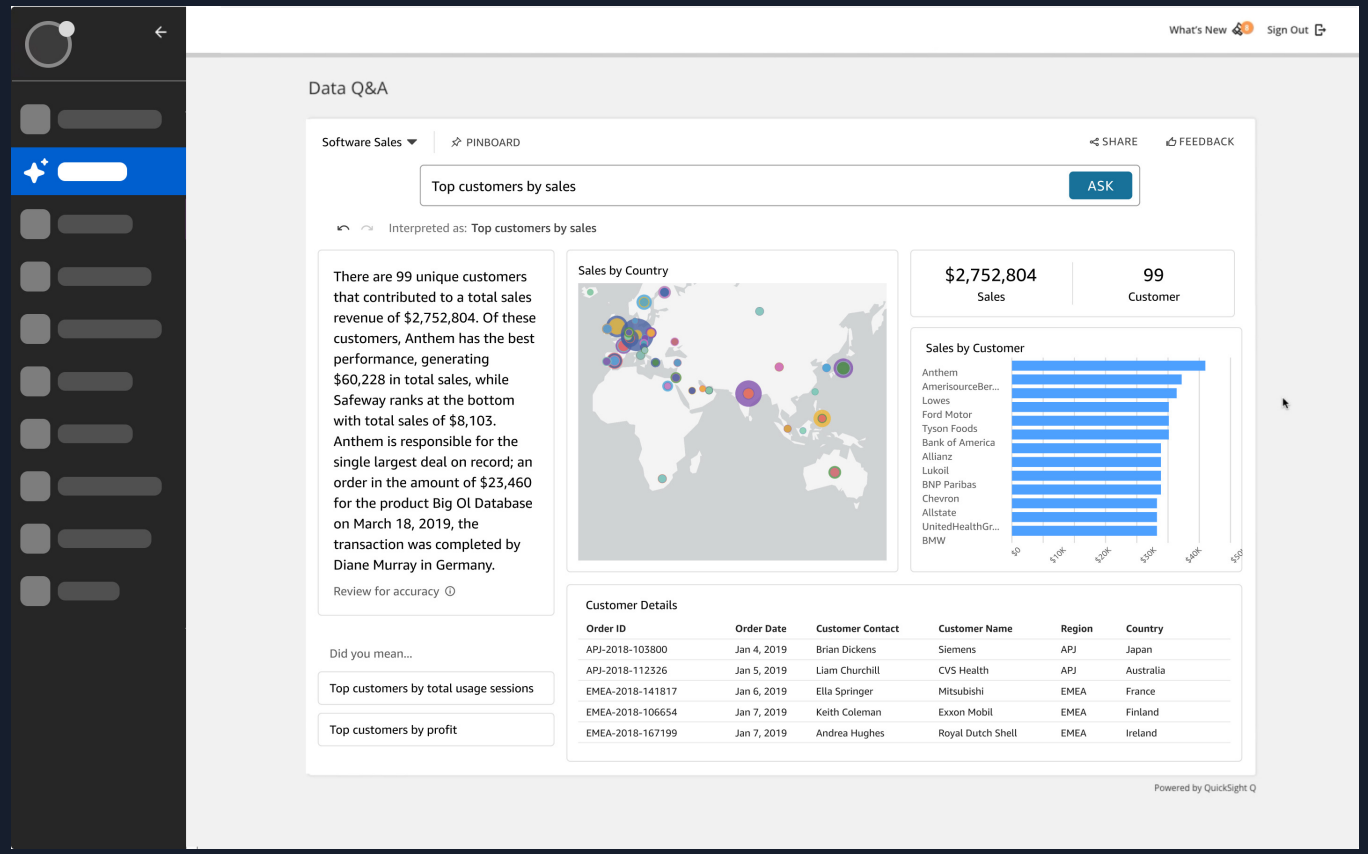
数行のコードのみで、AIの経験なても既存のWebアプリに複数ビジュアルを返すQ&A機能を追加

どのデータでも利用可能

特定のユースケースとデータのコンテキストを理解するため、質問を再定義したり、代替の質問が提案される。

シームレス・エクスペリエンス

ブランドに合わせ、サイズや場所、インタラクションをカスタマイズ可



まとめ

まとめ

なぜ自社データを生成 AI に活用させる必要があるのか？	他社との差別化や、自社のアプリを最適化するために自社の持つ1st party データを生成 AI に利用することで、AI が他社にはないドメイン知識を持ち、タスクの精度が上がる
プロンプトエンジニアリング / RAG	生成 AI へ適用するデータ活用で最も簡単に試せる方法はプロンプトエンジニアリング。RAGを利用することで質問された内容に類するドキュメントを検索して回答に利用できる。
基盤モデルのカスタマイズ	少量データで特定のタスクに特化させる Fine-tuning や、大量のラベルなしデータを用いてドメイン知識をつける Continued pre-training などの方法で既存モデルを自社データでカスタマイズできる。
可視化と生成 AI	QuickSight を利用することで自然言語によるダッシュボードの作成や、データレポートを自動で生成し、既存アプリケーションに埋め込むことが可能に。

Thank you!