



AWSの生成系AIサービス

アマゾン ウェブ サービス ジャパン 合同会社

執行役員 パブリックセクター技術統括本部長

瀧澤 与一

本セッションの内容

- スタートアップからエンタープライズまで、多くの組織が、生成系 AI を使用したイノベーションが生まれています。
- 本セッションでは、生成系AIの基本的な概要から最新のユースケース、生成系AIの動向について説明します。
- また、Amazon Bedrockを含めた、最新のAWSの生成系AIに関連するサービスについても紹介します。

Agenda

- AWSとは？
- 生成系AIの概要とユースケース
- AWSの生成系AIサービス



Amazon Web Services (AWS) とは？



Our mission

Amazon は、

地球上で最もお客様を大切にする企業、

そして地球上で最高の雇用主となり、

地球上で最も安全な職場を提供すること

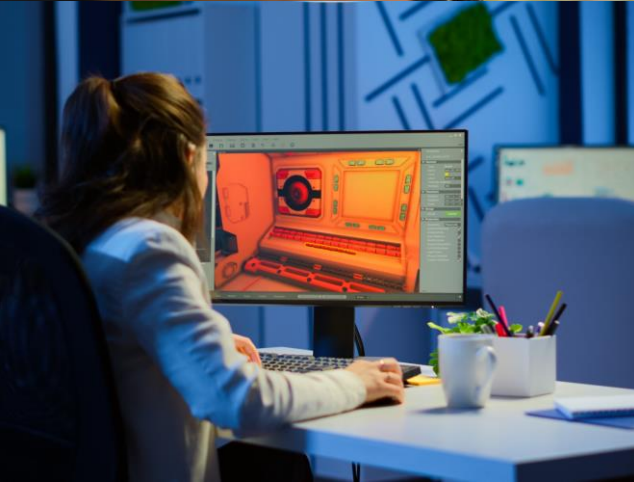
を目指しています。

AWS とは

- 2006 年より、他社にさきがけてクラウドサービスを提供
 - 245 の国と地域、世界数百万、日本では数十万以上のお客様
 - 全国をカバーするパートナーコミュニティ
 - 累計で 134 回以上値下げをして利益をお客様へ還元
-

※ お客様とはアクティブカスタマー数を指します。アクティブカスタマーとは、AWS クラウド無料利用枠を含むAWS アカウントの先月の使用状況のあるアマゾン会員でない対象アカウントです。



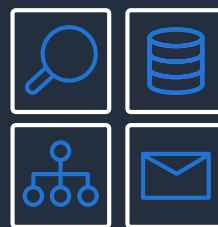


あらゆる場所で

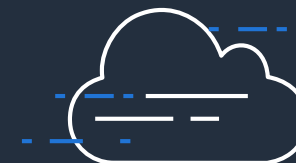
AWS が選ばれる理由



お客様の声をもとに
作られる機能やサービス



200 を超える
豊富なサービス



高い可用性を持つ
インフラストラクチャ



包括的な**セキュリティ**機能と
コンプライアンス認証の取得



世界中で**数百万を**
超えるお客様



最大規模の**コミュニティ**と
パートナーネットワーク

生成系AIとは？

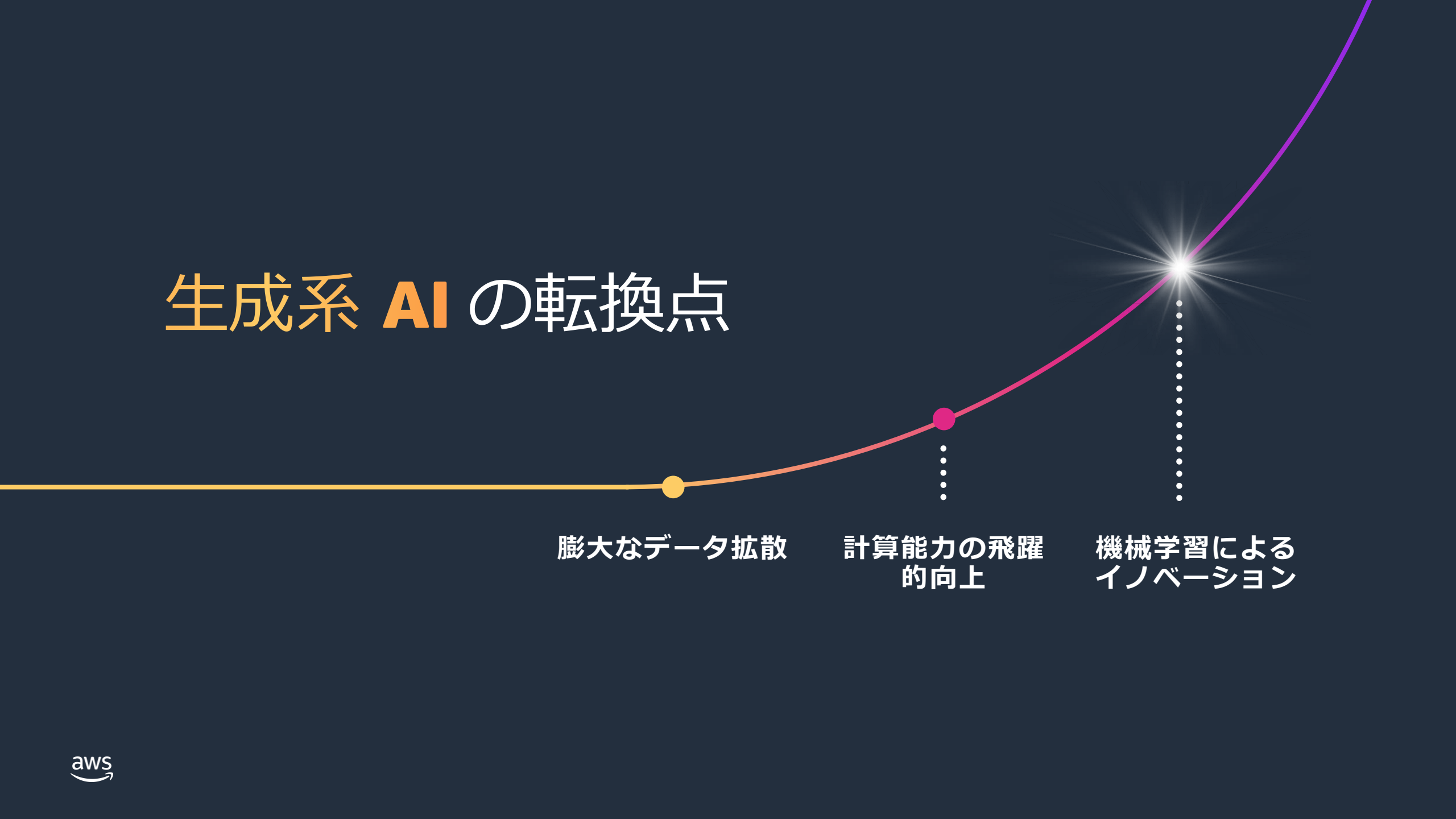


イノベーションは 産業を変革させる



生成系AI

生成系 AI の転換点



膨大なデータ拡散

計算能力の飛躍
的向上

機械学習による
イノベーション

AI

MACHINE
LEARNING

SIMPLE

INPUTS



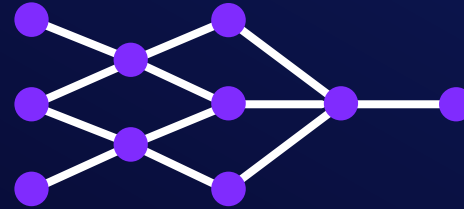
SIMPLE

OUTPUTS

DEEP
LEARNING

COMPLEX

INPUTS



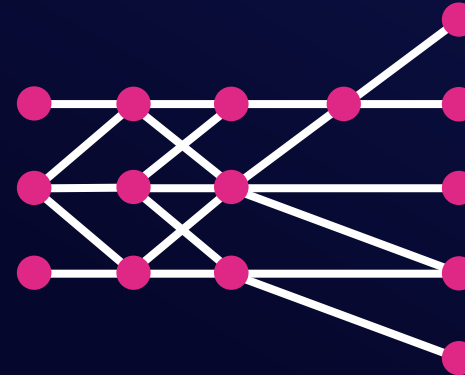
SIMPLE

OUTPUTS

FOUNDATION
MODELS

COMPLEX

INPUTS

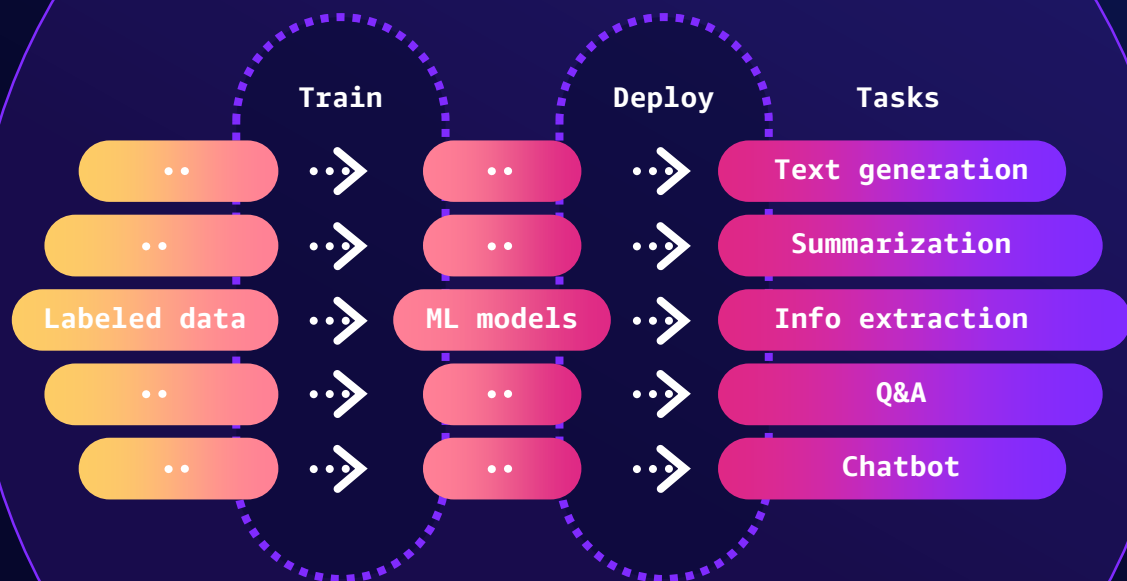


COMPLEX

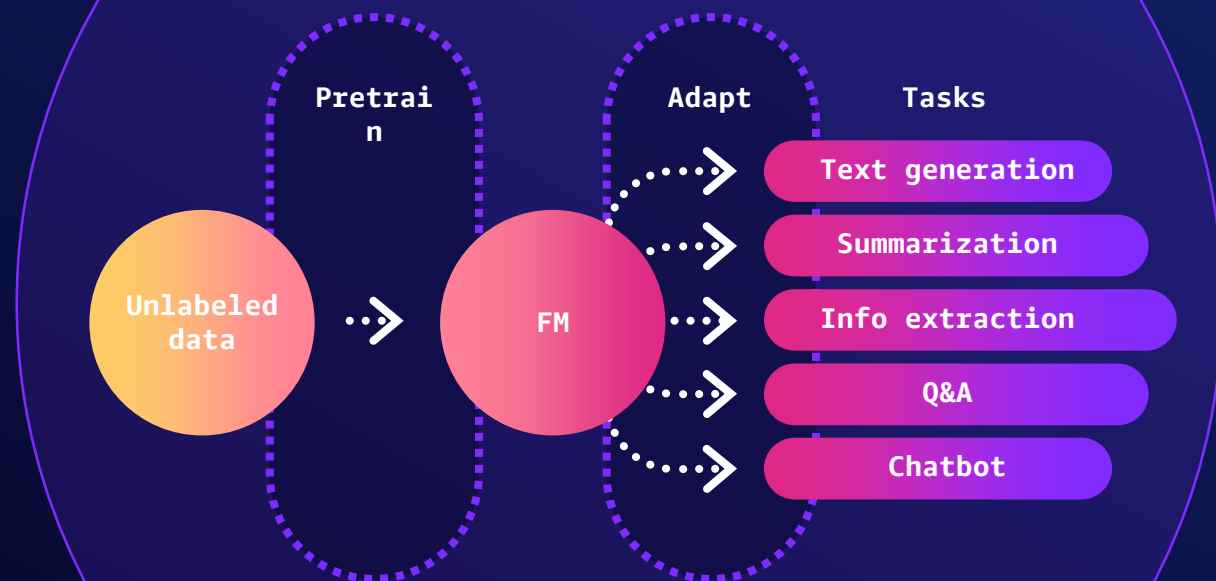
OUTPUTS



TRADITIONAL ML MODELS



FOUNDATION MODELS



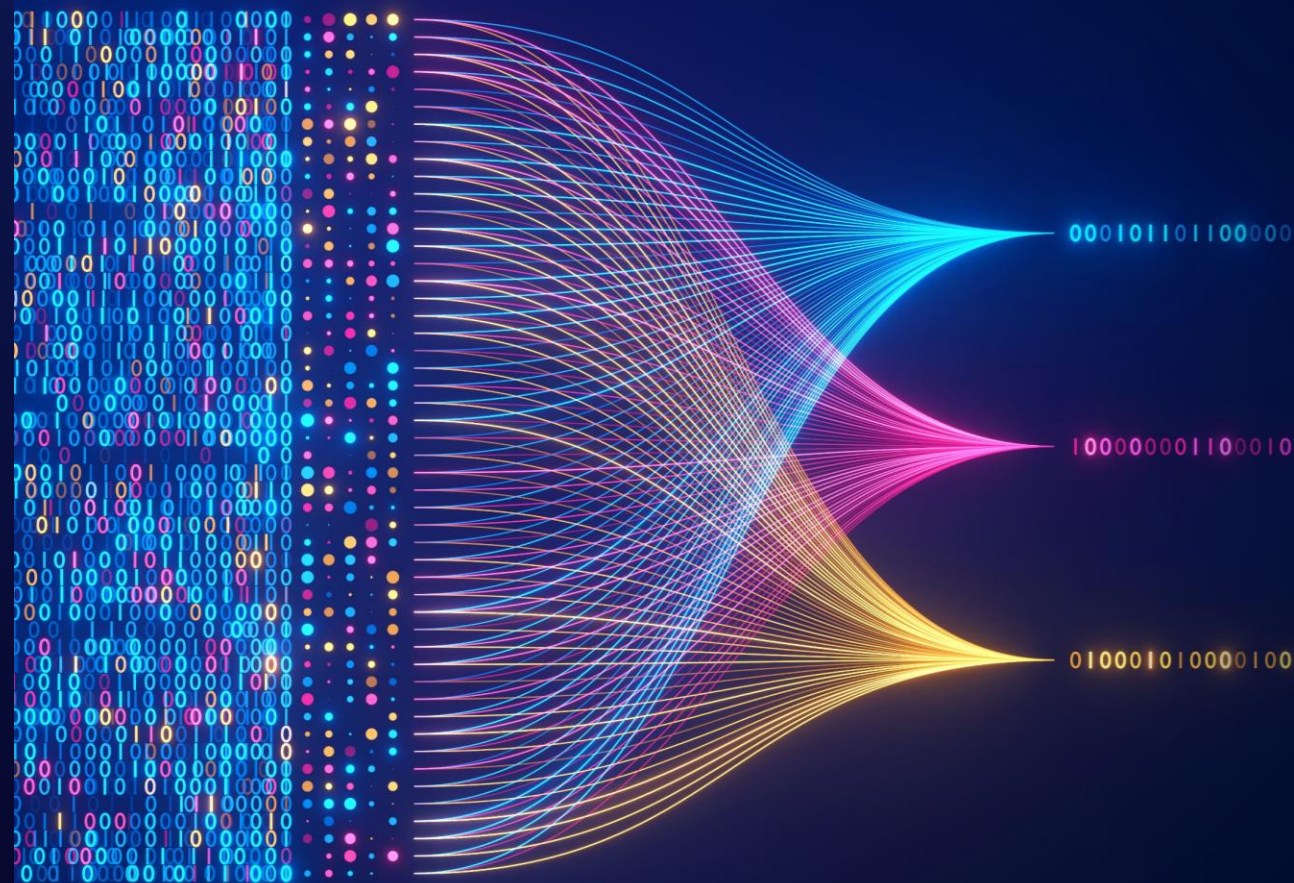
生成系AIは基盤モデルによって強化されています

膨大な量の事前トレーニング済み
非構造化データ

複雑な概念を学習できるようにする
多数のパラメーター

幅広いコンテキストに適用可能

特定の業界むけのデータを使用して
基盤モデルをカスタマイズ



様々な基盤モデルが利用されている

例

インプット



基盤モデル



アウトプット

“心臓の健康と歩行の相関性について
記事をまとめて”

Text-to-text (テキストからテキスト)
簡単な自然言語プロンプトから、
複数アプリケーションのためにテキスト生成

“健康な心臓を保つには
一日 10,000 歩のウォーキング
が有効です”

“コーヒー”

Text-to-embeddings (テキストを数値表現に)
検索や文書間の類似点検索などの
用途向けに、テキストの数値表現を生成

以下を数値で表現
“コーヒーフィルター、コーヒー豆、
コーヒーメーカー、コーヒーミル”

“宇宙飛行士が馬に乗っている写真”

Multimodal (複合的な情報を扱う)
自然言語プロンプトから、画像生成・編集



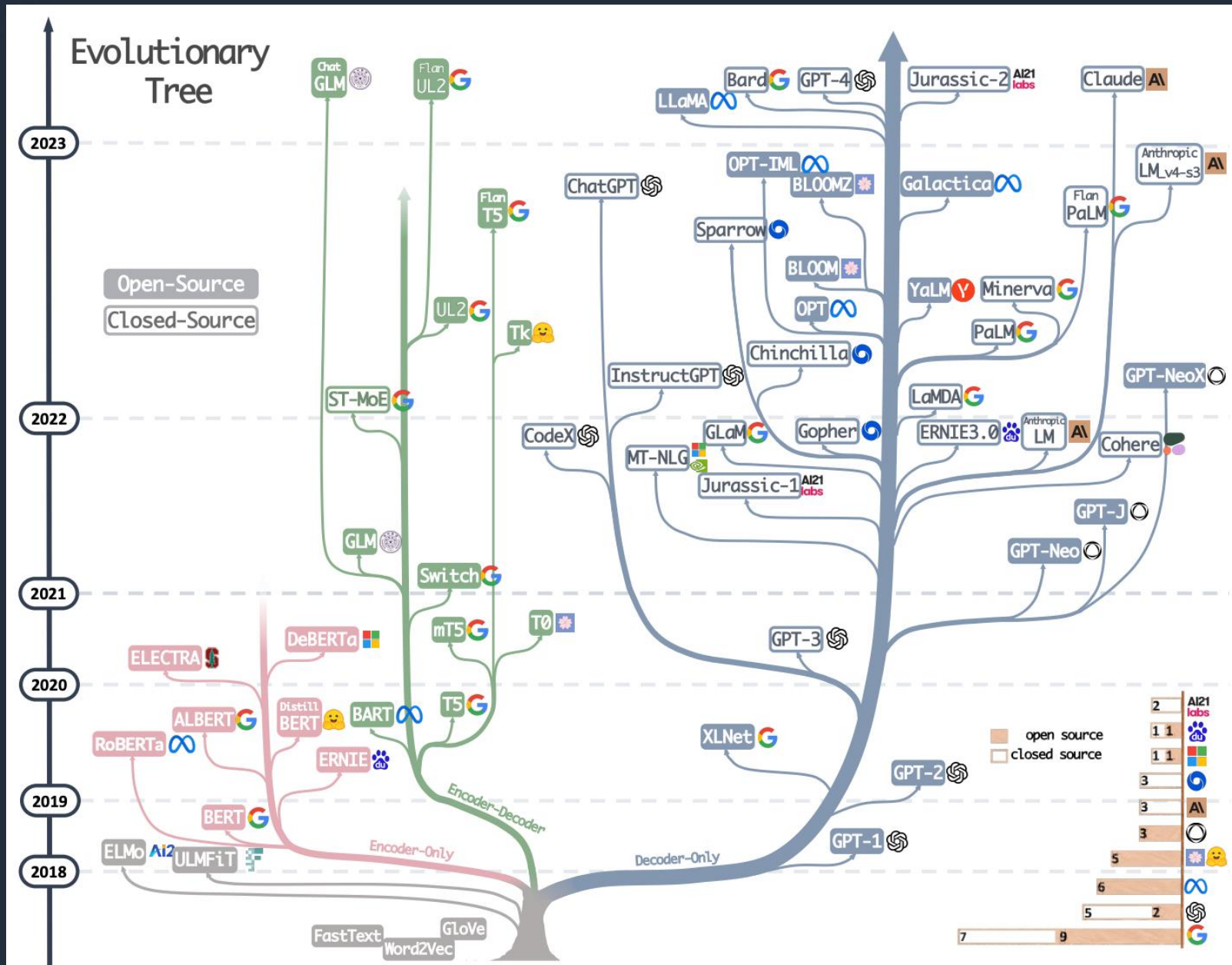
大規模言語モデルには 様々なものがある

時代の変化や用途に合わせて、
最適な大規模言語モデル(LLM)を
選択することが重要

Source: JingFeng et al. "Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond", arXiv:2304.13712 (2023)



© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.





生成系 AI で、
こんなことができる？

1

社内文書・規定を
簡単に探し、
わかりやすく要約したい



社内の情報収集を生成系 AI で効率化する

課題

- 社内のノウハウやマニュアルや外部の情報など、膨大な情報から、ユーザの関心事に関わる情報を抽出し、その要約を参照したい

要点

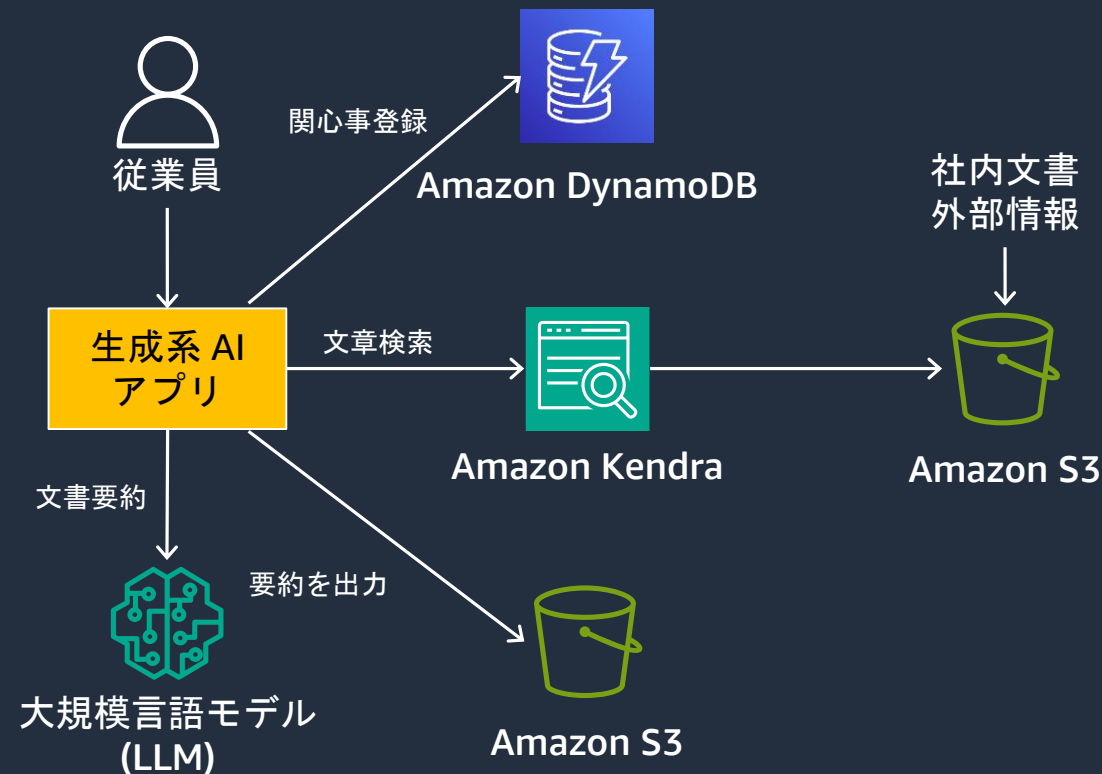
- あらかじめ関心事を登録することで社内情報や外部の最新情報に基づいた要約を自動生成。それを元に追加の情報収集に着手したり、情報源にアクセスできる
- 利用者の情報検索スキルに依存せず、自動的に要約を生成するため、誰でも・時間を掛けずに情報収集が可能

応用のアイデア

- 自動生成された要約からキーポイントを抽出し関心事として登録することで、関連情報を自動出力するよう拡張
- 情報の内容の関係性（主語-述語）から関係データを生成して関係図を生成。視覚的な情報理解を可能に

構成例

- 「従業員」が関心事に関する情報の要約の生成をリクエストする



app · Streamlit

localhost:8501

Clear Chat

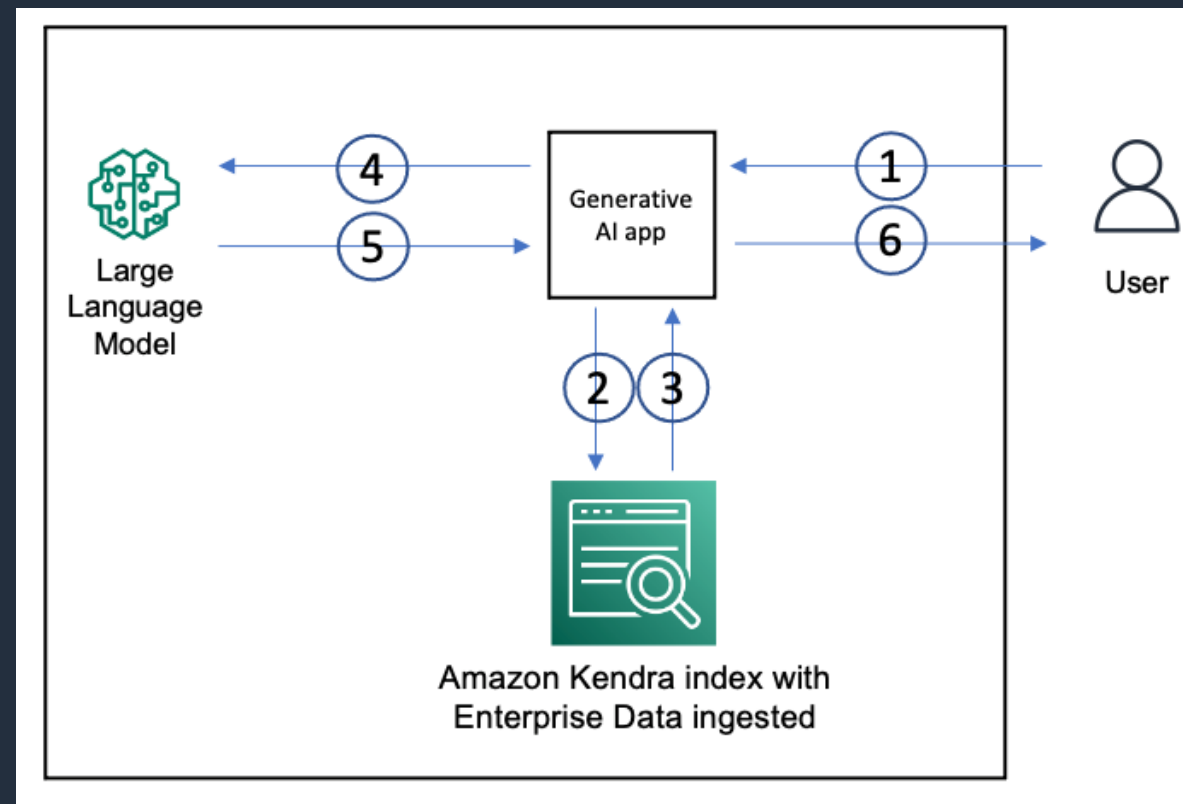
 **教えて！AI！**

AIに質問してみましょう☺

Made with Streamlit

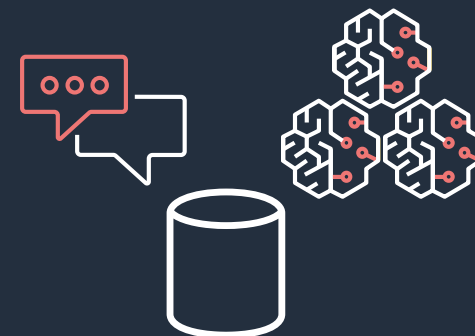
拡張検索生成 (RAG; Retrieval Augmented Generation)

- 生成系 AI アプリケーションの出力を
その組織のデータのみに制限
- 生成系 AI の課題であるハルシネーション
(事実でない回答を出力する) を抑制し、
回答の精度を向上
- Amazon Kendra の機能と LLM の機能を組み
合わせて RAG ワークフローを実装。組織内
のコンテンツに基づく会話型の
ユーザ体験を提供する最先端の
生成系 AI アプリケーションを作成できる



<https://aws.amazon.com/jp/blogs/news/quickly-build-high-accuracy-generative-ai-applications-on-enterprise-data-using-amazon-kendra-langchain-and-large-language-models/>

Amazon Kendra (エンタープライズ検索サービス)



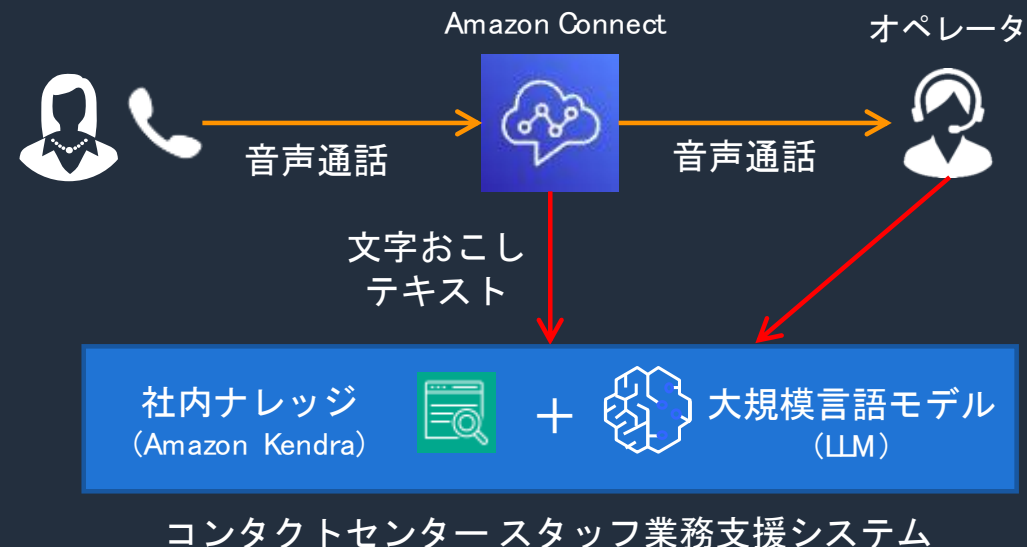
- 機械学習を利用した高精度な文書検索のためのサービス
 - 組織内の情報をあらかじめ処理し、検索リクエストに対応する回答を検索
 - 自然文の質問に対応できる
 - 関連文書をランキング形式で表示
- コネクタを利用して様々なデータソース (Amazon S3, Amazon RDS, SharePoint など)
- 学習なしで IT, 金融, 保険など、各業種の文書に対応



応用例

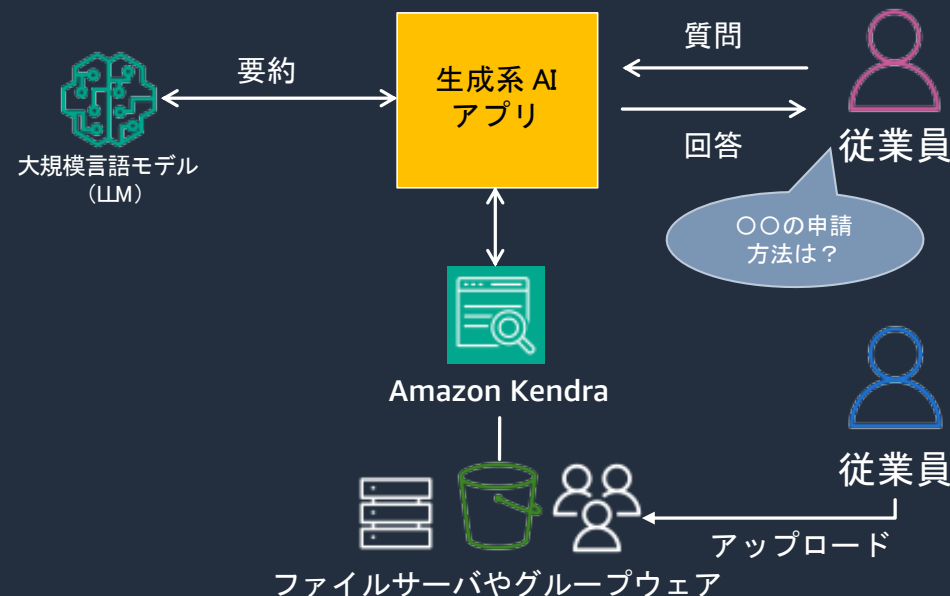
2

お客様からの
問い合わせに素早く、
正確に回答し
顧客体験を向上したい



3

抜け・漏れを減らして、
業務を
引き継ぎできるようにしたい





生成系 AI で、
こんなことができる？

4

申請書や契約書など、
文書作成を効率化したい

生成系 AI で文書作成業務を効率化する

課題

- 定型的な文書・書類作成業務を効率化することで人手不足に対処。自動化によってケアレスミスなど人的ミスの発生を抑制する

要点

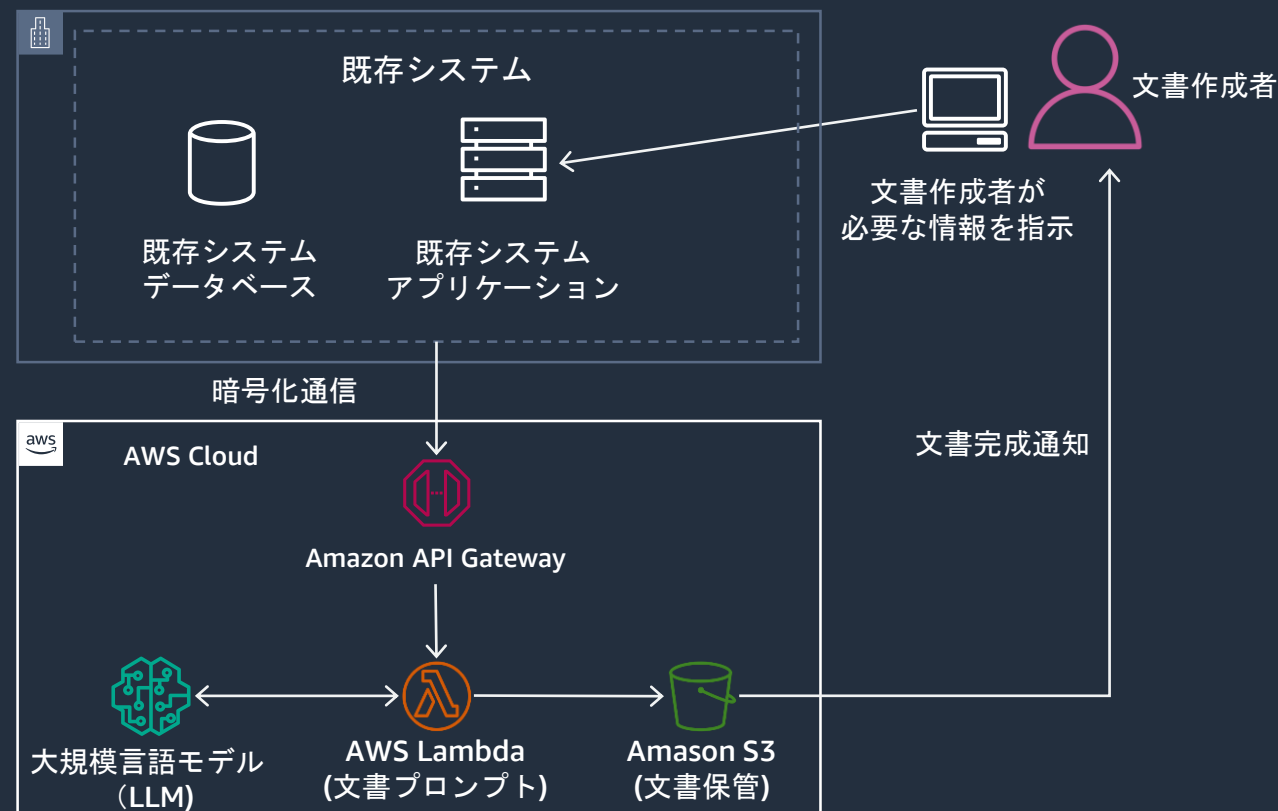
- 既存のシステムの情報AWSクラウドに連携するアプリケーションにより、書類作成時の手入力を減らしミスの発生要因を減らすことが期待できる
- 生成系 AI (大規模言語モデル)へ指示するプロンプトの工夫により複数のタイプの文章生成も可能

応用のアイデア

- 既存システムから自動的に書類生成をリクエストする仕組みにすれば、さらなる手間の削減も可能に
- 情報検索の仕組みと組み合わせることで、生成された文書の内容チェックを行うことも

構成例

- 従業員が作成すべき文書・必要な情報を指示し、既存システムが必要な情報を添えて文書化を AWS クラウド側にリクエスト





生成系 AI で、
こんなことができる？

5

患者さんの
待ち時間を減らして、
病院の診療効率を上げたい

NEW



AWS HealthScribe

HIPAA 準拠臨床用自動ノート生成サービス

IN PREVIEW NOW



臨床生産性を向上



AI を臨床現場で責任を持って
使用できるように設計



セキュリティ、プライバシー、
コンプライアンス機能搭載

テキストベース 音声書き起こし



臨床医 今日はどうしましたか？

患者 膝の後ろに固いしこりがあります。
原因はわかりませんが、治りません。

臨床医 なるほど。しこりに痛みはありますか？

患者 ええと、いいえ。痛みはありません。
でも、動くと痛む感じがします。

臨床医 なるほど。つまり、膝窩に痛みのない
硬い皮下腫瘍ができたのですね。この
しこりはどれくらい続いていますか？

患者 ああ、3ヶ月ほど前に気づきました。

臨床医 スポーツをしますか？

患者 はい、週に数回サッカーをします。
あとできれば月に2、3回は野球を
するようにしています。

臨床医 なるほど

生成系AI の洞察

CLINICAL NOTES

主な症状:膝の後ろの固いしこり

病歴:

その患者さんに今日、膝の後ろに固いしこりがあるかどうかを調べる。

患者はしこりの原因がわかりませんが、3か月前に初めて気づいて以来、しこりは消えていません。

動くとしこりを感じることができ、痛みも感じません。定期的な活動には、週に数回サッカーをしたり、月に数回野球をしたりすることが含まれます。

評価:ベイカー嚢胞

ベイカー嚢胞の処方：

膝に氷を塗り、激しい運動を避けてください。

膝に圧迫ラップを使用すると、腫れを抑え、安静時に膝を上げることができます。

それでも問題が解決しない場合は、さらなる体液の蓄積を防ぐためにコルチゾン注射を検討します。

生成系 AI の利用シーン

チャットボット
バーチャルアシスタント

エージェント支援

コンタクトセンター分析

パーソナライゼーション

**顧客体験
向上**

会話型検索

コンテンツ現地対応

テキスト、
画像動画生成

テキスト要約

コード生成

**社員生産性
向上**

文書処理

コンテンツ管理

データ作成・合成

メンテナンス支援

異常検知

**事業運営
改善**

ウェブページ用
画像生成

映像効果強化

音楽制作

高精細化
画像補正

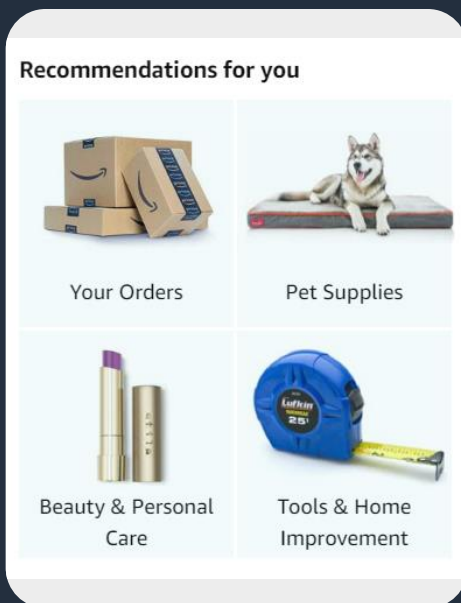
アニメーション製作

創造性

AWSの生成系AIサービス



20 年以上にわたる機械学習によるイノベーション



毎分 4,000 個
Amazon.com での取引



毎日 160 万個
のパッケージ発送

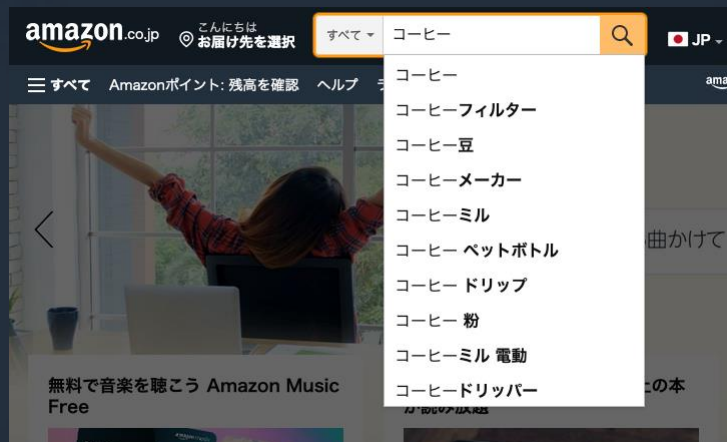


毎週 10 億件を超える
Alexa でのインタラクション



2016 年 12 月 7 日～
Prime Air ドローン配達

その取り組みは生成系 AI 技術の活用にも



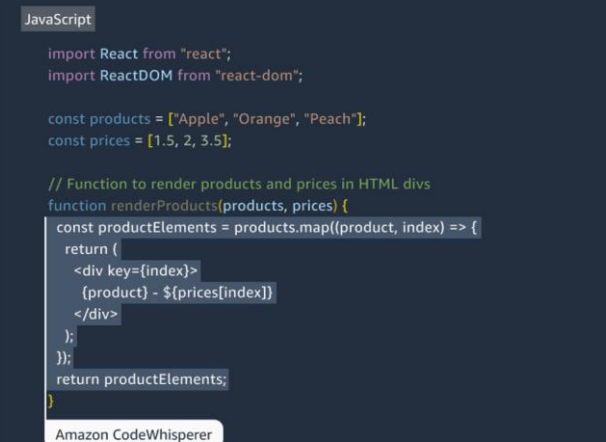
Amazon の検索機能

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

生成系 AI スタック

基盤モデルを活用したアプリケーション



Amazon Q



Amazon Q
in Amazon QuickSight



Amazon Q
in Amazon Connect



Amazon CodeWhisperer

基盤モデルを使って構築するためのツール



Amazon Bedrock

Guardrails | Agents | Customization Capabilities

トレーニングと推論のためのインフラストラクチャ



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2
Capacity Blocks



Nitro



Neuron

生成系 AI スタック

基盤モデルを活用したアプリケーション



Amazon Q



Amazon Q
in Amazon QuickSight



Amazon Q
in Amazon Connect



Amazon CodeWhisperer

基盤モデルを使って構築するためのツール



Amazon Bedrock

Guardrails | Agents | Customization Capabilities

トレーニングと推論のためのインフラストラクチャ



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2
Capacity Blocks



Nitro



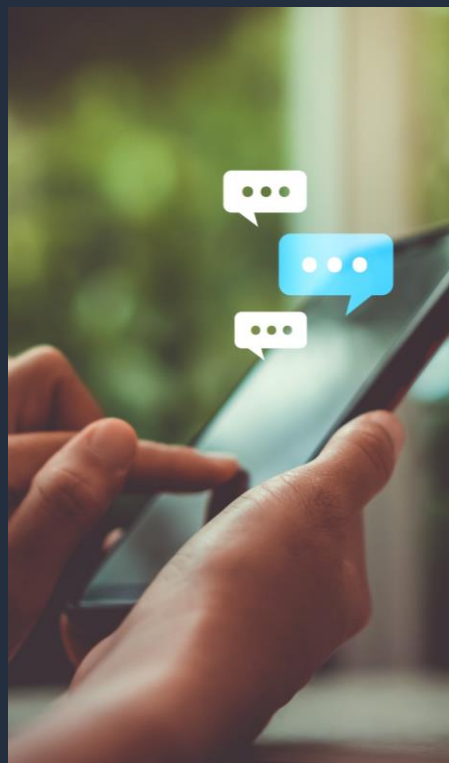
Neuron

生成系 AI を活用して多くの業務の生産性を向上できる

情報検索



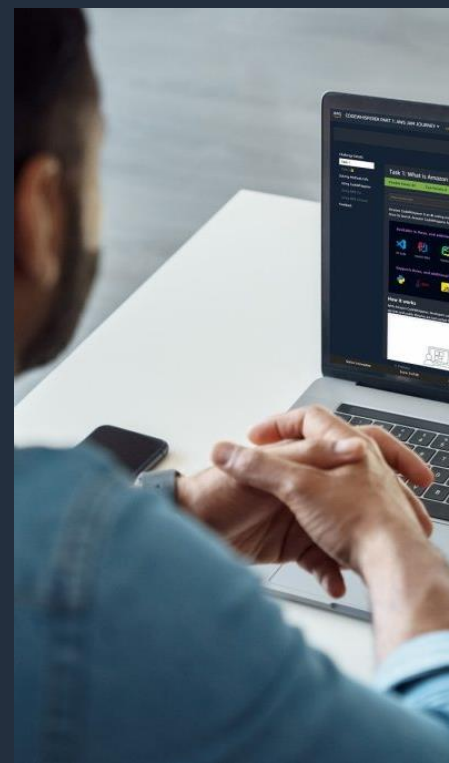
質問回答



データ分析



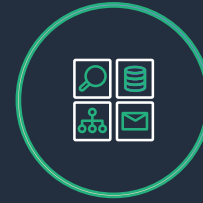
コーディング



生成系 AI を活用する際の課題



正しい回答を得られない
(ハルシネーション)



事業や顧客の情報を
反映していない



機密情報を安全に共有
する必要がある



利用する人の特性に
合っていない



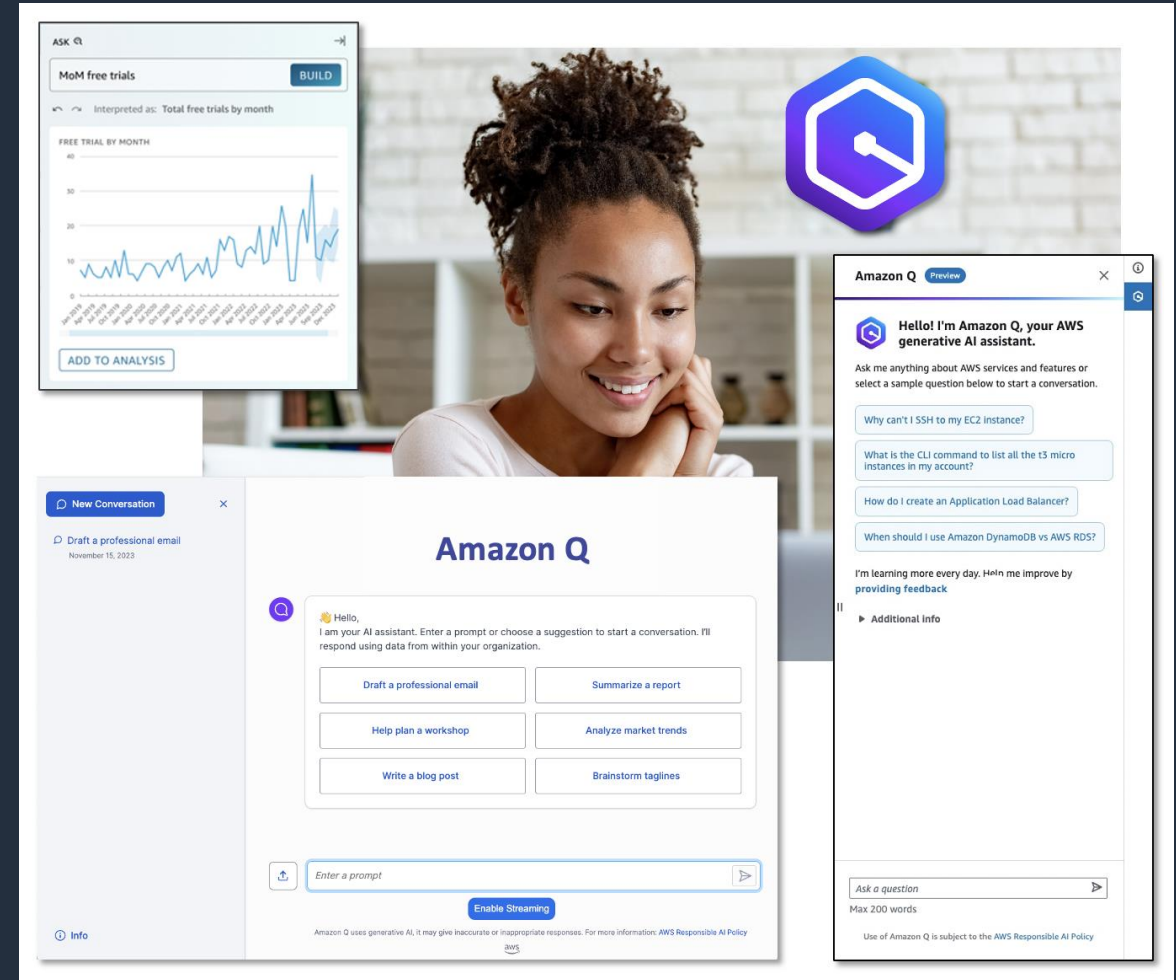
生成されるデータが
不十分または古い



保有しているデータと
連携できていない

Amazon Q

- 問題を解決したり、コンテンツを生成したり、アクションを実行するための対話を行うアシスタント
- 企業の情報、コード、システムと連携して、反応したりアクションを実行できる
- 利用者の役割や権限にもとづいた対話を提供する
- セキュアでプライベートな環境
- 2023/12 時点でプレビューで提供。対応言語は英語。



Amazon Q が対応する専門領域

Your
business



Building
on AWS



Amazon
QuickSight



Amazon
Connect

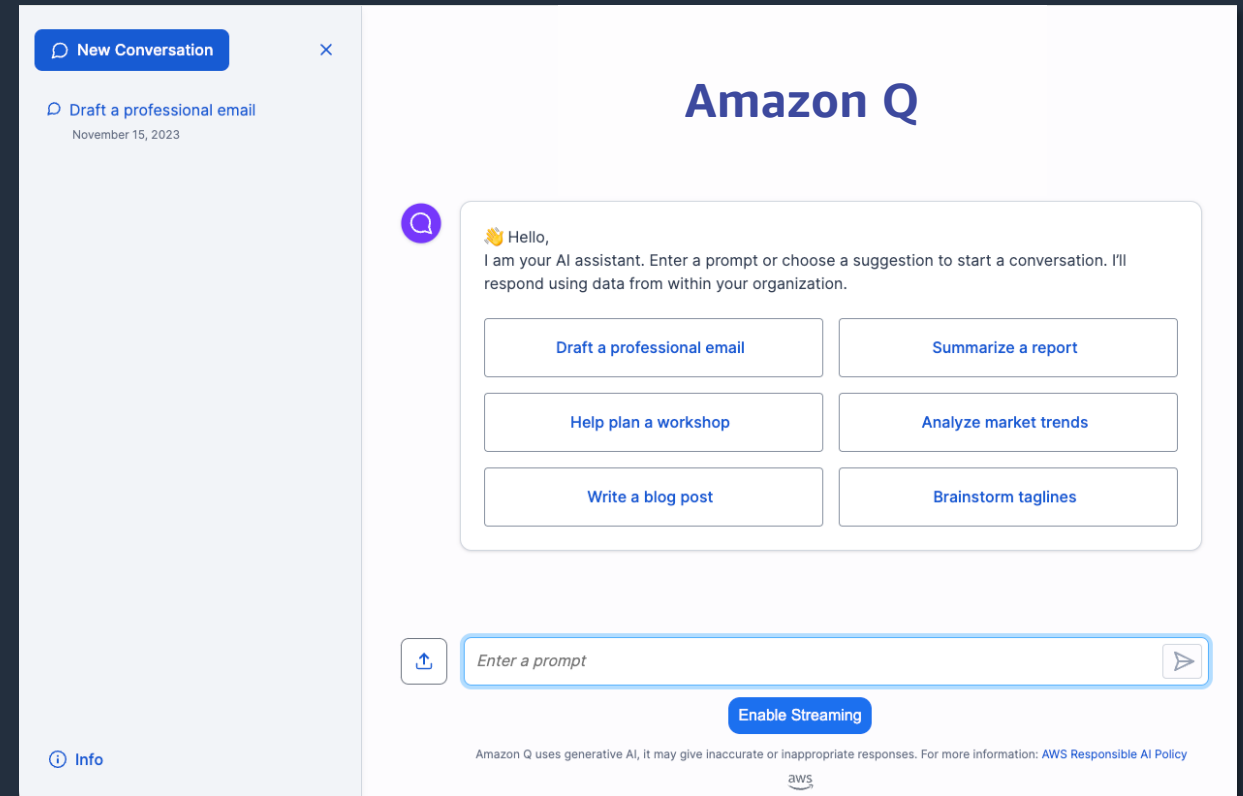


AWS
Supply Chain



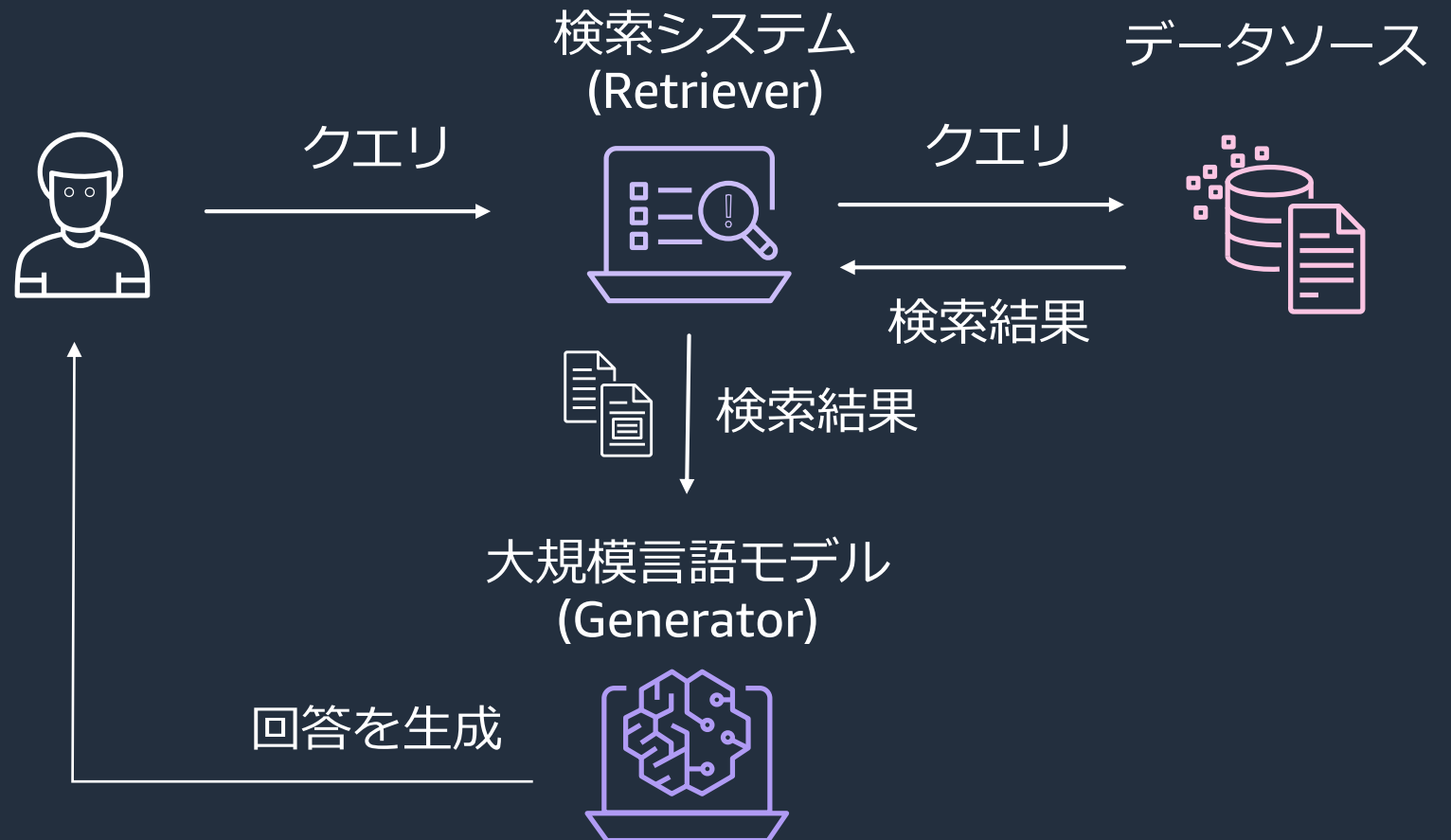
Amazon Q is your business expert (プレビュー)

- 質問に対する回答をすぐに正確に提供し、プライベートな文書レポジトリに安全に接続
- 回答の正確性を事実に沿って確認するために、回答の引用元を提供する
- ドキュメントに対する既存のアクセス制御を継承する
- 40 以上の企業アプリケーション、レポジトリに対するコネクタを提供
- レスポンスをカスタマイズしたり制御する機能 (ガードレール) を提供



Amazon Q のコンセプト: RAG (検索拡張生成)

- 大規模言語モデル (LLM) のみで回答を得るのではなく、既存のデータソースを活用
- 社内情報などの非公開なデータソースから、社内業務に合った回答を生成でき、正確さを向上



🔍 New Conversation



🔍 tell me about Example Corp. what do they do?

November 23, 2023

🔍 Tell me about Example Corp. What does the company do?

November 23, 2023

📘 Info

Amazon Q

Your AI assistant for work



👋 Hello, I am your AI assistant. Enter a prompt or choose a suggestion to start a conversation. I'll respond using data from within your organization.

架空の Example Corp に関する質問



tell me about Example Corp. what do they do?



Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)



🔍 New Conversation



🔍 tell me about Example Corp. what do they do?

November 23, 2023

🔍 tell me about Example Corp. what do they do?

November 23, 2023

🔍 Tell me about Example Corp. What does the company do?

November 23, 2023

対話の履歴を保存

📘 Info



tell me about Example Corp. what do they do?



Example Corp is a leading global fashion retailer known for its affordable and on-trend apparel and accessories. Founded in 1975 in New York City as a small women's boutique, Example Corp has grown into a multinational company with over 1,500 stores across 50 countries. At its core, Example Corp focuses on bringing accessible fashion styles to consumers. It offers apparel, footwear and accessories for women, men and kids. Example Corp has sophisticated supply chain operations and digital commerce capabilities. It markets through traditional and social media. Today, Example Corp retains its founding values while strategically expanding its global presence and sustainability programs. [1](#) [2](#) [3](#)

Sources ▾



参照元の情報を調べることができる



I

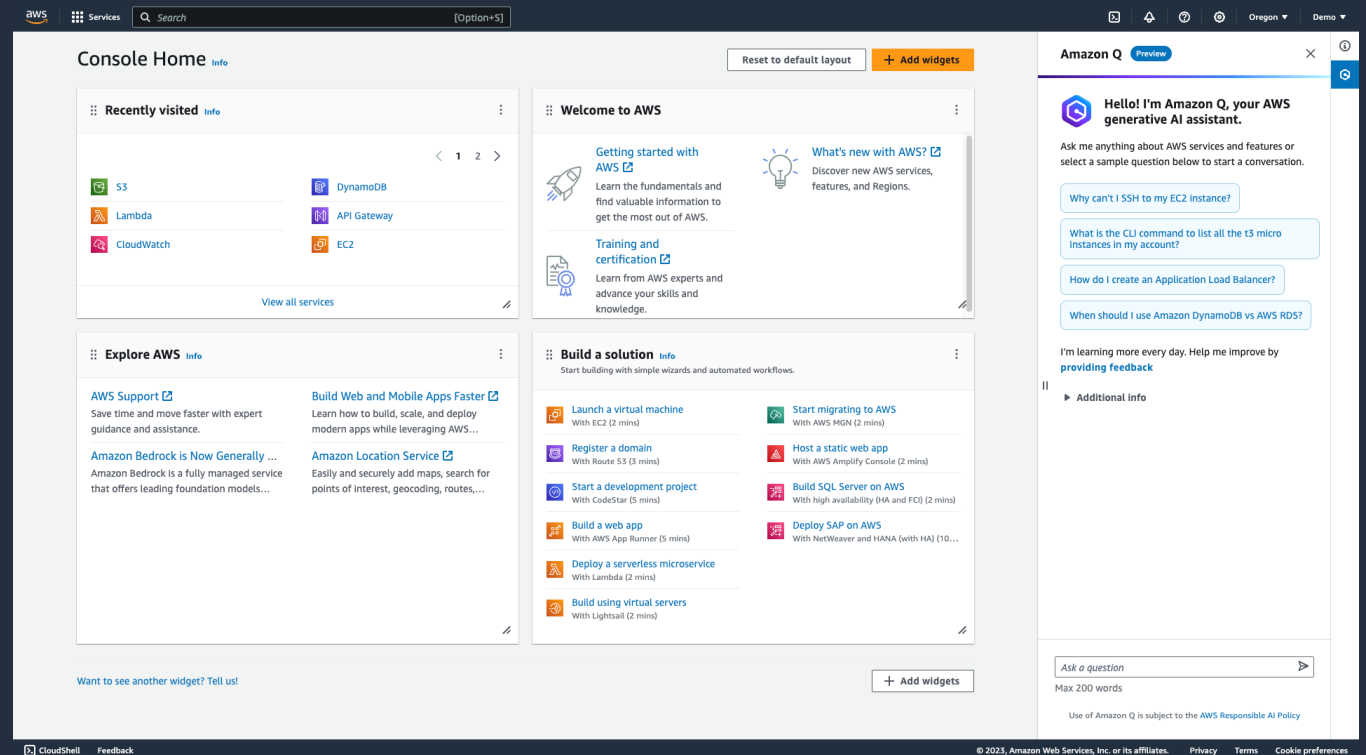


Amazon Q - Preview uses generative AI and it may give inaccurate or inappropriate responses. For more information: [AWS Responsible AI Policy](#)



Amazon Q is your AWS expert (プレビュー)

- 17年の AWS に関する知識にもとづいて構築されている
- AWS で作業するあらゆる場面で支援
- 新しい AWS の機能を調べたり、不慣れた技術を学んだり、ソリューションを構築するユーザと会話
- トラブルシューティング、新機能の構築などをサポート



Console Home [Info](#)

Reset to default layout

+ Add widgets

⋮ Recently visited [Info](#)

< 1 2 >

 API Gateway

 IAM

 CloudWatch

 Lambda

 S3

 EC2

[View all services](#)

⋮ Welcome to AWS



Getting started with AWS [↗](#)

Learn the fundamentals and find valuable information to get the most out of AWS.



What's new with AWS? [↗](#)

Discover new AWS services, features, and Regions.



Training and certification [↗](#)

Learn from AWS experts and advance your skills

⋮ AWS Health [Info](#)

Open issues

0

Past 7 days

Scheduled changes

0

Upcoming and past 7 days

⋮ Explore AWS [Info](#)

[AWS Support \[↗\]\(#\)](#)

Save time and move faster with expert guidance and assistance.

[Amazon Bedrock is Now Gener...](#)

Amazon Bedrock is a fully managed service that offers leading foundati...

[Build Web and Mobile Apps Fas...](#)

Learn how to build, scale, and deploy modern apps while leveraging AWS

[Build Apps Faster with GraphQL...](#)

AWS AppSync uses GraphQL APIs to query data from multiple data

 CloudShell

Feedback

© 2023, Amazon Web Services, Inc. or its affiliates.

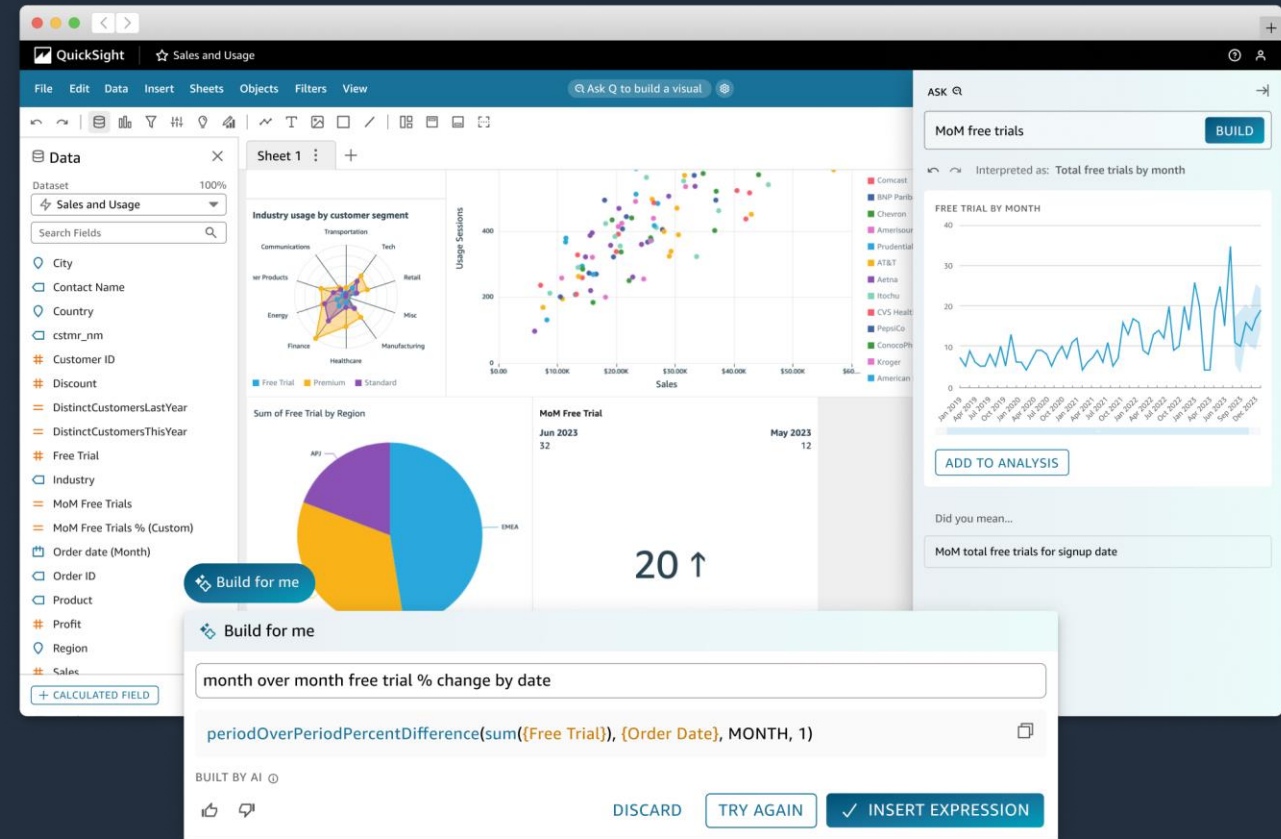
Privacy

Terms

Cookie preferences

Amazon Q in QuickSight (プレビュー)

- ビジネスアナリストやビジネスユーザが、すぐに構築できて、見た目をよくしたり、洞察の要約を作成したり、データに関する質問に回答できるようにする
- 特定の文法・作法を調べたり、学んだりすることなく、簡単な計算だけで分析を加速
- シンプルな UI で、見た目にわかりやすい表示で、洞察を解釈したりシェアできる





New experience

Data

Sheet 1

Dataset

SPICE SaaS-Sales_2... 100%

Fields list

Search fields

- City
- Contact Name
- Country
- Customer ID
- Customer Name
- Discount
- Free Trial
- Industry
- Order Date
- Order ID
- Product
- Profit
- Region
- Sales
- Segment
- Subregion
- Usage Sessions

+ CALCULATED FIELD



Total Sales

\$2,487,148

Total Profit

\$304,396

Total Usage Sessions

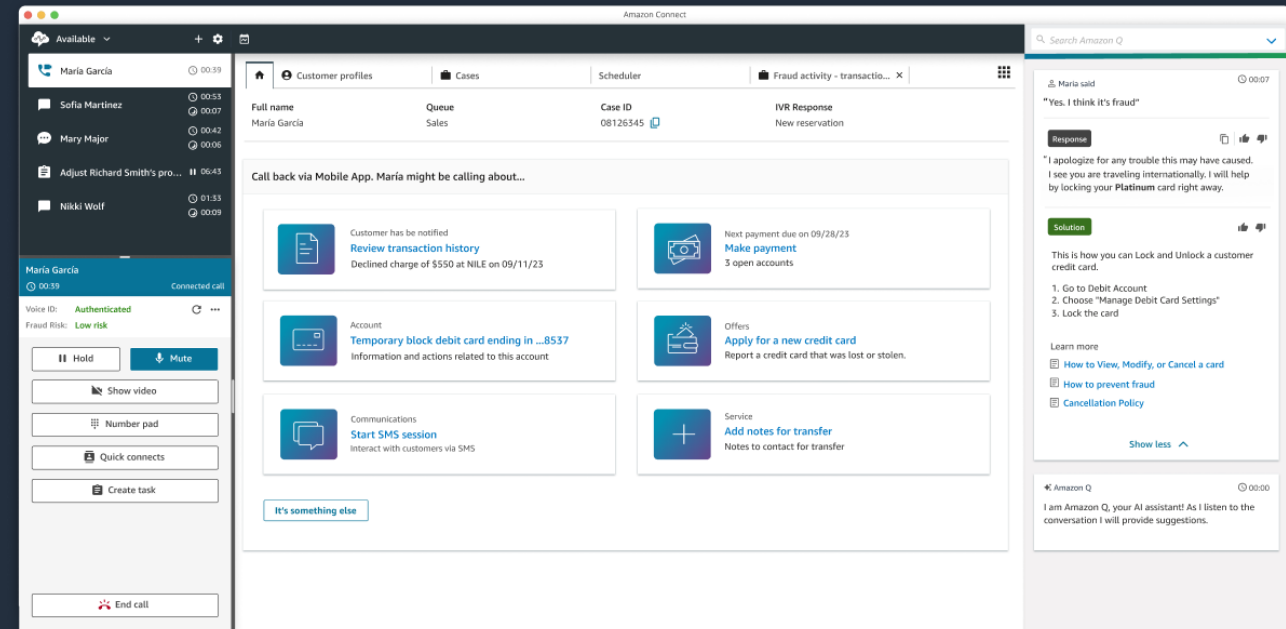
41,077

Total Free Trials

411

Amazon Q in Connect

- コンタクトセンターのオペレータに対して、顧客の会話にもとづくコンテキストから、リアルタイムにレスポンスを返したり、アクションを提案する
- オペレータは、企業の情報を横断的に検索し、とるべきアクションといった、必要な情報を知ることができる
- ソースとなる文書はオペレータに表示され、必要に応じて詳細な情報にアクセスできる

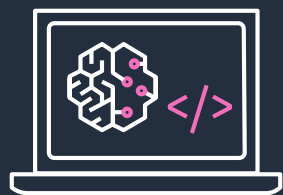


生成系 AI によるAWS の AI サービスの強化

- Amazon Transcribe: 音声からの文字の書き起こし
 - 基盤モデルによって100言語以上をサポートし、精度を改善 (バッチモード)
 - Call Analytics 機能における通話要約機能 (プレビュー)
- Amazon Lex: 音声・テキストによる会話型インタフェースの構築
 - 基盤モデルによる RAG を利用して企業のデータとの連携が可能に (プレビュー、英語対応)
 - 組み込みのスロット (都市名、日付など) の認識精度向上 (英語対応)
- Amazon Personalize: リコメンデーション
 - 類似商品のリコmendを行う際に、個々の類似商品の説明文などから、類似商品全体を表すテーマ文を生成。「類似商品のおすすめ」よりも、商品に合った文を生成できる。

Amazon CodeWhisperer（一般公開）

個人開発者は無料で使用可能



コードの候補を
リアルタイムで生成



コードをスキャンして、
発見の困難な脆弱性が
ないか調べる



オープンソースのトレーニ
ングデータに似たコードに
フラグをたて、フィルタリ
ングする



生成系 AI スタック

基盤モデルを活用したアプリケーション



Amazon Q



Amazon Q
in Amazon QuickSight



Amazon Q
in Amazon Connect



Amazon CodeWhisperer

基盤モデルを使って構築するためのツール



Amazon Bedrock

Guardrails | Agents | Customization Capabilities

トレーニングと推論のためのインフラストラクチャ



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2
Capacity Blocks



Nitro

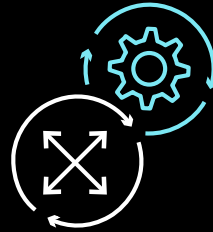


Neuron

生成系 AI を利用する上で重要な点



基盤モデルの
選択肢



アプリケーション
とのシームレスな
統合



データの
セキュリティと
プライバシー



Amazon Bedrock

基盤モデルによる 生成系 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

基盤モデルのカスタマイズ

検索拡張生成 (RAG)

複雑なタスクを実行可能なエージェント

セキュリティ・プライバシー・安全性

Amazon Bedrock は主要な基盤モデルをサポート



Amazon Titan

要約、画像とテキストの生成と
検索、Q&A

AI21labs

Jurassic-2

文脈に応じた回答、要約、
言い換え

ANTHROPIC

Claude 2.1 & Claude Instant

要約、複雑な推論、
ライティング、コーディング



Command & Embed

テキスト生成、検索、分類



Llama 2

対話のユースケースと
言語タスク

stability.ai

Stable Diffusion XL 1.0

高品質の画像とアート



Amazon Titan Models

Amazon による高性能な基盤モデル

テキストの生成と埋め込みに加えて
画像生成 (プレビュー) と画像・テキストの
同時埋め込みを新たにサポート

ラベル付きのデータによるファイン
チューニングと、ラベルなしデータ
による継続事前学習 (プレビュー)

知的財産権に関わる損害賠償請求への補償

新しく利用可能になった Third-party 基盤モデル

ANTHROPIC

Claude 2.1

- 20万トークンのサポート
- 50% 少ない
ハルシネーション
- Q&A タスクにおける誤りを30%低減

∞ Meta

Llama 2 70B

- 13B に加えて 70B のモデルを利用可能に
- Llama2 はチャットのユースケースにおいて性能のテストを実施
- ファインチューン可能

stability.ai

Stable Diffusion XL 1.0

- 品質の高い、芸術的な画像、リアルな画像の生成が可能
- 特に色使い、コントラスト、ライティングなどにうまくチューニングされている

自分のデータを使って基盤モデルをカスタマイズ

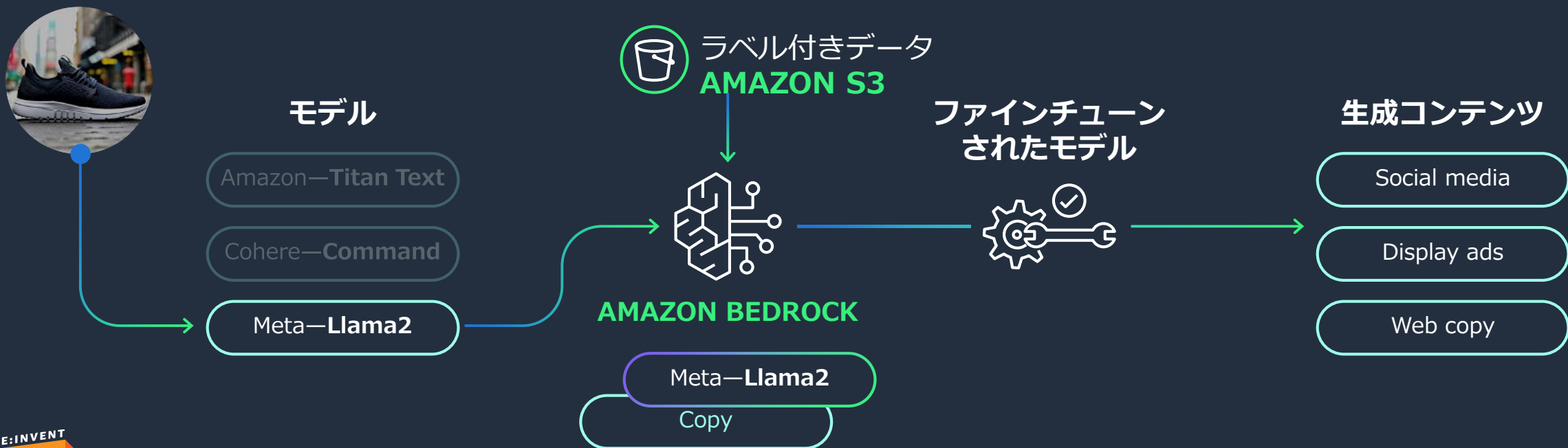
ファインチューニングおよび継続事前学習

基盤モデルを
カスタマイズすることで
より良いユーザエクスペ
リエンスを提供

ラベル付きデータを使
用して、Llama 2、
Command、Titan の基
盤モデルをカスタマイズ

継続事前学習により、
Titan Text の基盤モ
デルをお客様のドメ
インに適合 (Preview)

Amazon Bedrockへの
入力・出力は、
元の基盤モデルの改善
に使用されない



Amazon Bedrock でのモデル評価機能

基盤モデルを評価して、ユースケースに最適なモデルを選択する

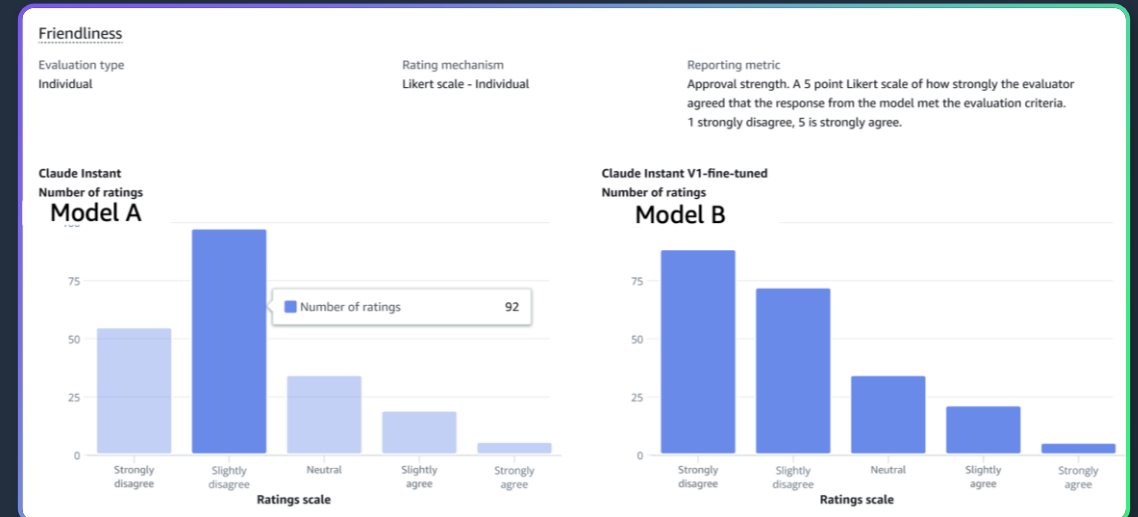
Preview

自動評価または人間による評価を実施できる

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
(リッカート尺度による5段階評価など)



自動評価レポート

Text summarization evaluation summary (3)

The results for text summarization consist of accuracy, toxicity, and robustness, which indicate the quality of the summaries generated by the model. [Learn more.](#)

Accuracy		Toxicity		Robustness	
Dataset	Value	Dataset	Value	Dataset	Value
CNN/DailyMail	.6	S3 URI	.5	CNN/DailyMail	.4
S3 URI 3	.4			S3 URI 2	.6

Knowledge bases for Amazon Bedrock

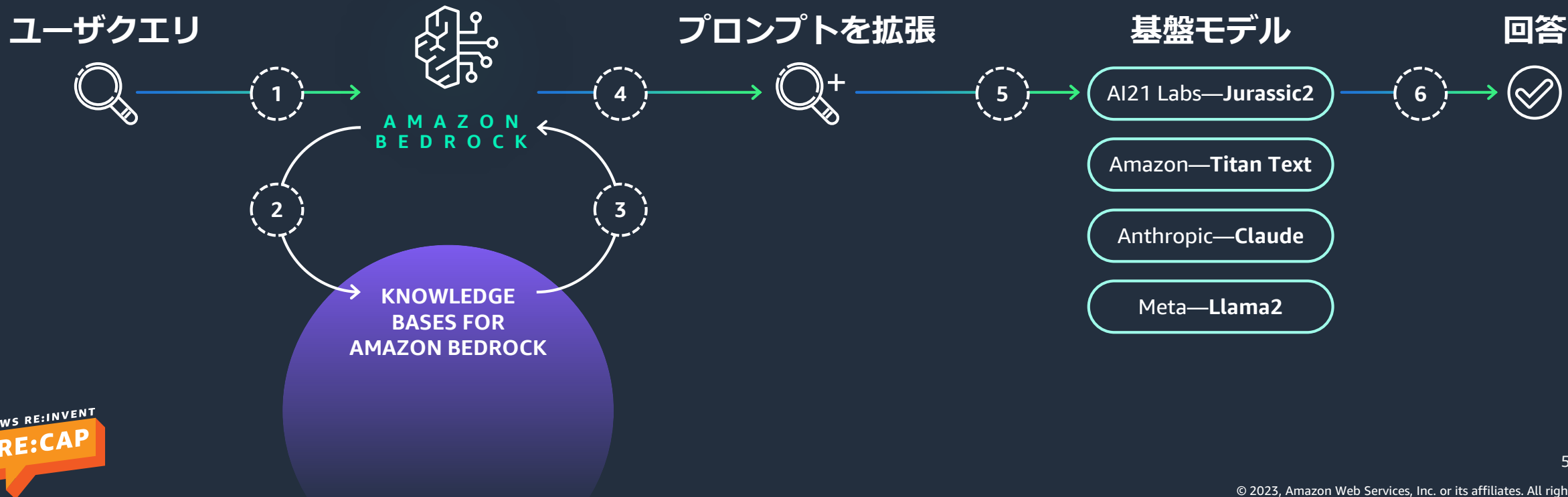
検索拡張生成（RAG）のネイティブサポート

基盤モデルをデータソースに安全に接続し、RAGにより適切な回答を提供

取り込み、検索、生成を含むフルマネージド RAG ワークフロー

対話を継続する際に、コンテキスト情報を組み込み・管理する

検索結果に引用元を自動的に付与し、透明性を改善する



Guardrails for Amazon Bedrock

アプリケーション要件と責任ある AI ポリシーに合わせた保護を実装する

Preview

ファインチューニングモデルやエージェントを含め、すべての基盤モデル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

個人を特定できる情報を検出して隠す
(近日公開予定)

The screenshot displays the Amazon Bedrock Guardrails console interface. The main configuration area is titled 'Working draft: antje-banking-assistant'. It includes several sections:

- Denied topics (1):** A table with columns 'Name' and 'Instructions'. The 'Name' column contains 'Investment advice', which is highlighted with a red box. The 'Instructions' column contains 'Investment advice refers to guidance or recommendations provided by a financial professional, adv...'. An 'Edit' button is visible.
- Content moderation: filter strengths:** A table with columns 'Prompt filters' and 'Response filters'. Both columns show 'ON' for 'Toxicity filter strength', 'Insults filter strength', 'Sexual filter strength', and 'Violence filter strength'. The 'Response filters' column also shows 'High' for 'Toxicity filter strength', 'Insults filter strength', 'Sexual filter strength', and 'Violence filter strength'. An 'Edit' button is visible.
- Default responses:** A table with columns 'Blocked prompts' and 'Blocked responses'. Both columns show 'Sorry, I can't comment on that.'.

On the right side, there is a 'Test' section. It shows a 'Prompt' input field with the text 'Should I open a credit card account?' highlighted by a red box. Below the prompt, the 'Model response' and 'Final response' are displayed. The 'Guardrail check' section shows a green checkmark and the text 'Passed View trace >'. A red arrow points to the 'View trace >' link. Below the 'Guardrail check' section is a 'Run' button.

生成系 AI スタック

基盤モデルを活用したアプリケーション



Amazon Q



Amazon Q
in Amazon QuickSight



Amazon Q
in Amazon Connect



Amazon CodeWhisperer

基盤モデルを使って構築するためのツール



Amazon Bedrock

Guardrails | Agents | Customization Capabilities

トレーニングと推論のためのインフラストラクチャ



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2
Capacity Blocks



Nitro



Neuron

NVIDIA、及び、Anthropicとのパートナーシップ

NVIDIAとは長年のパートナーシップ。

ANTHROPICとは、初期からの提携であり、CLAUDE2.1ではTRAINIUM, INFERENCEIAを利用

• NVIDIA

- 新しいパートナーシップを発表。AWSはGPUを初めてクラウドで提供した会社。200万GPU, 3000ZettaFlopを展開している。
- 全く新しいGPUファミリとしてH200、GH200、L40S、L4 GPUを提供予定。Nitroを活用して、仮想的なGPUのインスタンスを構築し、ultra clusterを構成できるようになる。
- NVIDIA DGX Cloud - NVIDIA GH200 NVL32 を初めて搭載

• Anthropic

- 数年前にAmazonとの提携を発表。AWSは、初期からの主要な提携先であり、Trainium, Inferentiaを活用して、4倍早いスピードを活用している。その結果、ユニークな機能を提供できている。
- 最も多いユースケースは、バイオメディカルの領域。創薬に関して、モデルを微調整して、リーガルサーチに活用できる。似たような業界でも活用できると思う。
- Claude2.1は、220万のトークンをサポートして、複数のドキュメントをサポート。ハルシネーションも減らしている。エンタープライズ向けの機能も含めている。より安全で、責任のあるGenAIを提供することが重要である。
- 他の競合のモデルに比べて、10倍安全である。コンプライアンスを考慮して、エコシステムで、テストや評価、測定をできるようにしたり、規制に対しても、責任を持って対応している。

AWS Trainium 2とEC2 Trn2インスタンスを発表

- AWS Trainium 2チップを発表
 - AWS Trainiumと比較してトレーニング処理が最大4倍高速、メモリ容量が3倍、エネルギー消費効率が2倍に
- AWS EFAによるEC2 Ultra Clusterで、最大100,000のTrainium 2で処理を実行可能
 - 最大65EFLOPSを発揮し、3,000億パラメータのLLMのトレーニングを数ヶ月から数週間の単位に短縮する
- 詳細は今後発表予定



Amazon SageMaker JumpStart

既存の基盤モデルにアクセスして、今日から開発・運用スタート



Products / Machine Learning / Amazon SageMaker JumpStart

Getting started with Amazon SageMaker JumpStart

Amazon SageMaker JumpStart is a machine learning (ML) hub that can help you accelerate your ML journey. Explore how you can get started with built-in algorithms with pretrained models from model hubs, pretrained foundation models, and prebuilt solutions to solve common use cases. To get started, see documentation or example notebooks that you can quickly execute.

[Reset Filters](#)

Search: foundation models

Sort By: Popularity

Product Type	Text Tasks	Vision Tasks
FOUNDATION MODEL	☐ End-to-end Solution ☐ Text Classification ☐ Text Embedding ☐ Text Generation ☐ Text Summarization ☐ Named Entity Recognition ☐ Question Answering ☐ Zero-Shot Classification	☐ End-to-end Solution ☐ Image Classification ☐ Image Embedding ☐ Instance Segmentation

Text Generation

Proprietary Models

Various Providers

Models from AI21 Labs, Cohere, and LightOn in preview. Sign-up for preview with JumpStart in us-east-1 or eu-west-1 SageMaker Console.

Deploy Only

Text to Image

Stability AI

Model ID: model-txt2img-stabilityai-stable-diffusion-v2. This is a text-to-image model from Stability AI and downloaded from HuggingFace. It takes a textual description as input and generates an image.

Fine-tunable

Text Generation

Alexa

Model ID: pytorch-textgeneration1-alexa20b. AlexaTM 20B is a multitask, multilingual, large-scale sequence-to-sequence (seq2seq) model, trained on a mixture of Common Crawl.

Deploy Only

Text Generation

Bloom 1b7

Huggingface

Model ID: huggingface-textgeneration-bloom-1b7. This is a Text Generation model built upon a Transformer model from Hugging Face. It takes a text string as input and predicts next words in the sequence. This model has BigScience Responsible AI License v1.0. Please read the [terms] (https://huggingface.co/spaces/bloom).

Deploy Only

Build with the broadest selection of foundation models

AVAILABLE IN AMAZON SAGEMAKER JUMPSTART

∞ Meta AI

Models

Llama 2 7B, 13B, 70B

Tasks

Question answering
Chat
Summarization
Paraphrasing
Sentiment analysis
Text generation

AI21 Labs

Models

Jurassic-2 Ultra, Mid
Contextual answers
Summarize
Paraphrase
Grammatical error
correction

Tasks

Text generation
Long-form generation
Summarization
Paraphrasing
Chat
Information extraction
Question answering
Classification

co:here

Models

Cohere
Command XL

Tasks

Text generation
Information extraction
Question answering
Summarization



Models

Falcon-7B, 40B
Open LLaMA
RedPajama
MPT-7B, Dolly
BloomZ 176B
Flan T-5 models (8 variants)
DistilGPT2
GPT NeoXT
Bloom models
(3 variants)

Tasks

Machine translation
Question answering
Summarization
Annotation
Data generation

Features

Fine-tuning on FLAN T5 models,
GPT-6B, Falcon-7B

stability.ai

Models

Stable Diffusion XL
2.1 base
Upscaling
Inpainting

Tasks

Generate photo-realistic
images from text input
Improve quality of generated
images

Features

Fine-tuning on Stable
Diffusion 2.1 base model

Light

Models

Lyra-Fr
10B, Mini

Tasks

Text generation
Keyword extraction
Information extraction
Question answering
Summarization
Sentiment analysis
Classification

alexa

Models

AlexaTM 20B

Tasks

Machine translation
Question answering
Summarization
Annotation
Data generation

まとめ

- 生成系 AI スタックを新たに発表し、生成系 AI のためのアプリケーションからインフラまでを幅広くカバーするサービス群を提供
- アプリケーションのレイヤーでは Amazon Q を発表し、情報検索、質問回答、データ分析、コーディングなど様々な業務への生成系 AI の適用をサポート
- Amazon Bedrock で、基盤モデルに簡単にアクセスし、生成系 AI アプリケーションの開発を容易に
- Amazon SageMaker や Trainium など、生成系 AI のためのインフラによって、基盤モデルの開発や運用を効率化