

# AI-3016 コースの内容について

- ・本コースでは、Azure AI Foundry（旧 Azure AI Studio）の基本的な機能や使い方についてご説明します。
- ・※2024/11/19 マイクロソフトのイベント「Ignite 2024」にて、Azure AI StudioはAzure AI Foundryにリブランディング（名称変更）されました。
- ・※ 「foundry」（ファウンドリ） = 鑄物（鋳造）工場、鑄物類、半導体製造工場

# コースの構成

モジュール1	Azure AI Foundry の概要
モジュール2	Azure AI Foundry でモデル カタログのモデルを調べてデプロイする
モジュール3	Azure AI Foundry でのプロンプト フローの概要
モジュール4	Azure AI Foundry を使用して独自のデータを使用する
モジュール5	Azure AI Foundry で言語モデルを微調整する
モジュール6	Azure AI Foundry で言語モデルのパフォーマンスを評価する
モジュール7	Azure AI Foundry での責任ある生成 AI

# モジュール1



## Azure AI Foundryの概要

48 分 • モジュール • 7 ユニット

Azure AI Foundry は、開発者、AI エンジニア、IT プロフェッショナルなど、組織全体が AI ソリューションをより簡単にカスタマイズ、ホスト、実行、管理できるように設計されています。この統一されたアプローチにより、開発と管理のプロセスが簡素化され、すべての利害関係者がイノベーションの推進と戦略的目標の達成に集中できるようになります。

# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# Azure AI Foundryとは？

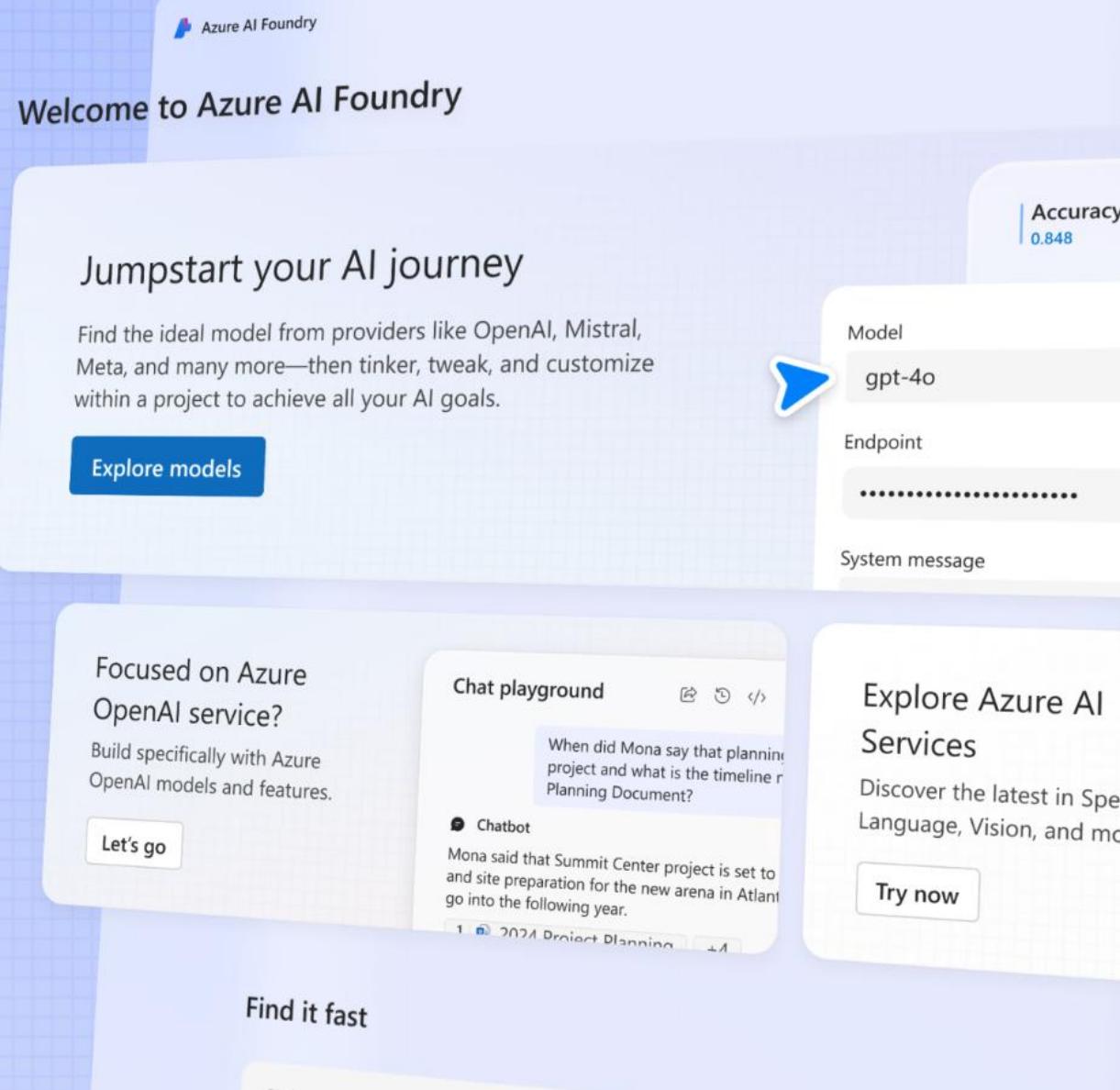
- 生成AIアプリの開発に役立つプラットフォーム
- Azure OpenAI Serviceが提供するGPTなどの生成AIモデル、従来からAzureで提供してきた言語・音声・視覚などのAIサービス、Azure Machine Learningの「プロンプトフロー」など、Azureの主要なAIサービスやツールをこの一つのプラットフォームで扱うことができる
- <https://ai.azure.com/> からアクセスできる



# 境界のない構築

Azure AI Foundry has everything you need to customize, host, run, and manage AI-driven applications built in GitHub, Visual Studio, and Copilot Studio, with APIs for all your needs.

開始するにはサインインしてください



Welcome to Azure AI Foundry

Jumpstart your AI journey

Find the ideal model from providers like OpenAI, Mistral, Meta, and many more—then tinker, tweak, and customize within a project to achieve all your AI goals.

Explore models

Model: gpt-4o

Endpoint: ....

System message

Focused on Azure OpenAI service?  
Build specifically with Azure OpenAI models and features.

Chat playground

When did Mona say that planning project and what is the timeline r Planning Document?

Chatbot

Mona said that Summit Center project is set to and site preparation for the new arena in Atlanta go into the following year.

Explore Azure AI Services

Discover the latest in Spe Language, Vision, and mo

Try now

Find it fast

## Jump into a project in Azure AI Foundry

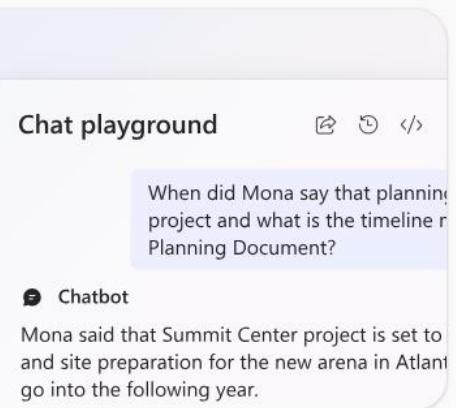
[すべてのプロジェクトを表示](#)[+ プロジェクトの作成](#)

プロジェクト	説明	作成日	場所	ハブ

## プロジェクトの外部で作業する

Azure OpenAI Service  
に重点を置いています  
か?

Azure OpenAI のモデルと機能を使用  
して具体的に構築します。

[始めましょ](#)

## ヘルプ

### Watch a tutorial

Learn how to create generative AI applications in this end-to-end video demoing Azure AI Studio's capabilities.

[AI Show: Build your own Copilot with Azure AI Studio](#)

### Overview ^

#### [What is Azure AI Studio?](#)

Azure AI Studio brings together capabilities from across multiple Azure AI Services.

#### [What are AI services?](#)

Azure AI services help rapidly create applications with prebuilt and customizable APIs and models.

#### [Azure AI Studio architecture](#)

AZURE STUDIO provides a unified experience to build, evaluate, and deploy AI models through a web portal, SDK, or CLI.

#### [Get started with Azure AI SDK](#)

Begin using Azure AI SDK to build generative AI applications.

### Quick starts ▼

### Tutorials ▾

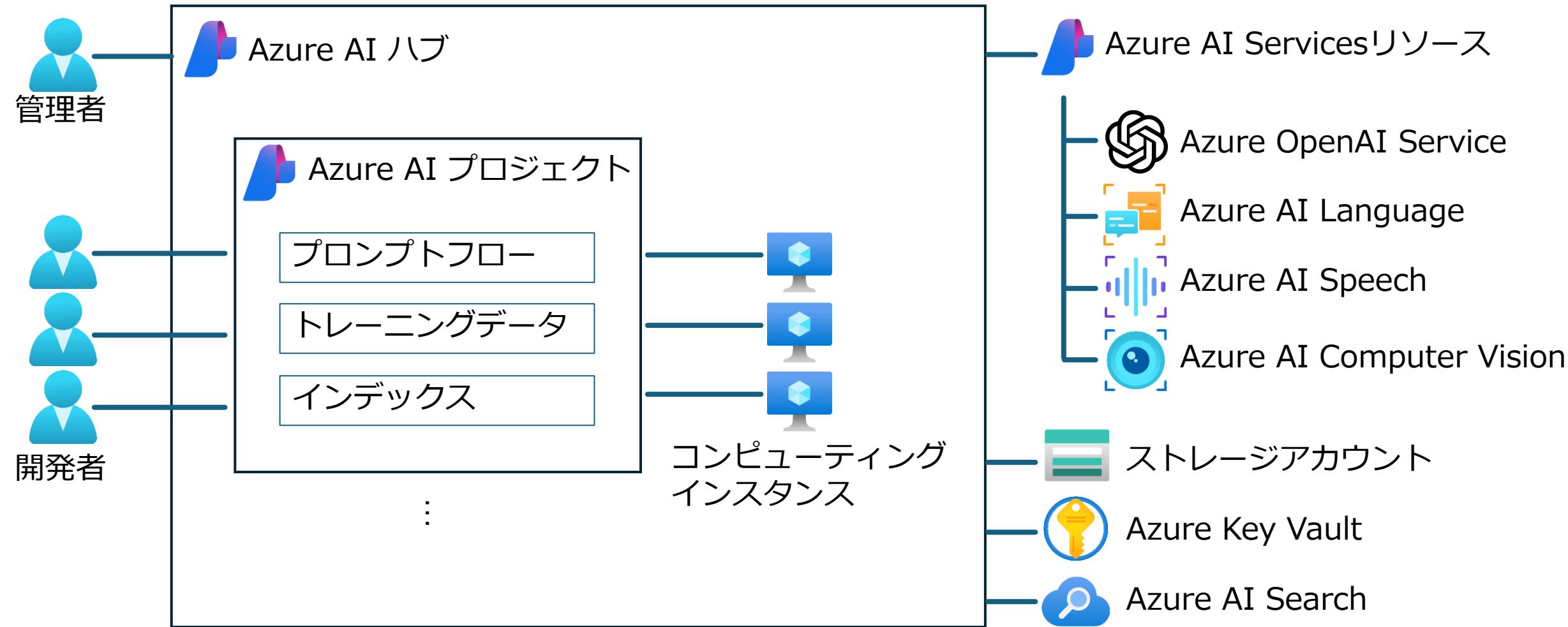
# Azure AI Foundryで利用できる主な機能

- ・**プレイグラウンド**での生成AIモデルの動作確認（モジュール1で解説）
- ・Azure AI Servicesの視覚・音声・言語などの**AIサービス**の動作確認（モジュール1で解説）
- ・**モデルカタログ**でのモデルの検索とデプロイ（モジュール2で解説）
- ・**プロンプトフロー**を使った、モデルの制御フローの構築とデプロイ（モジュール3で解説）
- ・**データの追加**（RAGアーキテクチャの実装）（モジュール4で解説）
- ・**微調整**によるモデルのファインチューニング（モジュール5で解説）
- ・モデルの**パフォーマンス評価**（モジュール6で解説）
- ・**コンテンツフィルター**による、モデルの入出力データのフィルタリング（モジュール7で解説）

# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# Azure AI Foundryの構成



# 「ハブ」とは？

- ・複数のプロジェクトを束ねるリソース
- ・ハブの中には、生成AIアプリの開発者が共同作業を行う場所である「プロジェクト」が作成される
- ・必要に応じて、ハブの中に複数のプロジェクトを作成できる。
- ・ハブには、Azure AI Servicesリソース、ストレージアカウントなどのリソースが接続される。これらはプロジェクトで使用できる。

# 「プロジェクト」とは？

- ・生成AIアプリの開発者が共同作業を行う場所
- ・生成AIアプリを構築するために必要なデータなどを保存するために使用される
- ・プロジェクトからは、 Azure AI Services リソースなどの、 ハブに接続されたリソースにアクセスできる
- ・プロジェクトでは、 必要に応じて「コンピューティングインスタンス」と呼ばれる仮想マシンが使用される
  - ・データを高速に検索するためのベクトルインデックスの作成や、 プロンプトフローの実行などに、 コンピューティングインスタンスが必要

# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

## プロジェクトの作成

プロジェクトを表示

+ 新しいプロジェクト

プロジェクトは、AI を使用して構築するために必要なすべてのものを共同作業、接続、整理するためのさらに優れた方法です。既存の資産は引き続き使用できます。コンテナーを簡単に管理できます。

プロジェクト名 \* ⓘ

Project45512651

カスタマイズ

&gt; Azure リソースが作成されました (新規: User1-45512651-6600\_ai + 4)

ⓘ セキュリティ またはストレージ リソースをカスタマイズする必要がありますか? [Azure portal にアクセスする](#)

プロジェクトを作成する

キャンセル

### Azure AI Studio

次世代のモデル開発、生成 AI アプリケーションを開発する方法について説明します。Web アプリを介して、または SDK を使用してコードで直接行います。

知識の取得、コードイングノリターや関数呼び出しなどのツールを使用して、パワー エージェントのようなエクスペリエンスを開発します。

ハフスーターを調整したり、独自のデータを基にしたりして、さまざまな基本モデルがユーザー入力にどのように応答するかを確認します。

### 大規模な言語モデル (LLM) のデプロイ

LLM またはプロンプト フローをデプロイし、その API をアプリケーション、Web サイト、またはその他の運用環境で使用できるようにします。

## プロジェクトの作成

✓ プロジェクトの詳細

② ハブの作成

③ レビューして終了

## プロジェクトのハブを作成する

ハブは、チームがプロジェクト作業、モデルエンドポイント、コンピューティング、接続、セキュリティ設定を共有するためのコラボレーション環境です。 [詳しく見る](#)

セキュリティまたはハブの [依存リソース](#) をカスタマイズする必要がありますか? [Azure portal にアクセスする](#)

ハブ名 \*

Hub45512651

サブスクリプション \* ⓘ

[新しいサブスクリプションの作成](#)

MOCOAI-lod49253210

リソース グループ \* ⓘ

[新しいリソース グループの作成](#)

ResourceGroup1

場所 \* ⓘ

East US

[選択に関するヘルプ](#)

Azure AI サービスまたは Azure OpenAI を接続する \* ⓘ

[新しいAI サービスの作成](#)

(新しい) ai-Hub45512651

Azure AI 検索に接続する ⓘ

[新しいAI Search の作成](#)

接続をスキップする

戻る

次へ

プロジェクトを作成する

キャンセル

## プロジェクトの作成

プロジェクトの詳細

ハブの作成

③ レビューして終了

## レビューして終了

必要な依存関係と共に、次のリソースが自動的に作成されます。最初のハブとプロジェクトの作成が完了するまで数分かかる場合があります。[ハブと依存関係に関する詳細情報](#).

## ハブ

名前: Hub45512651

サブスクリプション: MOCOAI-10d49253210

リソース グループ: ResourceGroup1

場所: eastus

## プロジェクト

名前: Project45512651

サブスクリプション: MOCOAI-10d49253210

リソース グループ: ResourceGroup1

## AI サービス

名前: ai-Hub45512651

プロジェクトを作成する

キャンセル

## プロジェクトの作成

 プロジェクトの詳細 ハブの作成 ③ レビューして終了

## レビューして終了

必要な依存関係と共に、次のリソースが自動的に作成されます。最初のハブとプロジェクトの作成が完了するまで数分かかる場合があります。[ハブと依存関係に関する詳細情報](#).

## リソース

⌚ Project45512651

⌚ Hub45512651

⌚ ai-hub45512651696892827091

⌚ sthub4551265696892827091

⌚ kv-hub45512696892827091

## タイプ

AI プロジェクト ⓘ

AI ハブ ⓘ

AI サービス ⓘ

ストレージ アカウント ⓘ

キー コンテナー ⓘ

⌚ リソースを作成しています...

プロジェクトを作成する

キャンセル

モデル (LLM) のデ

ンプト フローをデプ  
リをアプリケーショ  
、またはその他の運  
用環境で使用できるようになります。

← すべてのプロジェクトに戻る

現在のプロジェクト Project45512651

| **プロジェクトの概要**

設定

概要 ^

- モデル カタログ
- モデルのベンチマーク
- プロンプト カタログ
- AI サービス

プロジェクトのプレイグラウンド ^

- チャット
- アシスタント プレビュー
- リアルタイマー
- ムオーディオ プレビュー
- イメージ
- 入力候補
- 音声 プレビュー

ツール ^

- コード プレビュー
- プロンプト フロー
- 評価 プレビュー

# Project45512651

前回終了した位置から再開

モデルのデプロイ、データの接続、アセットの作成が完了すると、ここに表示されます... 中断したところからいつでも再開できます。

## ドキュメント

[Build and deploy a question and answer copilot with prompt flow in Azure AI Studio](#)

Use generative AI and prompt flow to build, configure, and deploy a copilot for your retail company called Contoso.

[Build a custom chat app in Python using the prompt flow SDK](#)

Develop a custom chat app in your local development environment, and learn to run and evaluate LLM prompts with the Prompt Flow SDK.

[Retrieval augmented generation and indexes](#)

This article talks about the importance and need for Retrieval Augmented Generation (RAG) and Index in generative AI.

その他を表示

[プロジェクトの削除](#)

プロジェクトユーザー 1 すべて表示 →

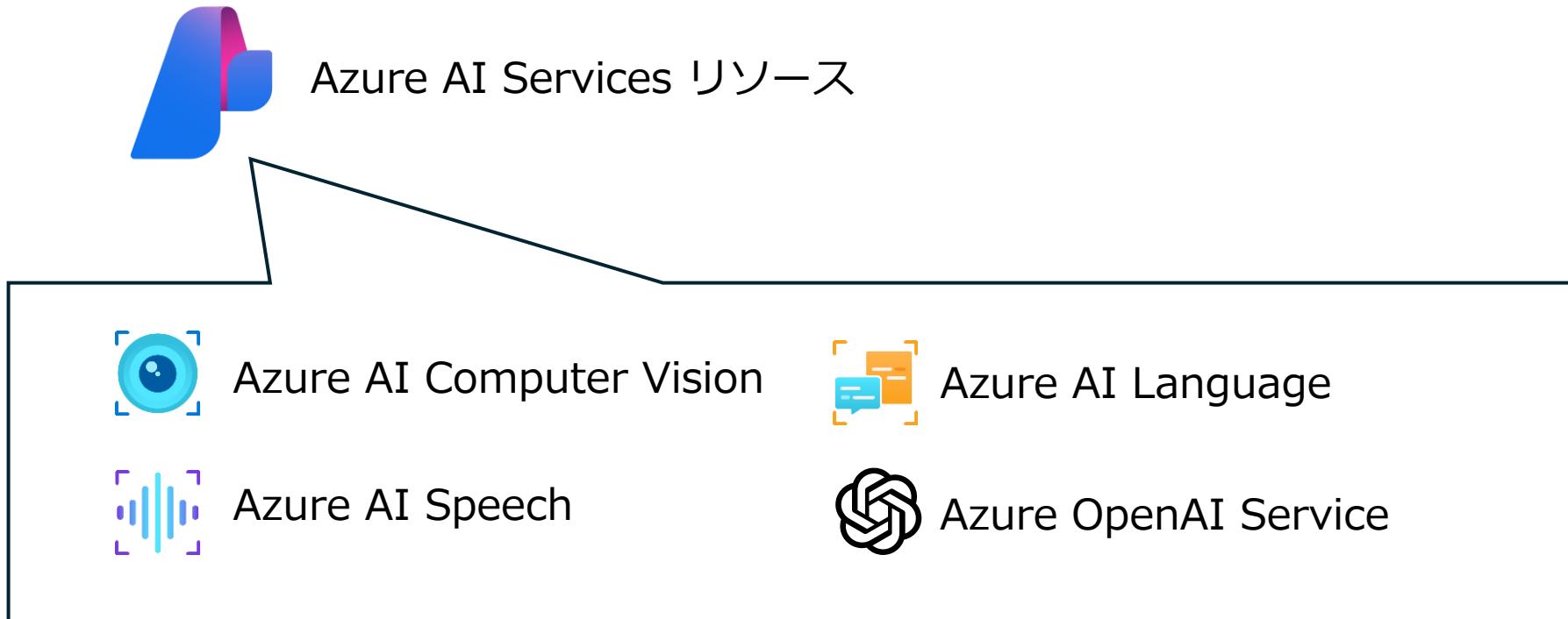
+ 新しいユーザー

# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# Azure AI Services リソースとは？

- Azureの視覚・音声・言語などのAIサービスとAzure OpenAI Serviceを使用できるリソース
- Azure AI Foundryでハブを作成する際に一緒に作成される
- 「マルチサービスアカウント」と似たようなもの



# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# チャットプレイグラウンドの利用例

- ・モデルカタログでGPT-4oモデルを選択し、デプロイ
- ・チャットプレイグラウンドでGPT-4oモデルを使用して質問に答えてもらう

## カスタム AI ソリューションを構築するための適切なモデルを見つける

フィルターの表示

すべてのフィルター

コレクション

業界

デプロイ オプション

推論タスク

タスクの微調整

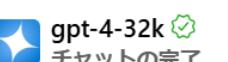
ライセンス

gpt-4

モデル 80



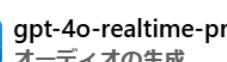
チャットの完了



チャットの完了



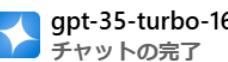
チャットの完了



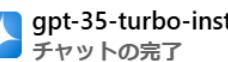
オーディオの生成



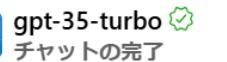
チャットの完了



チャットの完了



チャットの完了



チャットの完了



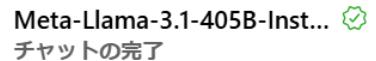
チャットの完了



チャットの完了



テキスト生成



チャットの完了



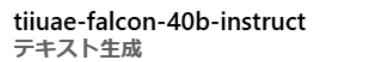
テキスト生成



テキスト生成



テキスト生成



テキスト生成



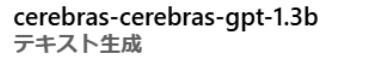
テキスト生成



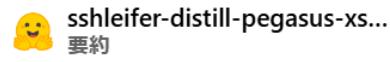
テキスト生成



テキストから画像へ



テキスト生成



要約



塗りつぶしマスク



テキスト生成



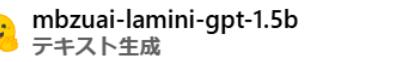
テキストの分類



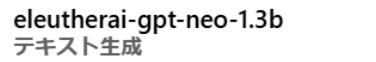
テキスト生成



テキスト生成



テキスト生成



テキスト生成

← すべてのプロジェクトに戻る

現在のプロジェクト  
Project45512651

プロジェクトの概要

設定

概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイラウンド

チャット

アシスタン

リアルタイ

ム オーディ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプト フロー

評価 プレビュー

← すべてのプロジェクトに戻る

現在のプロジェクト Project45512651

プロジェクトの概要

設定

概要

モデルカタログ

- モデルのベンチマーク
- プロンプトカタログ
- AIサービス

プロジェクトのプレイグラウンド

- チャット
- アシスタント プレビュー
- リアルタイム
- ムオーディ プレビュー
- イメージ
- 入力候補
- 音声 プレビュー

ツール

- コード プレビュー
- プロンプトフロー
- 評価 プレビュー

# ← gpt-4o

## 概要

タスク: チャットの完了 | 微調整タスク: chat-completion | ライセンス: custom

デプロイ | 更新 | 微調整

### 説明

GPT-4o offers a shift in how AI models interact with multimodal inputs. By seamlessly combining text, images, and audio, GPT-4o provides a richer, more engaging user experience.

Matching the intelligence of GPT-4 Turbo, it is remarkably more efficient, delivering text at twice the speed and at half the cost. Additionally, GPT-4o exhibits the highest vision performance and excels in non-English languages compared to previous OpenAI models.

GPT-4o is engineered for speed and efficiency. Its advanced ability to handle complex queries with minimal resources can translate into cost savings and performance.

The introduction of GPT-4o opens numerous possibilities for businesses in various sectors:

- Enhanced customer service:** By integrating diverse data inputs, GPT-4o enables more dynamic and comprehensive customer support interactions.
- Advanced analytics:** Leverage GPT-4o's capability to process and analyze different types of data to enhance decision-making and uncover deeper insights.
- Content innovation:** Use GPT-4o's generative capabilities to create engaging and diverse content formats, catering to a broad range of consumer preferences.

**Note: updated version 2024-08-06**

GPT-4o has been released under a new version 2024-08-06 which brings new functionalities and a larger output size (from 4,096 to 16,384).

### モデル ID

azureml://registries/azure-openai/models/gpt-4o/versions/2024-08-06

← すべてのプロジェクトに戻る

現在のプロジェクト Project45512651

プロジェクトの概要

設定

概要

モデルカタログ

- モデルのベンチマーク
- プロンプトカタログ
- AIサービス

プロジェクトのプレイグラウンド

- チャット
- アシスタント プレビュー
- リアルタイマー
- ムオーディ プレビュー
- イメージ
- 入力候補
- 音声 プレビュー

ツール

- コード プレビュー
- プロンプトフロー
- 評価 フロー

# ← gpt-4o

## 概要

タスク: チャットのデプロイ

### モデル gpt-4o をデプロイする

デプロイ名 \*

説明

GPT-4o offers a shift in combining text, images, and audio experience.

Matching the intelligence of text at twice the speed of vision performance and OpenAI models.

GPT-4o is engineered for queries with minimal reprocessing.

The introduction of GPT-4o in three sectors:

1. Enhanced customization: more dynamic and personalized responses.

2. Advanced analytics: different types of data integration and processing.

3. Content innovation: Use GPT-4o's generative capabilities to create engaging and diverse content formats, catering to a broad range of consumer preferences.

デプロイの詳細

モデルバージョン: 2024-05-13

デプロイの種類: グローバル標準

プロジェクト: Project45512651

接続されているAIリソース: ai-hub45512651696892827091\_aoai

容量: 10K 1分あたりのトークン数 (TPM)

認証の種類: キー

コンテンツの安全性: DefaultV2

リソースの場所: East US

カスタマイズ

デプロイ 取り消し

Note: updated version 2024-08-06

GPT-4o has been released under a new version 2024-08-06 which brings new functionalities and a larger output size (from 4,096 to 16,384).

# gpt-4o

← すべてのプロジェクトに戻る 現在のプロジェクト Project45512651 ▾

プロジェクトの概要 設定 概要 モデルカタログ モデルのベンチマーク プロンプトカタログ AIサービス プロジェクトのプレイグラウンド チャット アシスタント リアルタイム ムオーディオ ブラウジング イメージ 入力候補 音声 プレビュー ツール コード プレビュー プロンプトフロー 評価 プレビュー

詳細 メトリック リスクと安全性

プレイグラウンドで開く 編集 削除

### 配置情報

名前	gpt-4o	プロビジョニングの状態	成功
展開の種類	グローバル標準	作成日	2024-11-07T13:16:51.1280184Z
作成者	868ea4cd-8462-47de-b7f0-d20d820b1334	変更日	Nov 7, 2024 10:16 PM
更新者	868ea4cd-8462-47de-b7f0-d20d820b1334	バージョン更新ポリシー	新しい既定のバージョンが使用可能になったら
レート制限 (1分あたりのトークン数)	10,000	レート制限 (1分あたりの要求数)	60
モデル名	gpt-4o	モデルバージョン	2024-05-13
ライフサイクルの状態	GenerallyAvailable	作成日	May 13, 2024 9:00 AM
更新日	Aug 21, 2024 9:00 AM	モデルの廃止日	May 20, 2025 9:00 AM

### エンドポイント

ターゲット URI  
<https://ai-hub45512651696892827091.openai.azure.com/openai/deployments/gpt-4o>

認証の種類  
キー  
キー  
.....

### 監視と安全性

コンテンツ フィルター  
DefaultV2

### アプリケーション開発に役立つリンク

コードサンプル リポジトリ  
チュートリアル

# チャットプレイグラウンド

コードの表示 プロンプトフロー 評価 デプロイ インポート エクスポート プロンプトのサンプル ...

## セットアップ

デプロイ \* 新しいデプロイの作成  
gpt-4o (version:2024-05-13)

モデルに指示とコンテキストを与える 

あなたはAzureのプロフェッショナルです。日本語で300文字以内で回答してください。

保存 

+ セクションの追加

データを追加する パラメーター

チャット機能 JSONの表示

チャットを開始する

チャットのプレイグラウンドで、見る、聞く、話すことができるようになりました。チャット ウィンドウでマイクを選択し、音声入力を開始して、テキストを手動で入力せずにモデルにプロンプトを表示します。スピーカー アイコンを選択して、モデルの出力を聞くこともできます。

ここにユーザー クエリを入力します。(Shift + Enter で新しい行)

33/128000 送信するトークン   

# チャットプレイグラウンド

コードの表示 プロンプトフロー 評価 デプロイ インポート エクスポート プロンプトのサンプル ...

## セットアップ

デプロイ \* 新しいデプロイの作成  
gpt-4o (version:2024-05-13)

モデルに指示とコンテキストを与える ⓘ  
あなたはAzureのプロフェッショナルです。日本語で300文字以内で回答してください。

保存 + セクションの追加

データを追加する プレビュー  
パラメーター

チャットのクリア チャット機能 JSON の表示

チャットを開始する

チャットのプレイグラウンドで、見る、聞く、話すことができるようになりました。チャット ウィンドウでマイクを選択し、音声入力を開始して、テキストを手動で入力せずにモデルにプロンプトを表示します。スピーカー アイコンを選択して、モデルの出力を聞くこともできます。

Azure AI Studioとはどのようなものですか?| 48/128000 送信するトークン ⌂ ⌂ ⌂

# チャットプレイグラウンド

← すべてのプロジェクトに戻る

現在のプロジェクト Project45512651

プロジェクトの概要 設定

概要 モデル カタログ モデルのベンチマーク プロンプト カタログ AI サービス

プロジェクトのプレイグラウンド チャット

- アシスタント プレビュー
- リアルタイム プレビュー
- ムオーディオ プレビュー
- イメージ
- 入力候補
- 音声 プレビュー

ツール コード プレビュー プロンプト フロー 評価

## セットアップ

非表示

デプロイ \* 新しいデプロイの作成

gpt-4o (version:2024-05-13)

モデルに指示とコンテキストを与える ⓘ

あなたはAzureのプロフェッショナルです。日本語で300文字以内で回答してください。

保存 + セクションの追加

データを追加する プレビュー

パラメーター

チャットのクリア チャット機能 JSON の表示

Azure AI Studioとはどのようなものですか?

Azure AI Studioは、Microsoft Azureが提供するクラウドベースのAI開発プラットフォームです。このツールは、開発者やデータサイエンティストが機械学習モデルやAIソリューションを簡単に作成、トレーニング、デプロイできるように設計されています。Azure AI Studioは、Azure Machine Learning、Cognitive Services、Bot ServicesなどのさまざまなAzure AIサービスと統合されています。これにより、ユーザーは視覚的なインターフェースやコードベースのアプローチを利用して、AIプロジェクトのライフサイクル全体を効率的に管理することができます。さらに、データの前処理、モデルの評価、監視機能も提供しており、エンドツーエンドのAIソリューション開発をサポートします。

ここにユーザー クエリを入力します。(Shift + Enter で新しい行)

318/128000 送信するトークン   

# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# イメージプレイグラウンドの利用例

- ・モデルカタログでDALL-E（ダリ）モデルを選択し、デプロイ
- ・イメージプレイグラウンドでのDALL-Eモデルを使用して画像を生成する

## カスタム AI ソリューションを構築するための適切なモデルを見つける

すべてのフィルター コレクション 業界 デプロイ オプション 推論タスク タスクの微調整 ライセンス

dall-e

モデル 335

プロジェクトの概要 設定 概要 モデル カタログ モデルのベンチマーク プロンプト カタログ AI サービス プロジェクトのプレイ ラウンド チャット アシスタン ト リアルタイ ム オーディ オ イメージ 入力候補 音声 プレビュー ツール コード プレビュー プロンプト フロー 評価 プレビュー

dall-e-3 テキストから画像へ dall-e-2 テキストから画像へ text-embedding-3-small 埋め込み text-embedding-3-large 埋め込み Cohere-embed-v3-multilingual 埋め込み Cohere-embed-v3-english 埋め込み text-embedding-ada-002 埋め込み Cohere-rerank-v3-english テキストの分類 richieIo-small-e-czech-fin... トーケンの分類 intfloat-multilingual-e5-la... feature-extraction baai-bge-large-en-v1.5 feature-extraction OpenAI-CLIP-Image-Text... 埋め込み Facebook-DinoV2-Image... 埋め込み Facebook-DinoV2-Image... 埋め込み gokaygokay-flux-prompt... テキストからテキストへの生成 prometheus-eval-prometh... テキストからテキストへの生成 prometheus-eval-prometh... テキストからテキストへの生成 equall-saul-7b-instruct-v1 テキスト生成 alibaba-nlp-gte-large-en... sentence-similarity distilbert-base-uncased-fi... テキストの分類 helsinki-nlp-opus-mt-de-en トランスレーター distilbert-base-uncased-fi... テキストの分類 j-hartmann-emotion-engli... テキストの分類 helsinki-nlp-opus-mt-en-de トランスレーター helsinki-nlp-opus-mt-en-es トランスレーター helsinki-nlp-opus-mt-zh-en トランスレーター helsinki-nlp-opus-mt-en-zh トランスレーター

← すべてのプロジェクトに戻る

現在のプロジェクト Project45512651

プロジェクトの概要

設定

概要

モデルカタログ

モデルのベンチマーク

プロンプトカタログ

AIサービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイマー

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

## ← dall-e-3

概要

タスク: テキストから画像へ

▷ デプロイ (Red Box)

⟳ 更新

微調整

説明

**Model Family: DALL-E**

DALL-E 3 generates images from text prompts that are provided by the user. DALL-E 3 is generally available for use on Azure OpenAI.

The image generation API creates an image from a text prompt. It does not edit existing images or create variations.

Learn more at: <https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/models#dall-e>

モデル ID ⓘ

azureml://registries/azure-openai/models/dall-e-3/versions/3.0

← dall-e-3

すべてのプロジェクトに戻る

現在のプロジェクト Project45512651

プロジェクトの概要

設定

概要

モデルカタログ

モデルのベンチマーク

プロンプトカタログ

AIサービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイマー

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

## 概要

タスク: テキストから画像生成

デプロイ

### モデル dall-e-3 をデプロイする

デプロイ名 \*

説明

Model Family: DALL-E 3

DALL-E 3 generates images from text descriptions. It is generally available for prompt-based image generation.

The image generation process takes a text prompt and either generates a new image or creates a variation of an existing image or creates a variation of an existing image.

Learn more at: <https://services/openai/concepts/dall-e-3>

デプロイの詳細

カスタマイズ

モデルバージョン	3.0	デプロイの種類	Standard
プロジェクト	Project45512651	接続されている AI リソース	ai-hub45512651696892827091_aoai
容量	1 容量ユニット (CU)	認証の種類	キー
コンテンツの安全性	既定	リソースの場所	East US

デプロイ 取消し

openai/models/dall-e-3/versions/3.0

# dall-e-3

詳細 メトリック

 [プレイグラウンドで開く](#)  [編集](#)  [削除](#)

### 配置情報

名前	dall-e-3	プロビジョニングの状態	成功
展開の種類	Standard	作成日	2024-11-07T13:22:05.8938368Z
作成者	868ea4cd-8462-47de-b7f0-d20d820b1334	変更日	Nov 7, 2024 10:22 PM
更新者	868ea4cd-8462-47de-b7f0-d20d820b1334	バージョン更新ポリシー	新しい既定のバージョンが使用可能になったら

### 容量ユニット

レート制限 (1分あたりのトークン数)	--	レート制限 (1分あたりの要求数)	3
---------------------	----	-------------------	---

### モデル名

dall-e-3	モデルバージョン	3.0
----------	----------	-----

### ライフサイクルの状態

Preview	作成日	Aug 11, 2023 9:00 AM
---------	-----	----------------------

### 更新日

Aug 11, 2023 9:00 AM	モデルの廃止日	Apr 30, 2025 9:00 AM
----------------------	---------	----------------------

### エンドポイント

ターゲット URI  
<https://ai-hub45512651696892827091.openai.azure.com/openai/deployment...> 

認証の種類  
キー  
キー  
\*\*\*\*\*  

### 監視と安全性

コンテンツ フィルター  
既定

### アプリケーション開発に役立つリンク

[コードサンプルリポジトリ](#)  [チュートリアル](#) 

- セーレル カタログ
- モデルのベンチマーク
- プロンプト カタログ
- AI サービス

## プロジェクトのプレイグラウンド

- チャット
- アシスタンス プレビュー
- リアルタイム
- ムオーディオ プレビュー

## イメージ

- 入力候補
- 音声 プレビュー

## ツール

- コード プレビュー
- プロンプト フロー
- 評価 プレビュー
- 微調整 プレビュー

## コンポーネント

- データ
- インデックス
- デプロイ
- コンテンツ フィルタ

## 画像プレイグラウンド

デプロイ dall-e-3

フィードバックの送信

コードの表示

検索

タイルサイズ 中タイル

プロンプト ①

かわいらしい猫がふわふわの雲に寝そべっている

生成



かわいらしい猫がふわふわの雲に寝そべっている



# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# Azure AI Servicesの利用例

- Azure AI Servicesに含まれる視覚・音声・言語などのAIサービスの利用が可能
- ここでは例として、視覚（Computer Vision）の「画像の説明文の生成」を紹介

← すべてのプロジェクトに戻る現在のプロジェクト  
Project45512651

## プロジェクトの概要

設定

## 概要 ^

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

## プロジェクトのプレイグラウンド ^

チャット

アシスタント プレビュー

リアルタイ

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

## ツール ^

コード プレビュー

プロンプト フロー

訳語 プレビュー

## Project45512651

## 前回終了した位置から再開

モデルのデプロイ、データの接続、アセットの作成が完了すると、ここに表示されます... 中断したところからいつでも再開できます。

## ドキュメント

その他を表示

Build and deploy a question and answer copilot with prompt flow in Azure AI Studio

Use generative AI and prompt flow to build, configure, and deploy a copilot for your retail company called Contoso.

Build a custom chat app in Python using the prompt flow SDK

Develop a custom chat app in your local development environment, and learn to run and evaluate LLM prompts with the Prompt Flow SDK.

Retrieval augmented generation and indexes

This article talks about the importance and need for Retrieval Augmented Generation (RAG) and Index in generative AI.

## 説明

プロジェクトの説明 (オプション)

## プロジェクトのプロパティ

## ハブの表示名

Hub45512651

## ハブリソース名

hub45512651

## サブスクリプション

MOCOAI-lod49253210

## リソースグループ

ResourceGroup1

Azure Portal で管理する

API エンドポイントとキーを取得する

サブスクリプション クォータを表示する

プロジェクトの削除

## プロジェクトユーザー

2

すべて表示 →

project45512651

Azure AI Administrator

User1-45512651 (あなた)

所有者

## 生成 AI と統合する



### 音声分析 レビュー

オーディオとビデオの録音を文字起こしし、要約などの拡張された出力を生成したり、重要なトピック、個人を特定できる情報 (PII)、センチメントなどの重要な情報を抽出したりします。



### ドキュメント フィールドの抽出 レビュー

カスタム生成抽出モデルを使用して、ドキュメントとフォームからフィールドを抽出します。



### 生成 AI のプロンプト シールド

プロンプト シールドを使用すると、脱獄の試みを検出およびブロックすることで、生成 AI の安全性を確保できます。この重要な機能を使用して、コミュニティを安全で信頼できる状態に保ってください。



### 生成 AI を使用して要約する

シンプルな API 呼び出しを使用して高品質の要約を取得し、エンタープライズ規模で情報を簡素化します。

## AI 機能を使用してソリューションを活性化する



音声テキスト変換、テキスト読み上げ、音声翻訳の機能を通じて、カスタマーエクスペリエンスを強化します。

[Speech のすべての機能を表示する](#)



**Language + Translator**  
LLM が駆動する自然言語処理機能を使用して、分析、要約、翻訳を行います。  
[Language + Translator のすべての機能を表示する](#)



**Vision + Document**  
OCR とマルチモーダル AI を使用して、ドキュメント、画像、ビデオから情報と分析情報を検出します。  
[Vision + Document のすべての機能を表示する](#)



**Content Safety**  
テキスト、画像、マルチモーダル API など、アプリ内の有害、攻撃的、または不適切なユーザー生成コンテンツまたは AI 生成コンテンツを検出します。  
[Content Safety のすべての機能を表示する](#)



## Vision + Document

Give your apps the ability to read text, analyze images, process documents and detect faces with technology like optical character recognition (OCR) and machine learning.

### Integrate vision with generative AI

#### Document field extraction プレビュー

Extract fields from documents and forms using a custom generative extraction model.

[Try demo](#)

### View all other vision capabilities

Document

Face

Image

#### Common object detection

Recognize the location of objects of interest in an image and assign them a label.

[Try demo](#)

#### Image captioning

Generate a human-readable sentence that describes the content of an image.

[Try demo](#)

#### Dense captioning

Generate human-readable captions for all important object detected in your image.

[Try demo](#)

#### Image search

Retrieve specific moments within your photo album. For example, you can search for: a wedding you attended last summer, your pet or your favorite city.

[Try demo](#)

#### Common tag extraction

Detect the location of one or more human faces in images, along with attributes such as pose, face mask and facial landmarks.

[Try demo](#)

#### Optical character recognition

Extract printed and handwritten style text from images and documents for supported languages.

[Try demo](#)

#### Other image analysis capabilities

Generate a human-readable sentence that describes the content of an image.

[View more capabilities](#)

← すべてのビジョン  
とドキュメントへ

現在のプロジェクト  
Project45512651

ドキュメン  
ト ファイル プレビュー  
ドの抽出

ドキュメント

請求書

受け取り

ID ドキュメント

医療保険カード

米国税申告用紙

米国住宅ローンフォ  
ーム

結婚証明書

クレジットカード

コントラクト

読み取り

レイアウト

Face

画像内の顔を検出す  
る

イメージ

物体検出

画像のキャプション

### Connected Azure AI Services \*

ai-hub45512651696892827091 (eastus)

Use one of your own files or choose from a sample below.



Drag and drop video file here,  
or  
Browse for a file,  
or  
Take a photo



### Sample Image 2



Detected attributes

JSON

a man holding a surfboard on a rock

Azure AI ServicesのComputer Visionの機能により、画像の説明文が生成される。

A man holding a surfboard on a rock  
(岩の上でサーフボードを抱えている  
男性)

# モジュール1 Azure AI Foundryの概要

- Azure AI Foundryとは？
- Azure AI Foundryの構成
- 「ハブ」と「プロジェクト」の作成例
- Azure AI Servicesリソースとは？
- チャットプレイグラウンドの利用例（GPT-4oモデル）
- イメージプレイグラウンドの利用例（DALL-Eモデル）
- Azure AI Servicesの利用例（Azure AI Computer Vision）
- まとめ

# モジュール1まとめ

- Azure AI Foundryは、生成AIアプリの開発に役立つプラットフォーム。
- Azure OpenAI Serviceが提供するGPTなどの生成AIモデル、従来からAzureで提供されてきた言語・音声・視覚などのAIサービス、Azure Machine Learningの「プロンプトフロー」など、Azureの主要なAIサービスやツールをこの一つのプラットフォームで扱うことができる
- Azure AI Foundryでは「ハブ」と「プロジェクト」を作成する。「ハブ」の中で複数の「プロジェクト」が管理される。「ハブ」には、Azure AI Servicesリソースや、データを記録するためのストレージアカウントなどが接続される
- Azure AI Foundryで、生成AIモデルをデプロイし、プレイグラウンドを使用して動作を確認できる

# モジュール2



Azure AI Foundry でモデル カタログのモデルを調べてデプロイする

39 分 残り • モジュール • 1/7 ユニットが完了しました

Azure AI Foundry のモデル カタログで利用できる様々な言語モデルについて調べます。モデルを選択、デプロイ、テストする方法と、そのパフォーマンスを向上させる方法を理解します。

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# モデルカタログとは？

- Azure AI Foundryで利用できる、機械学習モデル・生成AIモデルのコレクション
- 1,600を超えるモデルが用意されている
  - OpenAI GPT-3.5, GPT-4 など
  - Cohere（コヒーア）Command R など
  - Meta Llama（ラマ）など
  - Microsoft Phi-3（ファイスリー）など

# カスタム AI ソリューションを構築するための適切なモデルを見つける

[フィルターの表示](#)

## アナウンス

### Cohere からのニュース



Embed 3 のマルチモーダル機能が Cohere のコレクションで利用できるようになり...

[モデルの表示](#)[ブログを読む](#)

### Mistral からの新しい SLM



Minstral 3B は、コンピューティング高効率で低遅延のソリューションを提供します。

[モデルの表示](#)[ブログを読む](#)

### Meta Llama 3.2 モデルがここに表示されます。



Llama 3.2 11B Vision Instruct と 90B Vision Instruct は、画像推論のユースケースに...

[モデルの表示](#)[ブログを読む](#)

### o1 モデルを体験する



o1 シリーズは、科学とコーディングの問題を解決するために強化された推論機能を...

[制限付きアクセスを試す](#)[ブログを読む](#)[すべてのフィルター](#)[コレクション](#)[業界](#)[デプロイ オプション](#)[推論タスク](#)[タスクの微調整](#)[ライセンス](#) 検索**モデル 1795**

gpt-4o-realtime-preview

オーディオの生成



openai-whisper-large-v3

音声認識



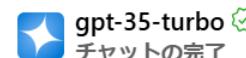
openai-whisper-large

音声認識



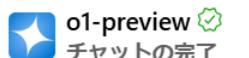
gpt-4

チャットの完了



gpt-35-turbo

チャットの完了



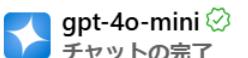
o1-preview

チャットの完了



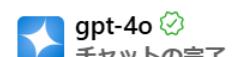
o1-mini

チャットの完了



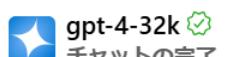
gpt-4o-mini

チャットの完了



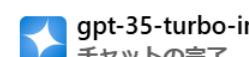
gpt-4o

チャットの完了



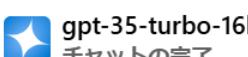
gpt-4-32k

チャットの完了



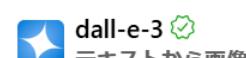
gpt-35-turbo-instruct

チャットの完了



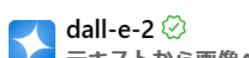
gpt-35-turbo-16k

チャットの完了



dall-e-3

テキストから画像へ



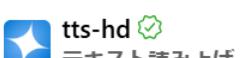
dall-e-2

テキストから画像へ



whisper

音声認識



tts-hd

テキスト読み上げ

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# デプロイとは？

- ・モデルカタログから使用したいモデルを選び、仮想マシンに配置して、モデルを利用できるようにする作業のこと
- ・モデルカタログのモデルの多くは仮想マシンに配置される
  - ・ただし「Azure OpenAI Serviceのモデル」と、「サーバーレスAPI」に対応したモデルについては、仮想マシンは使用されない
- ・デプロイが完了すると、モデルにアクセスするための**エンドポイント（ターゲットURIとキー）**が利用可能となる

← すべてのリソースへ

現在のハブ  
Hub45512651

ハブの概要

すべてのプロジェクト

概要   
モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プレイグラウンド   
チャット

アシスタント プレビュー

音声 プレビュー

共有リソース   
デプロイ

接続

コンピューティング インスタンス

ユーザー

コンテンツ フィルター

## モデル デプロイ

プライベート API キーとエンドポイント URI (Uniform Resource Identifier) を使用してモデルをデプロイします。

モデル デプロイ アプリのデプロイ サービスのデプロイ

+ モデルのデプロイ 最新の情報に更新 編集 プレイグラウンドで開く ビューのリセット

名前	モデル名	モデル バージョン	状態	モデルの廃止日	コンテンツ フィル
ai-hub45512651696892827091_aoui  Azure OpenAI	dall-e-3	dall-e-3	3.0	成功	既定
<b>gpt-4o</b> <span style="border: 2px solid red; padding: 2px;">gpt-4o</span>		gpt-4o	2024-05-13	成功	DefaultV2

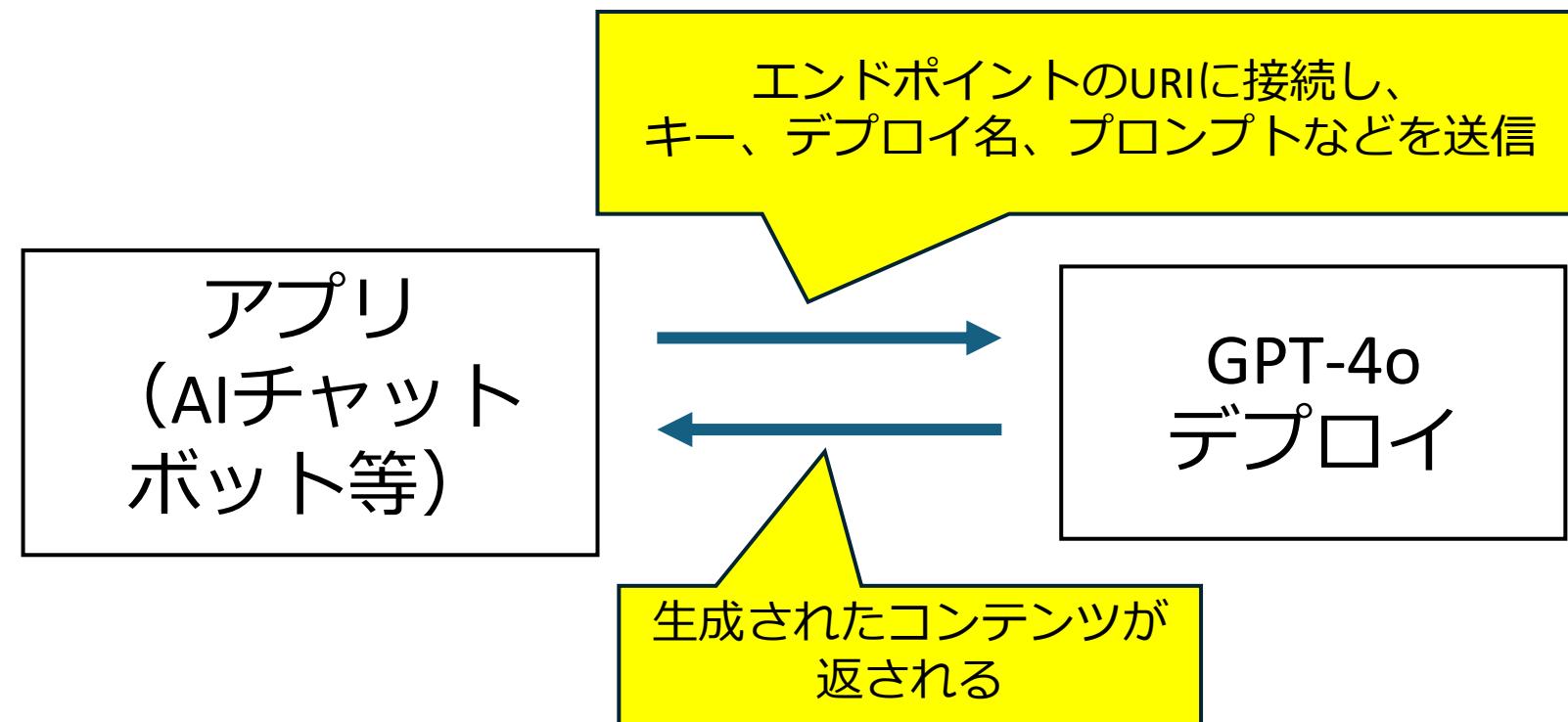
gpt-4o

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# アプリからのデプロイの利用

- 生成AIを使用するアプリは、エンドポイント（ターゲットURI）・キー・デプロイ名の3つの情報を使用して、生成AIを利用できる。



← すべてのリソースへ

現在のハブ Hub45512651

ハブの概要

すべてのプロジェクト

概要

- モデル カタログ
- モデルのベンチマーク
- プロンプト カタログ
- AI サービス

プレイグラウンド

- チャット
- アシスタント プレビュー
- 音声 プレビュー

共有リソース

- デプロイ
- 接続
- コンピューティング インスタンス
- ユーザー
- コンテンツ フィルター

# gpt-4o

詳細 メトリック

プレイグラウンドで開く 編集 削除

**配置情報**

名前	gpt-4o	プロビジョニングの状態	成功
展開の種類	グローバル標準	作成日	2024-11-07T13:16:51.1280184Z
作成者	868ea4cd-8462-47de-b7f0-d20d820b1334	変更日	Nov 7, 2024 10:16 PM
更新者	868ea4cd-8462-47de-b7f0-d20d820b1334	バージョン更新ポリシー	新しい既定のバージョンが使用可能になったら

---

レート制限 (1分あたりのトークン数)	10,000	レート制限 (1分あたりの要求数)	60
---------------------	--------	-------------------	----

---

モデル名	gpt-4o	モデル バージョン	2024-05-13
ライフサイクルの状態	GenerallyAvailable	作成日	May 13, 2024 9:00 AM
更新日	Aug 21, 2024 9:00 AM	モデルの廃止日	May 20, 2025 9:00 AM

**エンドポイント**

ターゲット URI  
[https://ai-hub45512651696892827091.openai.azure.com/openai/deployments/...](https://ai-hub45512651696892827091.openai.azure.com/openai/deployments/)

認証の種類  
キー  
キー  
.....

**監視と安全性**

コンテンツ フィルター  
DefaultV2

**アプリケーション開発に役立つリンク**

コードサンプル リポジトリ   
チュートリアル

# エンドポイント（ターゲットURIとキー）

- ターゲットURIの例:

`https://ai-hub45512651696892827091.openai.azure.com/openai/deployments/gpt-4o/chat/completions?api-version=2024-08-01-preview`

- キーの例:

`G7sPn32zd2MqCq` (中略) `w3AAAAAAC0GL5bb`

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# Azure OpenAI Serviceのモデルのデプロイ

- ・モデルカタログにはAzure OpenAI Serviceのモデルが含まれている。
  - GPT-3.5, GPT-3.5-turbo
  - GPT-4, GPT-4o
  - DALL-E (ダリ) 2, DALL-E3
  - Whisper (ウィスパー) -1, Whisper-2, Whisper-3
  - TTS-1, TTS-1-HD など
- ・使用したいモデルを選択して「デプロイ」する
- ・これらの「デプロイ」では、仮想マシンは使用されない
- ・従量課金で利用できる（1メガトークンの入力あたり\$2.5、1メガトークンの出力あたり\$10、などのように定義される）

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# マネージドコンピューティングを利用する デプロイ

- モデルカタログ内の（Azure OpenAI Serviceのモデルを除く）ほとんどのモデルは「マネージドコンピューティング」と呼ばれる仮想マシンにデプロイされる。
  - Microsoft Phi-3（ファイスリー）
  - Meta Llama（ラマ）
  - NVIDIA Nemotron（ネモトロン）
  - など
- 「マネージドコンピューティング」の特徴
  - Azure上に作成される
  - 仮想マシンのサイズやインスタンス数を指定する

[← すべてのリソースへ](#)現在のハブ  
Hub45512651[ハブの概要](#)[すべてのプロジェクト](#)

概要

## モデル カタログ

[モデルのベンチマーク](#)[プロンプト カタログ](#)[AI サービス](#)

プレイグラウンド

[チャット](#)[アシスタント プレビュー](#)[音声 プレビュー](#)

共有リソース

[デプロイ](#)[接続](#)[コンピューティング インスタンス](#)[ユーザー](#)[コンテンツ フィルター](#)[← facebook-s](#)[概要](#) バージョン 成果[タスク: 画像のセグ.](#)[▷ デプロイ](#)

説明

The Segment Anything Model is a generative model that can generate prompts such as points, lines, polygons, and text to segment objects in an image. It has been trained on a large dataset of images and masks, and has strong performance on various segmentation tasks.

The SAM model is made up of several components:

- The VisionEncoder: takes an image and produces point embeddings using a ViT-like architecture. A learned Positional Embedding is used to handle images of different sizes.
- The PromptEncoder: takes point embeddings and generates prompts based on user input.
- The MaskDecoder: takes prompts and generates a mask by combining them with point embeddings.
- The Neck: predicts the final output based on the generated mask.

## Training Data

### Training Data

See [here](#) for an overview.

## モデルのデプロイ

選択したモデルに対して、スコアリング スクリプトと環境が自動的に生成されます。

[詳細情報](#)[新しいプロジェクトの作成](#)

project45512651

仮想マシン \* ⓘ

Standard\_D16as\_v4 16 コア, 64 GB (RAM), 128 GB (ディスク), \$0.77/時間

インスタンス数 \* ⓘ

1

### エンドポイント

 新規  既存

エンドポイント名 \* ⓘ

hub45512651-nyfgd

ⓘ エンドポイント URL は、エンドポイントの作成後に生成されます。

<https://hub45512651-nyfgd.eastus.inference.ml.azure.com/score>

[実行方法を調べる](#)

デプロイ名 \* ⓘ

facebook-sam-vit-base-4

推論データ収集 ⓘ

 無効[デプロイ](#)[キャンセル](#)

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# サーバーレスAPIを利用するデプロイ

- ・モデルカタログ内の一部のモデルは「サーバーレスAPI」としてデプロイできる。
  - ・ Cohere (コヒーア) Command R
  - ・ Meta Llama (ラマ)
  - ・ Microsoft Phi-3 (ファイスリー)
  - ・ など
- ・ 「サーバーレスAPI」の特徴
  - ・ 仮想マシンなどのインフラを管理する必要がない
  - ・ 従量課金で利用できる

← すべてのリソースへ

## ← Cohere-command-r-08-2024

概要

タスク: チャットの

▷ デプロイ

説明

Command R 08-2024 is optimized for a variety of tasks, including answering questions, generating text, and providing information.

The model is optimized for English, Spanish, Italian, German, and Arabic.

Pre-training data includes data from multiple languages such as English, Spanish, Italian, German, French, Dutch, Portuguese, Polish, Czech, Hungarian, Swedish, Norwegian, Danish, Finnish, and many others.

Model Architecture

This is an auto-regressive model architecture. After pre-training on a large dataset, it has been fine-tuned to answer questions and generate text.

Tool use capabilities

Command R 08-2024 has been trained on a wide range of tools and capabilities. These have been trained into the model via a mixture of supervised fine-tuning and preference fine-tuning, using a specific prompt template. Deviating from this prompt template will likely reduce performance, but we encourage you to experiment and find what works best for your use case.

### Cohere-command-r-08-2024 用のサーバレス API デプロイ

概要 價格と使用条件

 Cohere Command R 08-2024 は、Azure Marketplace を通じて Cohere によって提供されています。価格と使用条件の詳細については、[価格と使用条件] タブを参照してください。

[サービスとしてのモデルの詳細について説明します。](#)

プロジェクトの選択 \*

新しいプロジェクトの作成

project45512651

① サーバレス API は以下のリージョンでご利用いただけます: [eastus, northcentralus, westus, southcentralus, eastus2, westus3, swedencentral]。

使用条件

"サブスクリプションとデプロイ" をクリックすることで、お客様は (a) 上記の各 Marketplace サービスに関する法律条項とプライバシーに関する声明に同意し、(b) サービスの利用を停止するまでの間、Microsoft より、その利用に伴う料金 (適用される税を含む) が、現在の支払い方法に対して Azure サブスクリプションと同じ請求頻度で課金または請求されることを認め、(c) Microsoft がお客様の連絡先情報および取引情報を収集・利用・分析・開示する権利を有する旨を含む。+ デtails

Azure Marketplace の使用条件

サブスクリプションとデプロイ キャンセル

[← すべてのリソースへ](#)現在のハブ  
Hub45512651

ハブの概要

すべてのプロジェクト

概要

**モデルカタログ**

モデルのベンチマーク

プロンプトカタログ

AIサービス

プレイグラウンド

チャット

アシスタント プレビュー

音声 プレビュー

共有リソース

デプロイ

接続

コンピューティング インスタンス

ユーザー

コンテンツ フィルター

**← Cohere-command-r-08-2024****概要**

タスク: チャットの

**デプロイ****説明**

Command R 08-2024 is optimized for a variety of use cases including reasoning, summarization, and question answering, with increased performance on coding, math, and reasoning.

Command R 08-2024 is an updated highly performant generative large language model. It is designed to excel across a variety of use cases including reasoning, summarization, and question answering, with increased performance on coding, math, and reasoning. The model has the capability for multilingual generation evaluated in 10 key business domains.

**価格**

paygo-inference-output-tokens:

\$0.0006 per 1000 tokens

paygo-inference-input-tokens:

\$0.00015 per 1000 tokens

paygo-inference-input-tokens:

\$0.00015 per 1000 tokens

**法的事項**[プライバシー ポリシー](#) [ライセンス契約](#)**Model Architecture**

This is an auto-regressive architecture. After pre-training on a large dataset, the model undergoes fine-tuning to learn specific tasks like coding, math, and reasoning. This approach allows it to generate text that is contextually appropriate and accurate.

**Tool use capabilities**

Command R 08-2024 has been specifically trained with conversational tool use capabilities. These have been trained into the model via a mixture of supervised fine-tuning and preference fine-tuning, using a specific prompt template. Deviating from this prompt template will likely reduce performance, but we encourage experimentation to find the best results for your use case.

**サブスクライブとデプロイ****キャンセル**

- 一部のモデルでは、デプロイ時に、「サーバーレスAPI」を使用するか「マネージドコンピューティング」を使用するかを選択できる。

The screenshot shows the 'Deployment Options' dialog box in the Azure AI studio interface. The dialog title is 'Deployment Options'. It contains two main options:

- Azure AI Content Safety を使ったサーバーレス API**: This option uses Azure AI Content Safety to filter content. It is described as using standard Azure AI Content Safety filters and is marked with a blue checkmark icon.
- Azure AI Content Safety を使用しないマネージド コンピューティング**: This option does not use Azure AI Content Safety, instead relying on user-managed hosting and model inference. It is marked with a yellow warning icon.

At the bottom right of the dialog is a 'Cancel' button. The background shows the rest of the studio interface, including a sidebar with project navigation and a main area showing training datasets and costs.

# モジュール2 モデルのデプロイ

- ・モデルカタログとは？
- ・デプロイとは？
- ・アプリからのデプロイの利用
- ・Azure OpenAI Serviceのモデルのデプロイ
- ・マネージドコンピューティングを利用するデプロイ
- ・サーバーレスAPIを利用するデプロイ
- ・まとめ

# モジュール2まとめ

- ・**モデルカタログ**では、1,600を超える、さまざまなモデルから、利用したいモデルを選び、デプロイできる
- ・デプロイされると、モデルの機能を利用するための**エンドポイントURI**、**キー**が作成される。アプリからは**エンドポイントURI**、**キー**、**デプロイ名**を使用してデプロイにアクセスできる
- ・Azure OpenAI Serviceのモデル（GPT-4等）は、デプロイの際に仮想マシンは不要
- ・多くのモデルは、仮想マシンにデプロイ（配置）される（**マネージドコンピューティング**）
- ・一部のモデルは、仮想マシンなしでのデプロイも可能（**サーバレスAPI**）
- ・一部のモデルは、マネージドコンピューティングとサーバーレスAPIの両方にに対応しており、デプロイ時にどちらを使用するかを選択できる

# モジュール3



## Azure AI Foundry でのプロンプト フローの概要

18 分 残り • モジュール • 4/8 ユニットが完了しました

プロンプトフローを使用して、Azure AI Foundry で言語モデルを利用するアプリケーションを開発する方法について学びます。

# モジュール3 プロンプトフロー

- プロンプトフローとは？
- フローの例
- フローの作成例
- フローの実行例
- まとめ

# プロンプトフローとは？

- Azure AI Foundryでは「プロンプトフロー」（prompt flow）を作成できる
- プロンプトフローでは「フロー」（オーケストレーションフロー）を開発する。フローは、複数のステップの流れを定義するもの。
- フローの中の各ステップは「ツール」と呼ばれる。
- ツールで、LLMにプロンプトを送信してコンテンツの生成を行ったり、Pythonコードを実行したりできる。
- あるツールの出力は、別のツールへの入力に渡すことができる。
- これらの仕組みを使用して、複雑なフロー（ロジック）を実装できる

# モジュール3 プロンプトフロー

- ・プロンプトフローとは？

- ・フローの例

- ・フローの作成例

- ・フローの実行例

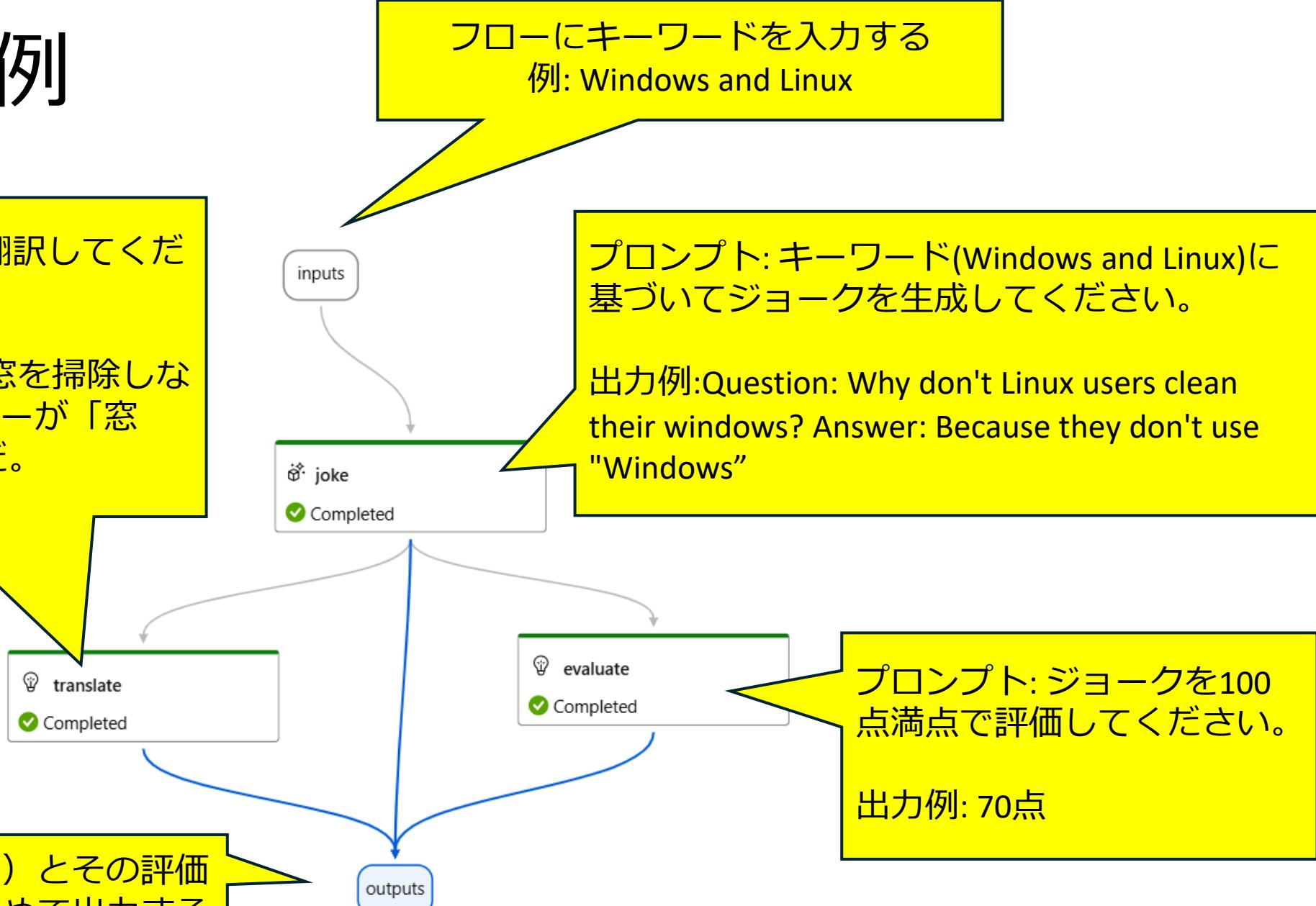
- ・まとめ

# フローの例

プロンプト: 文章を日本語に翻訳してください。

出力例: なぜLinuxユーザーは窓を掃除しないのか？それは、Linuxユーザーが「窓(Windows)」を使わないからだ。

フローにキーワードを入力する  
例: Windows and Linux



# モジュール3 プロンプトフロー

- ・プロンプトフローとは？
- ・フローの例
- ・フローの作成例
- ・フローの実行例
- ・まとめ

## オーケストレーションフローの作成、反復、デバッグ

フロー 実行

+ 作成

最新の情報に更新

アーカイブ

ビューのリセット

自分のフローのみを表示する

アーカイブ済みを含める



独自のフローをまだ作成していません

フローとは、アプリのAIロジックを実装する命令セットのことです。サンプルの複製、ローカルファイルまたは保存ファイルのインポート、またはゼロからのビルトによってフローを作成できます。

フローを使用した詳細を表示します。

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタン

プレビュー

リアルタイ

ムオーディ

オ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## 新しいフローの作成

種類別に作成

### 標準フロー

大規模な言語モデル、カスタマイズされた Python コードなどの機能を活用して、カスタマイズされたプロンプトフローを作成します。カスタムデータセットを使用してフローをテストし、エンドポイントとしてシームレスにデプロイして統合を容易にします。

### チャットフロー

標準フローに加え、このオプションでは、作成/デバッグ UI でチャット履歴のサポートとユーザーフレンドリなチャットインターフェイスが提供されます。

### 評価フロー

評価フローを作成して、出力が予想される条件と目標とどの程度一致するかを測定します。

ギャラリーを探す

All Standard flow Chat flow Evaluation flow

読み込んでいます...

インポート

ファイル共有からインポート

ローカルからアップロード

キャンセル

## 新しいフローの作成

種類別に作成

## 標準フロー

大規模な言語モデル、カスタマイズされた Python コードなどの機能を活用して、カスタマイズされたプロンプトフローを作成します。カスタムデータセットを使用してフローをテストし、エンドポイントとしてシームレスにデプロイして統合を容易にします。

**作成**

## チャットフロー

標準フローに加え、このオプションでは、作成/デバッグ UI でチャット履歴のサポートとユーザーフレンドリなチャットインターフェイスが提供されます。

**作成**

## ギャラリーを探す

**All** Standard flow Chat flow Evaluation flow

## Chat Multi-Round Q&amp;A on Your Data

Create a chatbot that uses LLM and data from your own indexed files to ground multi-round question and answering capabilities in enterprise chat scenarios.

**詳細の表示****クローン**

## Standard Q&amp;A on Your Data

Use LLM and data from your own indexed files to ground multi-round question and answering capabilities.

**詳細の表示****クローン**

## Chat Chat with Wikipedia

Create a chatbot that leverages Wikipedia data to...

## Chat Use GPT Function Calling

Learn how to use GPT function calling to extend the...

## 新しいフローの作成

フロー コード ファイルは、ワークスペース ファイル共有ストレージ内の特定のフォルダーに格納されます。このフォルダ名は、ユーザー設定に従ってカスタマイズできます。

フローを格納する場所 \*

Users/User1-45512651/promptflow

フォルダ名 \*

Flow-created-on-11-07-2024-23-16-57

**作成****キャンセル**

## 新しいフローの作成

種類別に作成

### 標準フロー

大規模な言語モデル、カスタマイズされた Python コードなどの機能を活用して、カスタマイズされたプロンプトフローを作成します。カスタムデータセットを使用してフローをテストし、エンドポイントとしてシームレスにデプロイして統合を容易にします。

**作成**

### チャットフロー

標準フローに加え、このオプションでは、作成/デバッグ UI でチャット履歴のサポートとユーザーフレンドリなチャットインターフェイスが提供されます。

**作成**

ギャラリーを探す

All Standard flow Chat flow Evaluation flow

#### Chat Multi-Round Q&A on Your Data

Create a chatbot that uses LLM and data from your own indexed files to ground multi-round question and answering capabilities in enterprise chat scenarios.

**詳細の表示**

**クローン**

#### Standard Q&A on Your Data

Use LLM and data from your own indexed files to ground multi-round question and answering capabilities.

**詳細の表示**

**クローン**

#### Chat Chat with Wikipedia

Create a chatbot that leverages Wikipedia data to

**詳細の表示**

**クローン**

#### Chat Use GPT Function Calling

Learn how to use GPT function calling to extend the

## 新しいフローの作成

フロー コード ファイルは、ワークスペース ファイル共有ストレージ内の特定のフォルダーに格納されます。このフォルダ名は、ユーザー設定に従ってカスタマイズできます。

フローを格納する場所 \* 

Users/admin/promptflow

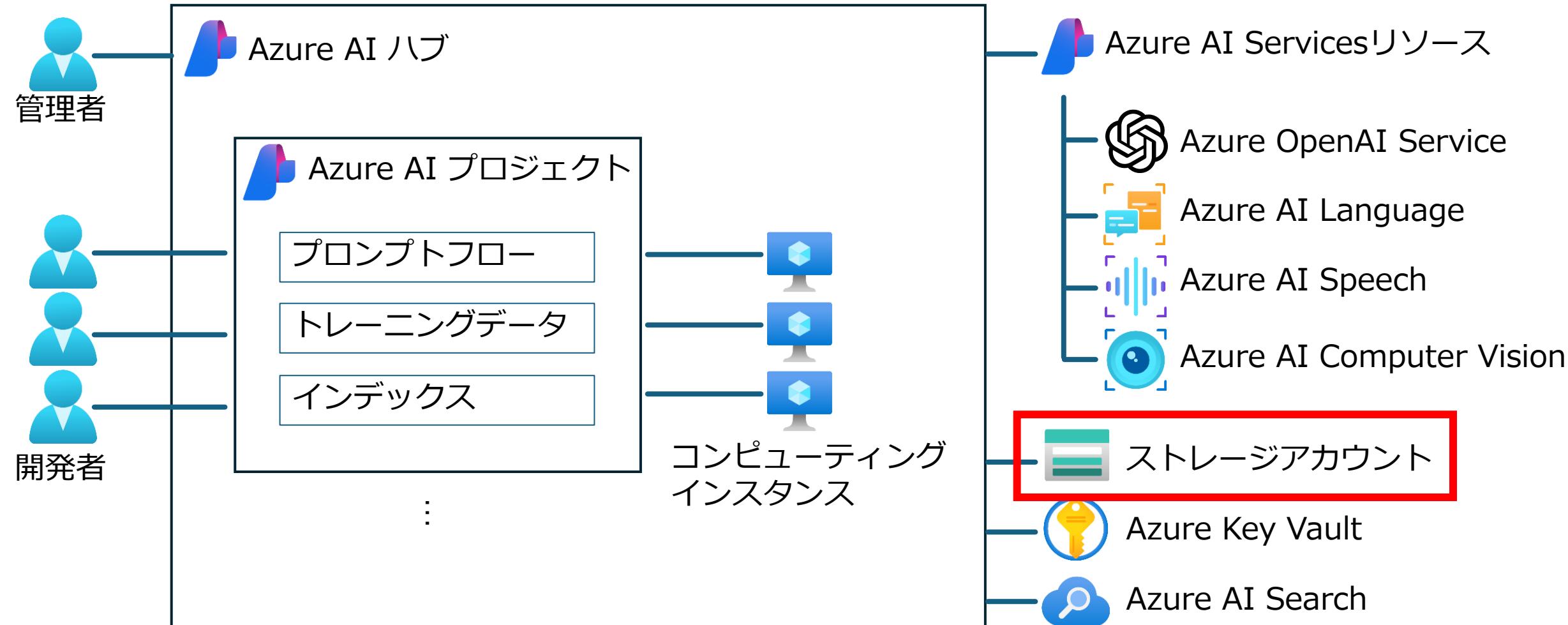
フォルダ名 \* 

Flow-created-on-11-08-2024-09-12-07

**作成**

キャンセル

# 作成した「フロー」はハブに接続されたストレージアカウントに格納される



モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス  
ト

リアルタイム

ムオーディオ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

Flow-created-on-11-08... クローン 保存 デプロイ 評価 コンピューティング セッションの開く 実行

Flow

+ LLM + Prompt + Python + その他のツール  生ファイルモード ...

▼ 入力 ①

名前	種類	値	アクション
topic	string	windows and linux	

+ 入力の追加

説明を表示する

▼ 出力 ①

名前	値	アクション
joke	\${echo.output}	

+ 出力を追加する

inputs → joke → echo → outputs

echo python

コード 参照先: echo.py

ファイル

グラフ

```

graph TD
    inputs((inputs)) --> joke[joke]
    joke --> echo[echo]
    echo --> outputs(outputs)
  
```

outputs

100% ...

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシstan  
トリアルタイ  
リアルタイムオーディ  
オ

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

### Flow-created-on-11-08...

Flow

+ LLM + Prompt + Python + その他のツール

ファイル

グラフ

joke

接続 \* ai-admin1863ai433032826210\_aoui API \* chat  
deployment\_name \* gpt-4o temperature 1 stop  
max\_tokens 256 response\_format {"type": "text"}

詳細

top\_p 1.0 presence\_penalty frequency\_penalty logit\_bias  
seed

> 関数呼び出し

プロンプト 参照先: joke.jinja2

```
1  # Prompt is a jinja2 template that generates prompt for LLM #
2
3  # system:
4
5  You are a bot can tell good jokes
6
7  # user:
8
9  A joke about {{topic}} please
10
```

inputs → joke → echo → outputs

# モジュール3 プロンプトフロー

- ・プロンプトフローとは？
- ・フローの例
- ・フローの作成例
- ・フローの実行例
- ・まとめ

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシstan  
ト  
リアルタイ

ムオーディ  
オ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

Flow-created-on-11-08... | ⏎ | ⏳ | + クローン | ⌂ 保存 | ⌂ デプロイ | ⌂ 評価 | コンピューティング セッションの開く | ⏮ 実行 | ⌂

**Flow**

+ LLM + Prompt + Python + その他のツール

生ファイルモード

説明を表示する

名前 種類 値 アクション

topic	string	windows and linux	編集 削除
-------	--------	-------------------	-------

+ 入力の追加

名前 値 アクション

joke	`\${echo.output}`	編集 削除
------	-------------------	-------

+ 出力を追加する

inputs → joke → echo → outputs

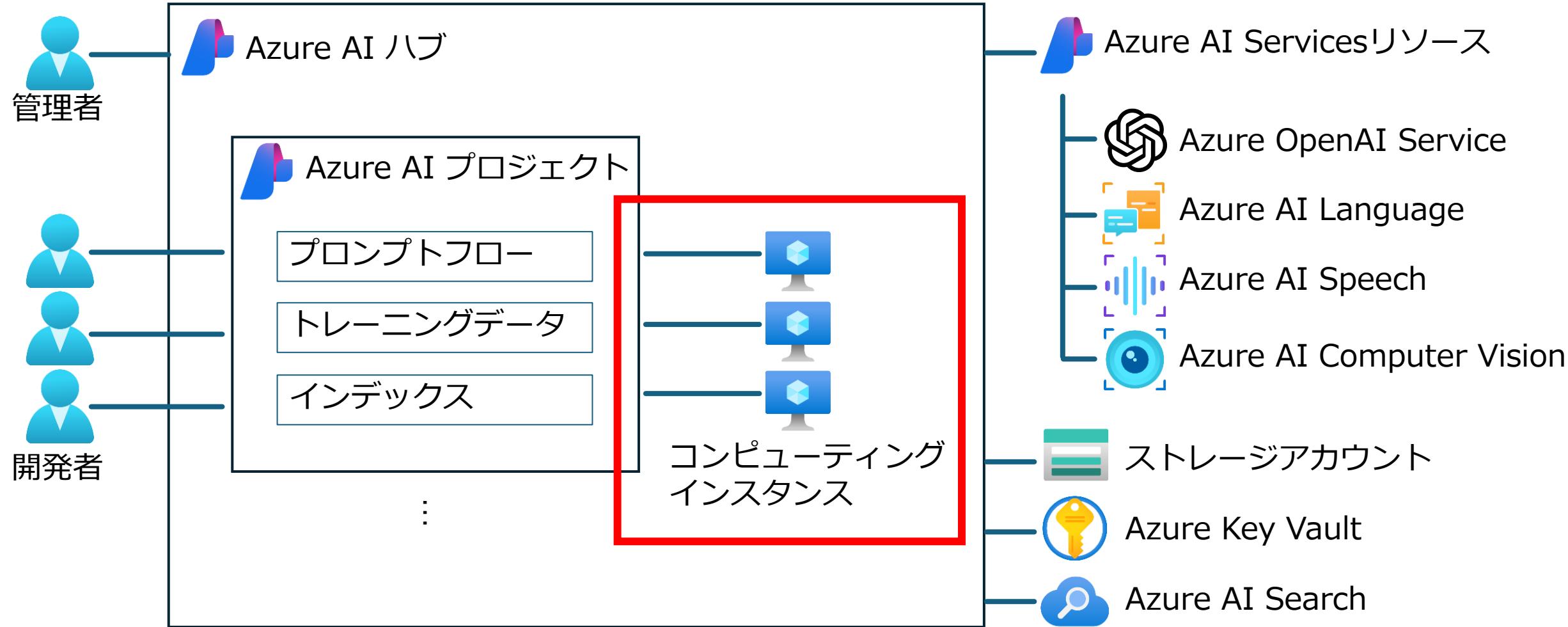
コード 参照先: echo.py

実行状況: 100% 実行

```

graph TD
    inputs((inputs)) --> joke[joke]
    joke --> echo[echo]
    echo --> outputs(outputs)
  
```

# 「フロー」はプロジェクトに動的に作成される 「コンピューティングインスタンス」で実行される



モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシstan  
ト  
リアルタイ

ムオーディ  
オ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

Flow-created-on-11-08...

フロー名: Flow-created-on-11-08...

操作ボタン: クローン, 保存, デプロイ, 評価, 実行 (実行ボタンが赤枠で囲まれている)

Flow ツールバー: + LLM, + Prompt, + Python, + その他のツール, 生ファイルモード

入力 (inputs): 名前: topic, 種類: string, 値: windows and linux

出力 (outputs): 名前: joke, 値: \${echo.output}

フロー構造:

```

graph TD
    inputs((inputs)) --> joke[joke]
    joke --> echo[echo]
    echo --> outputs(outputs)
  
```

echo コード (echo.py):

```

echo
python
  
```

コード (echo.py): 参照先: echo.py

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシstan  
ト

リアルタイ  
ム

ムオーディ  
オ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

variant\_0 実行が完了しました:

出力の表示

Flow-created-on-11-08...



\* Flow

+ LLM

+ Prompt

+ Python

+ その他のツール

生ファイルモード

joke



接続 \* ai-admin1863ai433032826210\_aoui API \* chat

deployment\_name \* gpt-4o

temperature 1

stop

max\_tokens 256

response\_format {"type": "text"}

詳細

top\_p 1.0

presence\_penalty

frequency\_penalty

logit\_bias

seed

関数呼び出し

プロンプト ① 参照先: joke.jinja2

```
1  {# Prompt is a jinja2 template that generates prompt for LLM #}
2
3  # system:
4
5  You are a bot can tell good jokes
6
7  # user:
```

ファイル

グラフ

inputs

joke

Completed

echo

Completed

100% 1:1

## 出力



出力 ログ メトリック トレース

↓ エクスポート ▾

↑ ↓ すべてのセルを開く

検索

フィルター

列

#	入力.topic	状態	joke
[ ] 0	windows and linux	✓ Completed	Sure, here's a lighthearted joke for you: Why do Linux users never need to fix their windows? Because they don't have any! 

# モジュール3 プロンプトフロー

- ・プロンプトフローとは？
- ・フローの例
- ・フローの作成例
- ・フローの実行例
- ・まとめ

# モジュール3 まとめ

- ・**プロンプトフロー**では「フロー」（オーケストレーションフロー）を開発する。
- ・フローの中の各ステップは「ツール」と呼ばれる。
- ・ツールでは、LLMにプロンプトを送信してコンテンツの生成を行ったり、Pythonコードを実行したりできる。
- ・あるツールの出力は、別のツールへの入力に渡すことができる。
- ・これらの仕組みを使用して、複雑なフロー（ロジック）を実装できる
- ・「フロー」は「**コンピューティングインスタンス**」で実行される

# モジュール4



## Azure AI Foundry を使用して独自のデータを使用する

4 分 残り • モジュール • 5/7 ユニットが完了しました

言語モデルはユーザーと連携して、提案を提供したり、コンテンツを生成したり、意思決定を支援したりできます。特定のデータソースから確実に情報を取得するために、Azure AI Foundryで独自のデータを追加することができます。

# モジュール4 独自のデータを使用する

- RAGアーキテクチャとは？
- Azure OpenAI on your dataとは？
- Azure AI Searchとは？
- 「埋め込み」モデルとは？
- Azure OpenAI on your dataの設定例
- まとめ

# RAGアーキテクチャとは？

- Retrieval-Augmented Generation
- 検索拡張生成（検索によって強化された生成）
- 取得拡張生成 とも
- 大規模言語モデル（GPT等）に、外部のデータを接続するアーキテクチャ
- 大規模言語モデルは必要に応じて外部データを検索し、回答に必要なデータを取得してから、回答の生成を行う
- 例
  - 企業が保有している内部資料をGPTに接続して、GPTがその企業の特有の情報に基づき回答できるようにする

# モジュール4 独自のデータを使用する

- RAGアーキテクチャとは？
- Azure OpenAI on your dataとは？
- Azure AI Searchとは？
- 「埋め込み」 モデルとは？
- Azure OpenAI on your dataの設定例
- まとめ

# Azure OpenAI **on your data**とは？

- Azure OpenAI Serviceにおいて、**RAGアーキテクチャ**を簡単に実装することができる仕組み。
- Azure AI Servicesでon your dataの設定を行うことができる。
  - データのアップロード
  - インデックスの作成
- Azure AI Servicesのチャットプレイグラウンドから利用できる。
- また独自のアプリ経由でも利用できる。

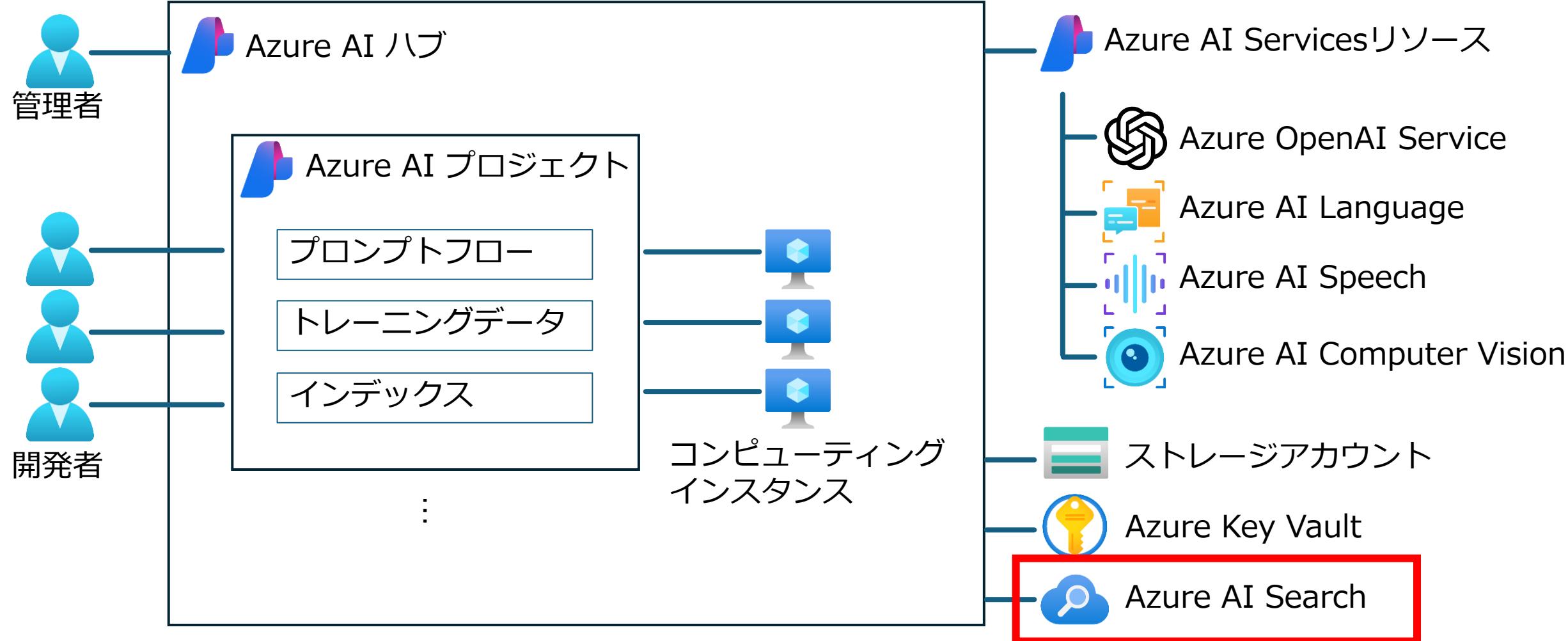
# モジュール4 独自のデータを使用する

- RAGアーキテクチャとは？
- Azure OpenAI on your dataとは？
- Azure AI Searchとは？
- 「埋め込み」モデルとは？
- Azure OpenAI on your dataの設定例
- まとめ

# Azure AI Searchとは？

- Azure AI Searchは、 Azureの検索サービス。
- たとえば大量のPDF文書に対し、 クエリ（検索）を実行して、 特定のキーワードが含まれるPDF文書をすばやく発見できる。

# Azure AI Search



# モジュール4 独自のデータを使用する

- RAGアーキテクチャとは？
- Azure OpenAI on your dataとは？
- Azure AI Searchとは？
- 「埋め込み」 モデルとは？
- Azure OpenAI on your dataの設定例
- まとめ

# 「埋め込み」モデルとは？

- ・「埋め込み」モデルは、テキストをベクトル化するためのモデル。
- ・2つのベクトルを比較することで、2つのテキストの「意味的な類似度」を判定できる。
- ・Azure AI Searchのベクトル検索機能により「ユーザーが指定した文字列（質問など）に意味的な類似度が高い文書」を探し出すことが可能となる。

# モジュール4 独自のデータを使用する

- RAGアーキテクチャとは？
- Azure OpenAI on your dataとは？
- Azure AI Searchとは？
- 「埋め込み」モデルとは？
- Azure OpenAI on your dataの設定例
- まとめ

# Azure OpenAI on your dataの設定例

- ・ハブ、プロジェクトをAzure AI Search付きで作成
  - ・または、すでに作成済みのハブに後からAzure AI Searchを追加
- ・埋め込みモデル（text-embedding-ada-002など）をデプロイ
- ・GPTモデル（gpt-4o等）モデルをデプロイ
- ・「データ」でPDFファイルをアップロード
- ・「インデックス」を作成
- ・「チャット」プレイグラウンドの「データを追加する」で「インデックス」を選択

## プロジェクトの作成

プロジェクトの詳細

ハブの作成

レビューして終了

## プロジェクトのハブを作成する

ハブは、チームがプロジェクト作業、モデルエンドポイント、コンピューティング、接続、セキュリティ設定を共有するためのコラボレーション環境です。 [詳しく見る](#)

セキュリティまたはハブの [依存リソース](#) をカスタマイズする必要がありますか? [Azure portal にアクセスする](#)

ハブ名 \*

Hub45531347

サブスクリプション \* ⓘ

[新しいサブスクリプションの作成](#)

MOCOAI-lod49253180

リソース グループ \* ⓘ

[新しいリソース グループの作成](#)

ResourceGroup1

場所 \* ⓘ

East US

[選択に関するヘルプ](#)

Azure AI サービスまたは Azure OpenAI を接続する \* ⓘ

[新しい AI サービス の作成](#)

(新しい) ai-Hub45531347

Azure AI 検索に接続する ⓘ

[新しい AI Search の作成](#)

接続をスキップする

戻る

次へ

プロジェクトを作成する

キャンセル

## プロジェクトの作成

プロジェクトの詳細

ハブの作成

レビューして終了

## プロジェクトのハブを作成する

ハブは、チームがプロジェクト作業、モデルエンドポイント、コンピューティング、接続、セキュリティ設定を共有するためのコラボレーション環境です。 [詳しく見る](#)

セキュリティまたはハブの [依存リソース](#) をカスタマイズする必要がありますか? [Azure portal にアクセスする](#)

ハブ名 \*

Hub45531347

サブスクリプション \* ⓘ

MOCOAI-lod49253180

[新しいサブスクリプションの作成](#)

リソース グループ \* ⓘ

ResourceGroup1

独自のデータにインデックスを作成するには、標準の Azure AI 検索が必要です。今すぐ作成しない場合や、カスタム設定を使用する場合は、後で独自の設定を作成し、プロジェクト設定でリンクすることができます。

場所 \* ⓘ

East US

[選択に関するヘルプ](#)

名前

aisearch923879234

[詳しく見る](#)

作成

キャンセル

Azure AI サービスまたは Azure OpenAI を接続する \* ⓘ

(新しい) ai-Hub45531347

[新しい AI Search の作成](#)

Azure AI 検索に接続する ⓘ

接続をスキップする

戻る

次へ

プロジェクトを作成する

キャンセル

## プロジェクトの作成

## ホーム

概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

Azure OpenAI

AI サービス

管理

すべてのリソース

モデル クォータ

VM クォータ

プロジェクトの詳細

ハブの作成

③ レビューして終了

## レビューして終了

必要な依存関係と共に、次のリソースが自動的に作成されます。最初のハブとプロジェクトの作成が完了するまで数分かかる場合があります。[ハブと依存関係に関する詳細情報](#).

## リソース

## タイプ

🕒 Project45531347

AI プロジェクト ⓘ

🕒 Hub45531347

AI ハブ ⓘ

🕒 ai-hub45531347046600165401

AI サービス ⓘ

🕒 aisearch923879234046600165401

AI Search ⓘ

🕒 sthub4553134046600165401

ストレージ アカウント ⓘ

🕒 kv-hub45531046600165401

キー コンテナー ⓘ

🕒 リソースを作成しています...

プロジェクトを作成する

キャンセル



ホーム > Azure AI services

Azure AI services | AI Search  ...

## Azure AI services

Azure AI Search を後から追加する場合は、まずは Azure portal で Azure AI Search リソースを作成する。

Azure AI Searchを後から追加する場合は、まずはAzure portalでAzure AI Searchリソースを作成する。

[← すべてのリソースへ](#)

現在のハブ  
admin-1863\_ai

### ハブの概要

[すべてのプロジェクト](#)

#### 概要

[モデル カタログ](#)[モデルのベンチマーク](#)[プロンプト カタログ](#)[AI サービス](#)

#### プレイグラウンド

[チャット](#)[アシスタント プレビュー](#)[音声 プレビュー](#)

#### 共有リソース

[デプロイ](#)[接続](#)[コンピューティング インスタンス](#)[ユーザー](#)[コンテンツ フィルター](#)

# admin-1863\_ai

# ハブ

### プロジェクト 1

[すべて表示 →](#)

リソース名	名前	作成日
admin-4961	admin-4961	Nov 7, 2024 10:22 AM

[+ 新しいプロジェクト](#)

### 説明

ハブの説明をここに追加してください

### 接続されたリソース 4

[すべて表示 →](#)

名前	種類
ai-admin1863ai433032826210_aoai	Azure OpenAI
ai-admin1863ai433032826210	AI Services
admin-4961/workspaceartifactstore	Azure Blob Storage <a href="#">プレビュー</a>
admin-4961/workspaceblobstore	Azure Blob Storage <a href="#">プレビュー</a>

[+ 新しい接続](#)

# 新しい接続

### ハブのプロパティ

#### 名前

admin-1863\_ai

#### 場所

eastus

#### サブスクリプション



#### リソース グループ

rg-admin-3411\_ai

[Azure Portal で管理する](#)[API エンドポイントとキーを取得する](#)[サブスクリプション クオータを表示する](#)[ハブの削除](#)

### ユーザー 2

[すべて表示 →](#)

admin-1863\_ai

Azure AI Administrator

グループとアプリケーションの  
アクセス許可  
+ 2

[Azure Portal で管理する](#)[+ 新しいユーザー](#)

## 外部資産への接続を追加する

接続するリソースの種類を選択してください。

## Azure AI



Azure OpenAI Service

AI モデル



Azure AI サービス

AI モデル



Azure AI 検索 プレビュー

検索とインデックス作成



サーバーレス モデル プレビュー

AI モデル



Azure AI Content Safety プレビュー

責任ある AI



Azure AI 音声 プレビュー

音声

## データ

## その他のリソースの種類



Serp

検索



Open AI

AI モデル



API キー

カスタム



カスタム キー

カスタム

閉じる

すべて表示 →

Azure AI Administrator

Azure Portal で管理する

## Azure AI Search リソースを接続する

[← 資産の種類の選択に戻る](#) リソースの参照  手動で入力する リソースを検索してください

(1) 個のリソースを表示しています

## 名前

[aisearch9272432](#)

## リソース グループ

rg-admin-3411\_ai

## 場所

eastus

## SKU

free

## サブスクリプション

ME-MngEnvMCAP959990-hiryamada-1

## セマンティック検索

disabled

[接続を追加する](#)

## 認証

 API key

① ハブに、このリソースへのアクセスが許可されます。プロジェクトまたはハブにアクセスできるすべてのユーザーがこのリソースを使用できるようになります。

[すべて表示 →](#)

Azure AI Administrator

[Azure Portal で管理する](#)[閉じる](#)

## Azure AI Search リソースを接続する

[← 資産の種類の選択に戻る](#) リソースの参照  手動で入力する リソースを検索してください

(1) 個のリソースを表示しています

## 名前

[aisearch9272432](#)

## リソース グループ

rg-admin-3411\_ai

接続済み



## 場所

eastus

## SKU

free

## サブスクリプション

## セマンティック検索

disabled

## 認証

API key

① ハブに、このリソースへのアクセスが許可されます。プロジェクトまたはハブにアクセスできるすべてのユーザーがこのリソースを使用できるようになります。

[すべて表示 →](#)

Azure AI Administrator

[Azure Portal で管理する](#)[閉じる](#)

[admin-1863\\_ai](#)

## プロジェクト 1

[すべて表示 →](#)

リソース名	名前	作成日
admin-4961	admin-4961	Nov 7, 2024 10:22 AM

[+ 新しいプロジェクト](#)

## 説明

ハブの説明をここに追加してください

## ハブのプロパティ

名前	場所
admin-1863_ai	eastus
サブスクリプション	リソース グループ
ME-MngEnvMCAP959990-hiryamada-1	rg-admin-3411_ai

[Azure Portal で管理する](#)[API エンドポイントとキーを取得する](#)[サブスクリプション クオータを表示する](#)[ハブの削除](#)

## 接続されたリソース 4

[すべて表示 →](#)

名前	種類
ai-admin1863ai433032826210_aoai	Azure OpenAI
ai-admin1863ai433032826210	AI Services
admin-4961/workspaceartifactstore	Azure Blob Storage <a href="#">プレビュー</a>
admin-4961/workspaceblobstore	Azure Blob Storage <a href="#">プレビュー</a>

[+ 新しい接続](#)

Azure AI Searchがハブに追加された

## ユーザー 2

[すべて表示 →](#)

アクセス許可	<a href="#">+ 2</a>	<a href="#">Azure Portal で管理する</a>
--------	---------------------	------------------------------------

[+ 新しいユーザー](#)

← すべてのリソースへ

現在のハブ  
admin-1863\_ai

**ハブの概要**

すべてのプロジェクト

概要

- モデル カタログ
- モデルのベンチマーク
- プロンプト カタログ
- AI サービス

プレイグラウンド

- チャット
- アシスタント プレビュー
- 音声 プレビュー

共有リソース

- デプロイ
- 接続
- コンピューティング インスタンス
- ユーザー
- コンテンツ フィルター

[← すべてのリソースへ](#)現在のハブ  
admin-1863\_ai

ハブの概要

すべてのプロジェクト

概要

**モデル カタログ**

モデルのベンチマーク

プロンプト カタログ

AI サービス

プレイグラウンド

チャット

アシスタント プレビュー

音声 プレビュー

共有リソース

デプロイ

接続

コンピューティング インスタンス

ユーザー

コンテンツ フィルター

# カスタム AI ソリューションを構築するための適切なモデルを見つける

すべてのフィルター × コレクション 業界 デプロイ オプション

推論タスク タスクの微調整 ライセンス

text-embedding

モデル 680

text-embedding-3-large 埋め込み

text-embedding-ada-002 埋め込み

text-embedding-3-small 埋め込み

Cohere-embed-v3-multilingual 埋め込み

Cohere-embed-v3-english 埋め込み

Deci-DeciCoder-1b テキスト生成

CodeLlama-13b-hf テキスト生成

tenyx-llama3-tenyxchat-70b テキスト生成

gradientai-llama-3-8b-instant... テキスト生成

cognitivecomputations-do... テキスト生成

Nemotron-3-8B-Chat-RLHF テキスト生成

Nemotron-3-8B-Chat-SFT テキスト生成

フィルター条件:

非表示

## コレクション

Azure AI によるキュレーション 91

Azure OpenAI 9

Microsoft 7

Meta 24

Mistral 4

NVIDIA 5

AI21 Labs

Deci AI 4

Nixtla

JAIS

Cohere 4

Databricks 2

Snowflake 1

Hugging Face 589

SDAIA

Paige AI

表示を減らす

## デプロイ オプション

マネージド コンピューティング 669

サーバーレス API 14

## 業界

ヘルス + ライフ サイエンス

← すべてのリソースへ

現在のハブ admin-1863\_ai

ハブの概要

すべてのプロジェクト

概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プレイグラウンド

チャット

アシスタント プレビュー

音声 プレビュー

共有リソース

デプロイ

接続

コンピューティング インスタンス

ユーザー

コンテンツ フィルター

## ← text-embedding-ada-002

概要

タスク: 埋め込み

デプロイ (赤枠)

更新

微調整

説明

This model can be deployed for inference, specifically for embeddings, but cannot be finetuned.

Model family: GPT-3

text-embedding-ada-002 outperforms all the earlier embedding models on text search, code search, and sentence similarity tasks and gets comparable performance on text classification. Embeddings are numerical representations of concepts converted to number sequences, which make it easy for computers to understand the relationships between those concepts.

Learn more at <https://learn.microsoft.com/en-us/azure/cognitive-services/openai/concepts/models#embeddings-models>

モデル ID

azureml://registries/azure-openai/models/text-embedding-ada-002/versions/2

← すべてのリソースへ

現在のハブ admin-1863\_ai

ハブの概要

すべてのプロジェクト

概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プレイグラウンド

チャット

アシスタント プレビュー

音声 プレビュー

共有リソース

デプロイ

接続

コンピューティング インスタンス

ユーザー

コンテンツ フィルター

## ← text-embedding-ada-002

概要

タスク: 埋め込み

▷ デプロイ

説明

This model can generate embeddings, b

Model family: G

text-embedding-ada-002 is used for search, code search, and on text classification. It has been converted to number relationships between words.

Learn more at <https://services/openai/concept>

### モデル text-embedding-ada-002 をデプロイする

デプロイ名 \*

text-embedding-ada-002

カスタマイズ

デプロイの詳細

モデル バージョン	2	デプロイの種類	Standard
AI ハブ	admin-1863_ai	接続されている AI リソース	ai-admin1863ai433032826210_aoai
容量	120K 1 分あたりのトークン数 (TPM)	認証の種類	キー
コンテンツの安全性	DefaultV2	リソースの場所	East US

デプロイ 取り消し

← すべてのリソースへ 現在のハブ  
admin-1863\_ai 

ハブの概要

すべてのプロジェクト

概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プレイグラウンド

チャット

アシスタント プレビュー

音声 プレビュー

共有リソース

デプロイ

接続

コンピューティング インスタンス

ユーザー

コンテンツ フィルター

## モデル デプロイ

プライベート API キーとエンドポイント URI (Uniform Resource Identifier) を使用してモデルをデプロイします。

モデル デプロイ アプリのデプロイ サービスのデプロイ+ モデルのデプロイ 最新の情報に更新 編集 プレイグラウンドで開く ビューのリセット 列

名前	モデル名	モデル バージョン	状態	モデルの廃止日	コンテンツ フィ.
ai-admin1863ai433032826210_aoai <a>Azure OpenAI</a>					
gpt-4o	gpt-4o	2024-05-13	成功		DefaultV2
text-embedding-ada-002	text-embedding-ada-002	2	成功		DefaultV2

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシstan  
ト  
リアルタイ

ムオーディ  
オ  
プレビュ

イメージ

入力候補

音声  
プレビュ

ツール

コード  
プレビュ

プロンプトフロー

評価  
プレビュ

微調整  
プレビュ

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## データを接続、作成、または管理する

+ 新しいデータ

アーカイブ

アーカイブされたデータを表示する

最新の情報に更新

ビューのリセット



### モデルが参照できるデータを追加する

ユースケースに必要なデータへのアクセス権をモデルに付与します。データはサポートされている種類のいずれかに設定する必要があります: ファイル (uri\_file) は保存されている任意の形式の単一ファイルを指し、フォルダー (uri\_folder) はファイルや非構造化データのフォルダーを指します。

[データ型とデータ接続に関する詳細情報](#)

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス

プレビュー

リアルタイム

ムオーディオ

プレビュー

イメージ

入力候補

音声

プレビュー

ツール

コード

プレビュー

プロンプト フロー

評価

プレビュー

微調整

プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## データを追加する

① データの選択

② 名前と完了日

## データを選択する

データソース \*

ファイル/フォルダーをアップロードする

+ 新しい接続

ファイルまたはフォルダーをアップロードする

↑ ファイルのアップロード

↗ フォルダーのアップロード



アップロードされたファイルはありません

開始するには、上の [ファイルまたはフォルダーのアップロード] メニューを選択します。

次へ

作成

キャンセル



開く

← → ▼ ↑📁 > ダウンロード > brochures

brochuresの検索

整理 ▼ 新しいフォルダー☰ ? 名前

更新日時

種類

▼ 今日

<input checked="" type="checkbox"/> Las Vegas Brochure.pdf	2024/11/08 8:40	Adobe Acrobat Docu...
<input checked="" type="checkbox"/> London Brochure.pdf	2024/11/08 8:40	Adobe Acrobat Docu...
<input checked="" type="checkbox"/> Margies Travel Company Info.pdf	2024/11/08 8:40	Adobe Acrobat Docu...
<input checked="" type="checkbox"/> New York Brochure.pdf	2024/11/08 8:40	Adobe Acrobat Docu...
<input checked="" type="checkbox"/> San Francisco Brochure.pdf	2024/11/08 8:40	Adobe Acrobat Docu...
<input checked="" type="checkbox"/> Dubai Brochure.pdf	2024/11/08 8:40	Adobe Acrobat Docu...

ファイル名(N): "Las Vegas Brochure.pdf" "London Brochure.pdf" "Margies Travel Company Info.pdf" すべてのファイル (\*.\*)

モバイルからアップロード

開く(O)

キャンセル

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

 アシスタンス プレビュー

リアルタイム

 ムオーディオ プレビュー

オ

イメージ

入力候補

 音声 プレビュー

ツール

 コード プレビュー

プロンプト フロー

 評価 プレビュー 微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## データを追加する

① データの選択

② 名前と完了日

## データを選択する

データソース \*

ファイル/フォルダーをアップロードする

+ 新しい接続

 ファイルまたはフォルダーをアップロードする 既に存在する場合は上書きする

## リストのアップロード

Las Vegas Brochure.pdf ✓ 556.56 KB/556.56 KB ...London Brochure.pdf ✓ 440.07 KB/440.07 KB ...Margies Travel Company Info.pdf ✓ 344.01 KB/344.01 KB ...New York Brochure.pdf ✓ 373.45 KB/373.45 KB ...San Francisco Brochure.pdf ✓ 376.3 KB/376.3 KB ...Dubai Brochure.pdf ✓ 379.51 KB/379.51 KB ...

次へ

作成

キャンセル

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンスト プレビュー

リアルタイ

ムオーディオ プレビュー

オ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## データを追加する

データの選択

② 名前と完了日

## 名前と完了日

データ名 \*

brochure

戻る

作成

キャンセル

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

 アシスタント プレビュー

リアルタイム

 ムオーディオ プレビュー

イメージ

入力候補

 音声 プレビュー

ツール

 コード プレビュー

プロンプトフロー

 評価 プレビュー 微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## brochure:1

### データの詳細

**名前**

brochure

**現在のバージョン**

バージョン: 1 (最新)

+ 新しいバージョン

**最新バージョン**

1

**種類**

フォルダー

**ファイル数**

6

**合計サイズ**

2.412 MiB

**作成日**

Nov 8, 2024, 8:42:14 AM

**変更日**

Nov 8, 2024, 8:42:14 AM

**作成者**

SA System Administrator

**タグ**

+ タグの追加

### データリンク

**接続名**

workspaceblobstore

**データ接続 URI**

azureml://subscriptions/1020c688-b80c-4b5c-a0aa-67a3dc7f99...

«

- 13bfbb5d-4944-42e1-a091-6fa2dd2cdceb-azureml-blo...
- UI
- 2024-11-07\_233759\_UTC
  - Dubai Brochure.pdf
  - Las Vegas Brochure.pdf
  - London Brochure.pdf
  - Margies Travel Company Info.pdf
  - New York Brochure.pdf
  - San Francisco Brochure.pdf

利用可能なレビューはありません

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシstan  
ト

リアルタイ

ムオーディ  
オ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス 選択

デプロイ

コンテンツ フィルタ

## インデックスを作成して生成 AI の応答をカスタマイズする

+ 新しいインデックス

最新の情報に更新

ビューのリセット



### インデックスの作成

インデックスは、データを効率的に検索できるデータストアです。これは、取得拡張生成 (RAG) で非常に便利です。ベクトル (埋め込みモデルを使用して数値シーケンスに変換されたテキスト/データ) を作成することで、LLM 向けにインデックスを最適化できます。

お困りですか? [ドキュメントを表示する](#)

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイラウンド ^

チャット

 アシスタン  
ト

リアルタイム

 ムオーディ  
オ

イメージ

入力候補

 音声 プレビュー

ツール ^

 </> コード プレビュー

プロンプトフロー

 評価 プレビュー 微調整 プレビュー

コンポーネント ^

データ

インデックス

デプロイ

コンテンツ フィルタ

## ベクトルインデックスを作成する レビュー

1 Source location

2 Index configuration

3 検索設定

4 Review and finish

Choose to location of the input data source that you want to index.

データソース \* ⓘ Azure AI Studio のデータ

名前	型	変更日
<input checked="" type="checkbox"/> brochure	UriFolder	Nov 8, 2024 8:42 AM

検索 フィルター 列

« < 1 ページ中の 1 ページ目 > » 25/ページ ▾

ⓘ データのインデックスを作成するには、Azure AI Search リソースと Azure Open AI 接続が必要です。新しい Azure AI Search リソースを作成する [\[+\] インデックス](#) の作成中に選択する接続を作成します。

次へ

ベクターインデックスを作成する

キャンセル

 モデルのベンチマーク プロンプト カタログ AI サービス**プロジェクトのプレイ**  ラウンド チャット アシstan

プレビュー

 リアルタイ ムオーディ

オ

 イメージ 入力候補 音声

プレビュ

**ツール**  コード

プレビュ

 プロンプトフロー 評価

プレビュ

 微調整

プレビュ

**コンポーネント**  データ インデックス デプロイ コンテンツ フィルタ

一

## ベクトルインデックスを作成する

 Source location Index configuration 検索設定 Review and finish

### Index settings

Configure your index

#### インデックスストレージ\*

Azure AI Search

#### Azure AI Search Service の選択\*

aisearch9272432

[新しい Azure AI Search リソースを作成する !\[\]\(dc556b23343d41f945574daaaa4d5a27\_img.jpg\)](#)

#### ベクターインデックス\*

brochure-index

#### 仮想マシン\*

 自動選択 推奨オプションから選択 すべてのオプションから選択

仮想マシンを選択すると、追加のコストが発生します。

[戻る](#)[次へ](#)[ベクターインデックスを作成する](#)[キャンセル](#)

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシスタン  
トリアルタイ  
ムムオーディ  
オ

イメージ

入力候補

音声 プレビュー

ツール

</> コード プレビュー

プロンプトフロー

評価 プレビュー微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

-

ベクトルインデックスを作成する レビュー

Source location

Index configuration

検索設定

Review and finish

## 検索設定の設定

ハイブリッド取得とセマンティック ランク付け (ハイブリッド + セマンティック) を組み合わせると、生成 AI アプリケーションに対して最も正確な検索結果が得られます。ベクトルインデックスを生成するには、埋め込みモデルが必要です。

## ベクトル設定

 ベクトル検索をこの検索リソースに追加する

## Azure OpenAI 接続 \*

ai-admin1863ai433032826210\_aoai

ⓘ このリソースには埋め込みモデルが必要です。まだお持ちでない場合は、[text-embedding-ada-002 \(バージョン 2\)](#) が展開されます。ベクター埋め込みを使用すると、アカウントの使用が生じます。 [View Azure OpenAI Service pricing](#)

戻る

次へ

ベクターインデックスを作成する

キャンセル

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド

チャット

アシスタン  
ト

リアルタイ

ムオーディ

オ

イメージ

入力候補

音声 プレビュー

ツール

</> コード プレビュー

プロンプトフロー

評価 プレビュー微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

ベクトルインデックスを作成する レビュー

Source location

## Review and finish

Review the configurations you set for your index

Index configuration

## ベクターインデックス

brochure-index

検索設定

## インデックスストレージ

Azure AI Search

Review and finish

## Azure AI Search 接続

aisearch9272432

## ベクトル設定を含める

はい

## コンピューティング

サーバーレスコンピューティング(自動選択)

戻る

ベクターインデックスを作成する

キャンセル

モデルのベンチマーク

ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド ^

チャット

アシスタンス  
ト  
リアルタイ

ムオーディ  
オ  
イメージ

入力候補

音声 プレビュー

ツール ^

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント ^

データ

インデックス

デプロイ

コンテンツ フィルタ

## brochure-index

### 状態

Running

最新の情報に更新

#### バージョン

-

モデルを使用して埋め込む

いいえ

#### ソースの種類

Azure AI On Your Data

ベクトルストア

-

#### インデックス作成時間の合計

-

コンピューティング

-

#### 作成日

Nov 8, 2024, 8:45:24 AM

#### 作成者

System Administrator

ジョブの詳細

テストデータ

### ソースデータ

名前	種類	サイズ
Dubai Brochure.pdf	.pdf	379.51 KB
Las Vegas Brochure.pdf	.pdf	556.56 KB
London Brochure.pdf	.pdf	440.07 KB
Margies Travel Company Info....	.pdf	344.01 KB

### 状態

手順 1 / 3

コンピューティング リソースの割り当て - 処理中

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド ^

チャット

アシスタンス  
ト リアルタイ

ムオーディオ

イメージ

入力候補

音声 プレビュー

ツール ^

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント ^

データ

インデックス

デプロイ

コンテンツ フィルタ

## brochure-index

### 状態

Completed

最新の情報に更新

### バージョン

-

### モデルを使用して埋め込む

いいえ

### ソースの種類

Azure AI On Your Data

### ベクトルストア

-

### インデックス作成時間の合計

9 分

### コンピューティング

サーバーレス コンピューティング

### 作成日

Nov 8, 2024, 8:45:24 AM

### 作成者

System Administrator

ジョブの詳細

テストデータ

### ソースデータ

名前	種類	サイズ
Dubai Brochure.pdf	.pdf	379.51 KB
Las Vegas Brochure.pdf	.pdf	556.56 KB
London Brochure.pdf	.pdf	440.07 KB
Margies Travel Company Info....	.pdf	344.01 KB

### 状態

手順 1 / 3

解読とチャンク - 完了



手順 2 / 3

Azure AI 検索インデックスの作成 - 完了



手順 3 / 3

インデックスの登録 - 完了



## チャットプレイグラウンド

on your dataなしの場合、質問に対して、大規模言語モデルがもともと持っている知識に基づいて、一般的な回答を返す

十 セクションの追加 ▼

## ✓ データを追加する プレビュー

自分のデータについて質問します。データは Azure サブスクリプションに安全に格納されます。データの保護方法について説明します。

## 使用可能なプロジェクトインデックスの選択\*

## 使用可能なプロジェクトインデックスの選択

#### 十 新しいデータソースを追加する

## › パラメーター

→ インポート → エクスポート プロンプトのサンプル ...

ヤツトのクリア

チャット機能

JSON の表示

Published online in New Mathematics and Natural Science, 2020, 2(1)

New York City offers a wide range of accommodations to suit various preferences and budgets. Here are some popular neighborhoods and hotels to consider:

## 1. Manhattan

- **Times Square/Theater District:** This area is vibrant and central, close to many major attractions. Consider hotels like the Marriott Marquis, The Knickerbocker, or The Times Square EDITION.
  - **Midtown:** Convenient for sightseeing, shopping, and dining. Options include The Plaza, The Langham, or The Peninsula.
  - **Upper West Side:** Quieter and more residential, close to Central Park and museums. The Lucerne Hotel and Hotel Beacon are good choices.
  - **Upper East Side:** Luxurious and close to Central Park and upscale shopping. Consider The Carlyle, The Mark, or The Surrey.
  - **SoHo/Tribeca:** Trendy neighborhoods with great shopping and dining. The James New York, The Roxy Hotel, or The Greenwich Hotel are excellent options.

ここにユーザー クエリを入力します。(Shift + Enter で新しい行)

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド ^

チャット

アシstan  
ト プレビュ

リアルタイ  
ム プレビュ

オーディオ プレビュ

イメージ プレビュ

入力候補 プレビュ

音声 プレビュ

ツール ^

コード プレビュ

プロンプト フロー

評価 プレビュ

微調整 プレビュ

コンポーネント ^

データ

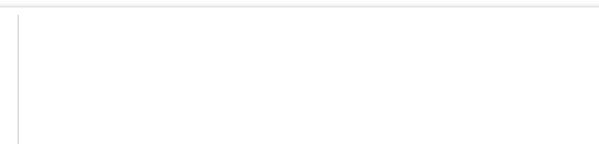
インデックス

デプロイ

コンテンツ フィルタ

# チャットプレイグラウンド

コードの表示 プロンプト フロー 評価 デプロイ インポート エクスポート プロンプトのサンプル ...



保存



+ セクションの追加

▽ データを追加する プレビュ

自分のデータについて質問します。データは Azure サブスクリプションに安全に格納されます。[データの保護方法について説明します。](#)

使用可能なプロジェクトインデックスの選択 \*

使用可能なプロジェクトインデックスの  
選択

brochure-index

> パラメーター

チャットのクリア

チャット機能



JSON の表示



## チャットを開始する

チャットのプレイグラウンドで、見る、聞く、話すことができるようになりました。チャットウィンドウでマイクを選択し、音声入力を開始して、テキストを手動で入力せずにモデルにプロンプトを表示します。スピーカー アイコンを選択して、モデルの出力を聞くこともできます。

ここにユーザー クエリを入力します。(Shift + Enter で新しい行)

21/128000 送信するトークン

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## チャットプレイグラウンド

on your dataありの場合、質問に対して、関連するPDF文書を検索し、その情報に基づいて回答を返す

+ セクションの追加

① ソースデータが追加されている場合、一部のショットの例は使用されません。以前の例はすべてクリアされました。データを削除すると、例を追加できます。

▼ データを追加する プレビュー

自分のデータソースへの分析情報を取得します。データは Azure サブスクリプションに安全に保存されます。[データの保護方法について説明します。](#)

インデックス:

[brochure-index](#)

検索の種類:

ハイブリッド (ベクトル + キーワード) [さまで](#)

[また検索の種類に関する詳細情報](#)

[詳細設定](#)

ロード インポート エクスポート プロンプトのサンプル ...

チャットのクリア

チャット機能

JSON の表示

I'd like to trip to New York. Where should I stay?

...

Margie's Travel offers several accommodation options in New York:

1. **The Manhattan Hotel:** Stay in the heart of New York City, within an easy walk to Times Square and Broadway ^1^.
2. **The Grand Central Hotel:** Comfortable mid-town hotel, close to Grand Central Station, the Chrysler Building, and the Empire State Building ^1^.
3. **The Park Hotel:** Luxurious accommodation in upper Manhattan, with views of Central Park ^1^.

To book your trip to New York, you can visit [Margie's Travel website](#) ^1^.

1 参照

ここにユーザー クエリを入力します。(Shift + Enter で新しい行)

164/128000 送信するトークン

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

## チャット

 アシスタン  
ト プレビュー リアルタイ  
ム プレビュー ムオーディ  
オ プレビュー

イメージ プレビュー

入力候補 プレビュー

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー プレビュー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ フィルタ

インデックス フィルタ

デプロイ フィルタ

コンテンツ フィルタ フィルタ

## チャットプレイグラウンド

コードの表示 プロンプトフロー 評価 デプロイ インポート エクスポート プロンプトのサンプル ...

保存

セクションの追加

ソースデータが追加されている場合、一部のショットの例は使用されません。以前の例はすべてクリアされました。データを削除すると、例を追加できます。

## データを追加する プレビュー

自分のデータソースへの分析情報を取得します。データは Azure サブスクリプションに安全に保存されます。[データの保護方法について説明します。](#)

インデックス:

[brochure-index](#)

検索の種類:

ハイブリッド(ベクトル+キーワード) さまざ

[まな検索の種類に関する詳細情報](#) [詳細設定](#)

チャットのクリア | チャット機能

JSON の表示

New York City, within an easy walk to Times Square and Broadway ^1^ .

2. **The Grand Central Hotel:** Comfortable mid-town hotel, close to Grand Central Station, the Chrysler Building, and the Empire State Building ^1^ .
3. **The Park Hotel:** Luxurious accommodation in upper Manhattan, with views of Central Park ^1^ .

To book your trip to New York, you can visit [Margie's Travel website](#) ^1^ .

1 参照

 1 [New York Brochure.pdf - Part 1](#)

ここにユーザー クエリを入力します。(Shift + Enter で新しい行)

164/128000 送信するトークン

## 引用

Title: New York Brochure.pdfMargie's Travel Presents...  
New York

The City of New York, often called New York City or simply New York, is the most populous city in the United States. With an estimated 2015 population of 8,550,405 distributed over a land area of about 302.6 square miles, New York City is also the most densely populated major city in the United States. Located at the southern tip of the state of New York, the city is the center of the New York metropolitan area, one of the most populous urban agglomerations in the world. A global power city, New York City exerts a significant impact upon commerce, finance, media, art, fashion, research, technology, education, and entertainment, its fast pace defining the term New York minute. Home to the headquarters of the United Nations, New York is an important center for international diplomacy and has been described as the cultural and financial capital of the world .

New York Hotels  
Margie's Travel offers the following accommodation options in New York : The Manhattan Hotel

# モジュール4 まとめ

- **RAG (Retrieval-Augmented Generation, 検索拡張生成) アーキテクチャ**は、大規模言語モデル（GPT等）に、外部のデータを接続するアーキテクチャ。大規模言語モデルは必要に応じて外部データを検索し、回答に必要なデータを取得してから、回答の生成を行う。これにより、企業内の資料などにもとづく回答が可能となる。
- **Azure AI Search**は、Azureの検索サービス。たとえば大量のPDF文書に対し、クエリ（検索）を実行して、特定のキーワードが含まれるPDF文書をすばやく発見できる。
- 「**埋め込み** モデル」は、テキストを**ベクトル化**するためのモデル。2つのベクトルを比較することで、2つのテキストの「意味的な類似度」を判定できる。**ベクトル検索**により、「意味的な類似度が高い文書」を探し出すことが可能となる。
- **Azure OpenAI on your data**は、**RAGアーキテクチャ**を簡単に実装することができる仕組み。チャットプレイグラウンドから利用できる。また独自のアプリ経由でも利用できる。

# モジュール5



## Azure AI Foundryで言語モデルを微調整する

38 分 残り • モジュール • 0/7 ユニットが完了しました

基本言語モデルをトレーニングし、Chat Completionsで利用できます。Azure AI Foundryのモデルカタログには、モデルのふるまいの具体的なニーズに合わせて微調整できる多くのオープンソースモデルが用意されています。

# モジュール5 フайнチューニング

- 微調整（ファインチューニング）とは？
- トレーニングデータの形式
- 微調整（ファインチューニング）に対応しているモデル
- 微調整（ファインチューニング）とRAGではどちらを使うべきか？
- 微調整（ファインチューニング）の実施例
- まとめ

# 微調整（ファインチューニング）とは？

- ・微調整（ファインチューニング）では、プロンプトに収まらない多くの例をトレーニングすることで、モデルが改善され、多数のタスクでより良い結果が得られるようになる
- ・**トレーニングデータ**を使用して**ベースモデル**（gpt-4などのモデル）をトレーニングする
- ・トレーニングが終わると**カスタムモデル**ができる
- ・**カスタムモデル**をデプロイして利用する

# モジュール5 フайнチューニング

- 微調整（ファインチューニング）とは？
  - トレーニングデータの形式
  - 微調整（ファインチューニング）に対応しているモデル
  - 微調整（ファインチューニング）とRAGではどちらを使うべきか？
  - 微調整（ファインチューニング）の実施例
  - まとめ

# トレーニングデータの形式: JSONL

```
{  
  "messages": [  
    {"role": "system", "content": "(システムメッセージ)"},  
    {"role": "user", "content": "(ユーザーの入力例)"},  
    {"role": "assistant", "content": "(生成AIの出力例)"}  
  ]  
}  
...
```

# トレーニングデータの実際の例

```
{"messages": [{"role": "system", "content": "あなたは Xbox カスタマー サポート エージェントであり、主な目的は Xbox デバイスで発生している問題についてユーザーを支援することです。あなたはフレンドリーで簡潔です。質問に対しては事実に基づいた回答のみを提供し、Xbox に関係のない回答は提供しません。"}, {"role": "user", "content": "Xbox は PlayStation より優れていますか?"}, {"role": "assistant", "content": "申し訳ありませんが、個人的な意見を提供することはできません。"}]}
```

```
{"messages": [{"role": "system", "content": "あなたは Xbox カスタマー サポート エージェントであり、主な目的は Xbox デバイスで発生している問題についてユーザーを支援することです。あなたはフレンドリーで簡潔です。質問に対しては事実に基づいた回答のみを提供し、Xbox に関係のない回答は提供しません。"}, {"role": "user", "content": "PC ゲームを Xbox でプレイできますか?"}, {"role": "assistant", "content": "Xbox コンソールは PC ゲームをネイティブにサポートしていませんが、さまざまな Xbox ゲームをプレイできます。その多くは PC でも利用できます。"}]}
```

# モジュール5 フайнチューニング

- ・微調整（ファインチューニング）とは？
- ・トレーニングデータの形式
- ・微調整（ファインチューニング）に対応しているモデル
- ・微調整（ファインチューニング）とRAGではどちらを使うべきか？
- ・微調整（ファインチューニング）の実施例
- ・まとめ

# 微調整（ファインチューニング）に対応するモデル

次のモデルでは、微調整がサポートされます。

- babbage-002
- davinci-002
- gpt-35-turbo (0613)
- gpt-35-turbo (1106)
- gpt-35-turbo (0125)
- gpt-4 (0613)\*
- gpt-4o (2024-08-06)
- gpt-4o-mini (2024-07-18)

# モジュール5 フайнチューニング

- ・微調整（ファインチューニング）とは？
- ・トレーニングデータの形式
- ・微調整（ファインチューニング）に対応しているモデル
- ・微調整（ファインチューニング）とRAGではどちらを使うべきか？
- ・微調整（ファインチューニング）の実施例
- ・まとめ

# 微調整（ファインチューニング）とRAG ではどちらを使うべきか？

- ・ファインチューニングは手間と時間、追加の費用がかかる。
  - ・微調整実施前のパフォーマンス評価
  - ・ファインチューニング用のトレーニングデータの準備
  - ・ファインチューニングの実施（カスタムモデルが作成される）
  - ・カスタムモデルのデプロイ
  - ・微調整実施後のパフォーマンス評価
- ・RAGは比較的簡単に実装でき、効果もすぐに確認できる。
- ・マイクロソフトのドキュメントでは、微調整を実施する前に、まずはプロンプトエンジニアリング（プロンプトの書き方の改善）やRAGを試してみることを推奨している。

<https://learn.microsoft.com/ja-jp/azure/ai-studio/concepts/fine-tuning-overview>

<https://learn.microsoft.com/ja-jp/azure/ai-services/openai/concepts/prompt-engineering?tabs=chat>

# モジュール5 フайнチューニング

- ・微調整（ファインチューニング）とは？
- ・トレーニングデータの形式
- ・微調整（ファインチューニング）に対応しているモデル
- ・微調整（ファインチューニング）とRAGではどちらを使うべきか？
- ・微調整（ファインチューニング）の実施例
- ・まとめ

# 微調整の実施例

- ・「微調整」画面へ移動
- ・ベースモデルを選択
- ・トレーニングデータをアップロード
- ・トレーニングを開始
- ・トレーニングが終わったら、カスタムモデルをデプロイ
- ・プレイグラウンドでデプロイの動作を確認

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス

プレビュー

リアルタイム

ムオーディオ

プレビュー

イメージ

入力候補

音声

プレビュー

ツール

コード

プレビュー

プロンプト フロー

評価

プレビュー

微調整

プレビュー

コンポーネント

データ

インデックス

デプロイ

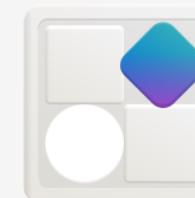
コンテンツ フィルタ

## 独自のデータでトレーニングしてモデルを微調整します

プレビュー

パフォーマンスと精度を向上させるために、より小さいタスク固有のデータセットでトレーニングすることにより、事前にトレーニングされたモデルを特定のタスクに最適化します。このメソッドではプロンプトで必要な例が少なくなる傾向があるため、一般的に、呼び出しごとに送信されるテキストと処理されるトークンの数が少なくなります。

+ モデルの微調整



### 最初のモデルを微調整します

モデルのタスク機能を更新するには、最初にトレーニングされた重みを使用して新しいデータを追加します。このメソッドではプロンプトで必要な例が少なくなる傾向があるため、一般的に、呼び出しごとに送信されるテキストと処理されるトークンの数が少なくなります。

[微調整に関する詳細情報 ↗](#)

## モデルの選択

表示されるモデルには、現在のプロジェクトと同じリージョンで微調整できるモデルのみが含まれます。 [微調整のためのリージョンの制約に関する詳細情報](#)

表示されているモデルの一覧には、現在のプロジェクトがあるリージョン内で微調整するためにアクセスできるモデルのみ反映されることにご注意ください。

検索

名前	コレクション	タスク
gpt-35-turbo	Azure OpenAI	チャットの完了
gpt-4o-mini	Azure OpenAI	チャットの完了
babbage-002	Azure OpenAI	完了
davinci-002	Azure OpenAI	完了

&lt; 前へ 次へ &gt;

確認

キャンセル

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo の微調整

① 基本設定

② トレーニング データ

③ 検証データ  
オプション④ タスク パラメーター  
オプション

⑤ レビュー

## 独自のトレーニング データを使用して、このモデルをカスタマイズします

すべての微調整されたモデルは、パフォーマンスと実行コストの両方に影響を与える基本モデルから開始されます。

## モデルバージョン \*

0125

## モデル サフィックス

ft-travel

## Azure OpenAI 接続 \*

ai-hub45518660444955840808\_aoai

選択したリソースにアクセスできるすべてのユーザーは、この微調整されたモデルにアクセスできます。 [ロールベースのアクセス制御の詳細情報](#)

次へ

送信

キャンセル

モデルのベンチマーク

プロンプトカタログ

AIサービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo の微調整

基本設定

トレーニングデータ

検証データ  
オプションタスクパラメーター  
オプション

レビュー

## トレーニングデータ

② ヘルプが必要ですか?

モデルをパーソナライズするデータセットを選択します。トレーニングデータは jsonl ファイルである必要があります、チャット入力候補の形式である必要があります。[データの準備に関する詳細をご覧ください。](#)

## トレーニングデータ\*

ファイルのアップロード

① 選択したリソースにアクセスできるすべてのユーザーは、このトレーニングファイルにアクセスできます。[ロールベースのアクセス制御の詳細情報](#)

ローカルドライブからアップロードするファイルを選択します。ファイルは 200 MB 未満である必要があります。

↑ ファイルのアップロード

 既に存在する場合は上書きする

## リストのアップロード

次のファイルの種類がサポートされています: *jsonl*

戻る

次へ

送信

キャンセル

のタスクに最適化します。  
ります。



整理 ▾

□ ?

▽ 推奨

アクティビティに基づいて、おすすめのファイルがここに表示されます。

▽ クイック アクセス



▽ お気に入り

いくつかのファイルをお気に入りに追加すると、ここに表示されます。

▽ 最近使用した項目

ファイルをいくつか開いた後、ここに最近使ったファイルが表示されます。

> PC

ファイル名(N): <https://raw.githubusercontent.com/MicrosoftLearning/mslearn-ai-studio/main/data/travel-finetune.jsonl>

JSONL ファイル (\*.jsonl)

開く(O)

キャンセル

モードルのベンチマーク

プロンプトカタログ

AIサービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー  
リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo の微調整

基本設定

トレーニングデータ

検証データ  
オプションタスクパラメーター  
オプション

レビュー

## トレーニングデータ \*

ファイルのアップロード

① 選択したリソースにアクセスできるすべてのユーザーは、このトレーニング ファイルにアクセスできます。 [ロールベースのアクセス制御の詳細情報](#)

ローカル ドライブからアップロードするファイルを選択します。 ファイルは 200 MB 未満である必要があります。

↑ ファイルのアップロード

 既に存在する場合は上書きする

## リストのアップロード

travel-finetune[1].jsonl

6.14 KB/6.14 KB

...

## messages

[{"role": "system", "content": "You are an AI travel assistant that helps people plan their trips. Your objective is t..."}]

[{"role": "system", "content": "You are an AI travel assistant that helps people plan their trips. Your objective is t..."}]

[{"role": "system", "content": "You are an AI travel assistant that helps people plan their trips. Your objective is t..."}]

戻る

次へ

送信

キャンセル

のタスクに最適化します。  
ります。

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo の微調整

✓ 基本設定

✓ トレーニング データ

③ 検証データ  
オプション④ タスク パラメーター  
オプション

⑤ レビュー

## 検証データ

② ヘルプが必要ですか?

モデルをパーソナライズするデータセットを選択します。トレーニング データは jsonl ファイルである必要があります、チャット入力候補の形式である必要があります。データの準備に関する詳細をご覧ください。

## 検証データ \*

なし

戻る

次へ

送信

キャンセル

のタスクに最適化します。  
ります。

は、ここをクリック... ▾ X

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス  
ト  
リアルタイム  
ムオーディオ

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo の微調整

- 基本設定
- トレーニング データ
- 検証データ  
オプション
- タスク パラメーター  
オプション
- レビュー

## タスク パラメーター

- Batch size i
- Default  Custom
- Learning rate multiplier i
- Default  Custom
- Number of epochs i
- Default  Custom
- Seed i
- Random  Custom

戻る

次へ

送信

キャンセル

のタスクに最適化します。  
ります。は、ここをクリッ...  

モデルのベンチマーク

プロンプトカタログ

AIサービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo の微調整

- 基本設定
- トレーニングデータ
- 検証データ  
オプション
- タスク パラメーター  
オプション
- レビュー

### レビュー

送信する前に、ジョブを確認または変更してください。

#### 基本設定

微調整されたモデル名



gpt-35-turbo-Finetune

モデル サフィックス

ft-travel

モデル バージョン

0125

#### タスクの設定

タスクの種類

Chat

#### トレーニングデータ

travel\_finetune\_1\_jsonl\_2024-11-07\_153039\_UTC  
ユーザー データ

#### 検証データ

なし

戻る

送信

キャンセル

モードルのベンチマークプロンプトカタログAIサービスプロジェクトのプレイグラウンドチャットアシスタンス  
リアルタイムオーディオイメージ入力候補音声ツールコードプロンプトフロー評価微調整コンポーネントデータインデックスデプロイコンテンツ フィルタ

## 独自のデータでトレーニングしてモデルを微調整します

プレビュー

パフォーマンスと精度を向上させるために、より小さいタスク固有のデータセットでトレーニングすることにより、事前にトレーニングされたモデルを特定のタスクに最適化します。このメソッドではプロンプトで必要な例が少なくなる傾向があるため、一般的に、呼び出しごとに送信されるテキストと処理されるトークンの数が少なくなります。

✓ 成功: travel\_finetune\_1\_jsonl\_2024-11-07\_153039\_UTC データ資産が正常に作成されました。リストが更新されるまでに数秒かかる場合があります。このデータセットに戻るには、[ここをクリック](#)...

+ モデルの微調整最新の情報に更新ビューのリセット検索フィルター列

モデル名	基本モデル	状態	作成日
ai-hub45518660444955840808_aoui (1)	ftjob-2bceb6a0532a46868471c6c947e04d4d	gpt-35-turbo-0125 Queued	Nov 8, 2024 12:33 AM

モードルのベンチマークプロンプトカタログAIサービスプロジェクトのプレイグラウンドチャットアシスタンス  
リアルタイムオーディオイメージ入力候補音声ツールコードプロンプトフロー評価微調整コンポーネントデータインデックスデプロイコンテンツフィルタ

## 独自のデータでトレーニングしてモデルを微調整します

プレビュー

パフォーマンスと精度を向上させるために、より小さいタスク固有のデータセットでトレーニングすることにより、事前にトレーニングされたモデルを特定のタスクに最適化します。このメソッドではプロンプトで必要な例が少なくなる傾向があるため、一般的に、呼び出しごとに送信されるテキストと処理されるトークンの数が少なくなります。

✓ 成功: travel\_finetune\_1\_jsonl\_2024-11-07\_153039.UTC データ資産が正常に作成されました。リストが更新されるまでに数秒かかる場合があります。このデータセットに戻るには、[ここをクリック](#)...

+ モデルの微調整最新の情報に更新ビューのリセット検索フィルター列

モデル名	基本モデル	状態	作成日
ai-hub45518660444955840808_aoui (1) <a href="#">ftjob-2bceb6a0532a46868471c6c947e04d4d</a>	gpt-35-turbo-0125	Running	Nov 8, 2024 12:33 AM

 モデルのベンチマーク プロンプト カタログ AI サービスプロジェクトのプレイグラウンド  チャット アシスタン

ト

 リアルタイム ムオーディ

オ

 イメージ 入力候補 音声

プレビュー

ツール  コード

プレビュー

 プロンプトフロー 評価

プレビュー

 微調整

プレビュー

コンポーネント  データ インデックス デプロイ コンテンツ フィルタ

## 独自のデータでトレーニングしてモデルを微調整します

パフォーマンスと精度を向上させるために、より小さいタスク固有のデータセットでトレーニングすることにより、事前にトレーニングされたモデルを特定のタスクに最適化します。このメソッドではプロンプトで必要な例が少なくなる傾向があるため、一般的に、呼び出しごとに送信されるテキストと処理されるトークンの数が少なくなります。

 モデルの微調整 最新の情報に更新 ビューのリセット 検索 フィルター 列

モデル名	基本モデル	状態	作成日
 ai-hub45518660444955840808_aoai (1)			
<a href="#">gpt-35-turbo-0125.ft-2bceb6a0532a46868471c6</a>	gpt-35-turbo-0125	Completed	Nov 8, 2024 12:33 AM



モデルのベンチマーク

プロンプトカタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス  
ト

リアルタイム

ムオーディオ

イメージ

入力候補

音声  
プレビュー

ツール

コード  
プレビュー

プロンプトフロー

評価  
プレビュー

微調整  
プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## gpt-35-turbo-0125.ft-2bceb6a0532a46868471c6c947e04d4d-ft-travel

詳細 メトリクス ログ チェックポイント

デプロイ

最新の情報に更新

削除

継続的な微調整

### モデル属性

#### ID

ftjob-2bceb6a0532a46868471c6c947e04d4d

#### 状態

完了

#### 作成日

Nov 8, 2024 12:33 AM

#### 基本モデル

gpt-35-turbo-0125

#### トレーニング ファイル

-

#### 検証ファイル

-

#### Azure OpenAI リソース

ai-hub45518660444955840808\_aoai

### タスクパラメーター

#### バッチ サイズ

1

#### 学習率の乗数

既定値

#### エポックの数

10

#### シード

335020092

結果のダウンロード

トレーニング ファイルをダウンロードする

検証ファイルのダウンロード

## gpt-35-turbo-

詳細 メトリクス ▾

▷ デプロイ ▾

## モデル属性

ID  
ftjob-2bceb6a0532a46プロジェクト  
Project45518660基本モデル  
gpt-35-turbo-0125トレーニング ファイル  
UI/2024-11-07\_153035  
finetune%5B1%5D.jsonAzure OpenAI リソース  
ai-hub4551866044495トレーニングの完了日  
Nov 8, 2024 1:16 AM請求済みのトレーニング  
12,000

↓ 結果のダウンロード

## モデル gpt-35-turbo-0125.ft-

## 2bceb6a0532a46868471c6c947e04d4d-ft-travel をデプロイする

## 現在のプロジェクト リソース

project45518660

## デプロイ名 \*

gpt-35-turbo-0125-ft-travel

## デプロイの種類

Standard

Standard: Pay per API call with lower rate limits. Adheres to Azure data residency promises. Best for intermittent workloads with low to medium volume. Learn more about [Standard deployments](#).

## 接続された Azure OpenAI リソース

ai-hub45518660444955840808\_aoui

① デプロイで使用可能な 1 分あたりの 250K トークンクォータ

## 1分あたりのトークン数レート制限 ①

50K

対応する 1 分あたりの要求数 (RPM) = 300

## コンテンツ フィルター ①

DefaultV2

## 動的クオータを有効にする ①

 有効

デプロイ

取り消し

モデルのベンチマーク  
プロンプトカタログ  
AI サービス

プロジェクトのプレイグラウンド

チャット  
アシスタンス プレビュー  
リアルタイム  
ムオーディオ プレビュー  
イメージ  
入力候補  
音声 プレビュー

ツール  
</> コード プレビュー  
プロンプトフロー  
評価 プレビュー

微調整 プレビュー  
コンポーネント  
データ  
インデックス  
デプロイ  
コンテンツ フィルタ

## gpt-35-turbo-0125.ft-2bceb6a0532a46868471c6c947e04d4d-ft-travel

詳細 メトリクス ログ チェックポイント

▷ デプロイ ⏪ 最新の情報に更新 削除 継続的な微調整

### モデル属性

ID  
ftjob-2bceb6a0532a46868471c6c947e04d4d

プロジェクト  
Project45518660

基本モデル  
gpt-35-turbo-0125

トレーニング ファイル  
UI/2024-11-07\_153039.UTC/travel-finetune%5B1%5D.jsonl

Azure OpenAI リソース  
[ai-hub45518660444955840808\\_aoui](#)

状態  
完了

作成日  
Nov 8, 2024 12:33 AM

検証ファイル

-

### タスクパラメーター

バッチ サイズ  
1

学習率の乗数  
既定値

エポックの数  
10

シード  
335020092

トレーニングの完了日  
Nov 8, 2024 1:16 AM

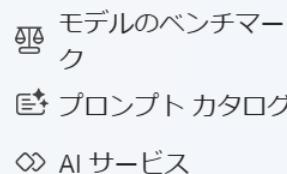
期間  
42 分 44 秒

請求済みのトレーニング トークン  
12,000

↓ 結果のダウンロード

トレーニング ファイルをダウンロードする

検証ファイルのダウンロード



# gpt-35-turbo-0125-ft-travel

## 詳細 メトリック

▶ プレイグラウンドで聞く

編集

 削除

## 配置情報

<b>名前</b>	gpt-35-turbo-0125-ft-travel	<b>プロビジョニングの状態</b>
		成功
<b>展開の種類</b>	Standard	<b>作成日</b>
		2024-11-07T17:37:43.0386635Z
<b>作成者</b>	3468d260-6c03-4a54-ab2d-9d9bb17217c5	<b>変更日</b>
	Nov 8, 2024 2:37 AM	
<b>更新者</b>	3468d260-6c03-4a54-ab2d-9d9bb17217c5	<b>バージョン更新ポリシー</b>
		モデルのバージョンは自動的にアップグレードされません
<b>レート制限 (1分あたりのトークン数)</b>	50,000	<b>レート制限 (1分あたりの要求数)</b>
		300
<b>モデル名</b>	gpt-35-turbo-0125.ft-travel	<b>モデル バージョン</b>
		1
<b>作成日</b>	Nov 8, 2024 1:06 AM	<b>更新日</b>
		Nov 8, 2024 2:37 AM
<b>モデルの廃止日</b>	Feb 15, 2025 9:00 AM	

エンドポイント

#### ターゲット URI

22

### 認証の種類

三

十一

監視と安全性

コンテンツ フィルター

## DefaultV?

### アプリケーション開発に役立つリンク

コード サンプル リポジトリ

チコートリアル [1]



# モジュール5 フайнチューニング

- ・微調整（ファインチューニング）とは？
- ・トレーニングデータの形式
- ・微調整（ファインチューニング）に対応しているモデル
- ・微調整（ファインチューニング）とRAGではどちらを使うべきか？
- ・微調整（ファインチューニング）の実施例
- ・まとめ

# モジュール5まとめ

- gpt-35-turbo のような一部のモデルは、**微調整（ファインチューニング）** が可能。
- 微調整により、デフォルトのモデル（**ベースモデル**）には含まれていない知識をトレーニングで与え、**カスタムモデル**を構築できる。
- トレーニングのデータはJSONL形式で作成する。この中にはシステムメッセージ、ユーザーの入力例、生成AIの出力例を含める。
- **カスタムモデル**は、デフォルトのモデルと同様に、デプロイして、プレイグラウンドや、外部のアプリから利用することができる。
- 微調整は時間と手間、コストがかかるため、まずはプロンプトエンジニアリングや、RAGアーキテクチャを試してみることが推奨されている。

# モジュール6



## Azure AI Foundry で言語モデルのパフォーマンスを評価する

20 分 残り • モジュール • 4/7 ユニットが完了しました

言語モデルがユーザーのニーズを満たし、正確な応答を提供し、時間の経過とともに継続的に改善されていることを確認するには、評価の仕組みが不可欠です。Azure AI Foundry で言語モデルのパフォーマンスを評価し、最適化する方法について調べます。

# モジュール6 パフォーマンス評価

- ・パフォーマンス評価とは？
- ・パフォーマンス評価の観点
- ・パフォーマンス評価の実施例
- ・まとめ

# パフォーマンス評価とは？

- ・生成AIモデルや、カスタムモデルに質問を与え、一貫性のある（筋道の通った）回答ができるかどうか、文法的に正しい回答ができるかどうかなどを評価するしくみ。

# モジュール6 パフォーマンス評価

- ・パフォーマンス評価とは？
- ・パフォーマンス評価の観点
- ・パフォーマンス評価の実施例
- ・まとめ

# パフォーマンス評価の観点

- コヒーレンス (coherence)
  - 言語モデルが一貫性のある（筋道の通った）出力を生成することを評価する。
- 流暢性 (fluency)
  - 生成型 AI の予測される回答の言語能力を評価する。生成されたテキストが文法規則、構文構造、ボキャブラリの適切な使用方法にどの程度準拠しているかを評価する。

# モジュール6 パフォーマンス評価

- ・パフォーマンス評価とは？
- ・パフォーマンス評価の観点
- ・パフォーマンス評価の実施例
- ・まとめ

# パフォーマンス評価の実施例

- 「評価」のページに移動
- 新しい「評価」を作成
  - 評価に使用するデータを指定
  - 評価対象のデプロイを選択

モデルのベンチマーク

プロンプトカタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## Assess and compare AI application performance

[Automated evaluations プレビュー](#) [Manual evaluations](#) [Evaluator library プレビュー](#)

Evaluate the quality and safety of your generative AI applications with industry standard metrics to compare and choose the best version based on your need. [Learn more about metrics.](#)

[+ New evaluation](#)

最新の情報に更新

取り消す

削除

Compare

Show batch runs

### > Evaluation process

#### Deep dive into your results

Investigate the details of your app's performance across each data point. You can search for specific terms within the table and customize your view by adding filters and managing columns.

[View documentation](#) 

#### Detailed metrics result

Index	Status	inputs.message	Groundedness	Relevance	Coherence	Fluency	Retrieval score	Violence	Violence reason	Self-harm
0	Completed	<pre>[{"content": "will my compass work in patagonia", "role": "User", "context": "", "customer_info": "#customer_info\n\\n\\tname: Jane Doe\\n\\tage: 28\\n\\tphone_number: 555-987-6543\\n\\titem_number: 7\\n\\tcha..."}, {"content": "will my compass work in patagonia", "role": "System", "context": "", "customer_info": "#customer_info\n\\n\\tname: Jane Doe\\n\\tage: 28\\n\\tphone_number: 555-987-6543\\n\\titem_number: 7\\n\\tcha..."}]</pre>	3.67	3.33	5	5	2	Very low	The system's responses provide factual information about a compass and a sleeping bag, with no mention or implication of violence or physical harm.	Very low



## Assess and compare AI application performance

### Automated evaluations

Evaluate the quality and

+ New evaluation

> Evaluation p

### What do you want to evaluate?



#### Model and prompt

選択したモデルとユーザー定義プロンプトによって生成された出力を評価します。



#### Dataset

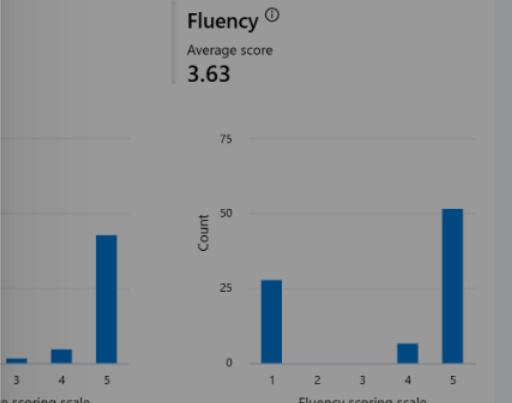
Evaluate an existing output dataset.



#### Prompt flow

Select a flow you've created and evaluate its output.

Cancel



モデルのベンチマーク  
プロンプトカタログ  
AI サービス

プロジェクトのプレイグラウンド

チャット  
アシスタント  
リアルタイム  
ムオーディオ  
イメージ

入力候補  
音声

ツール  
コード  
プロンプトフロー

評価  
微調整

コンポーネント  
データ  
インデックス

デプロイ  
コンテンツ フィルタ

モデルのベンチマーク

プロンプトカタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント

プレビュー

リアルタイム

ムオーディオ

プレビュー

イメージ

入力候補

音声

プレビュー

ツール

コード

プレビュー

プロンプトフロー

評価

プレビュー

微調整

プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## Create a new evaluation プレビュー

### 基本情報

評価名 \*

nice\_kitchen\_xl5szv129c

デプロイ \*

gpt-35-turbo-0125-ft-travel (version:1)



### パラメーター

最大応答

800

温度

0.7

Top P

0.95

システム メッセージ \*

You are an AI assistant that helps people find information.

+ 例を追加する

### テストデータの構成

評価するデータを選択する \*

サンプルデータの生成

GPT モデルを使用してサンプルの質問を生成し、回答を評価します

既存のデータセットを使用する

確立されたデータセットのコレクションから選択する

データセットの追加

ファイルのアップロード

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタン  
ト プレビューリアルタイ  
ムムオーディ  
オ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

## 評価するデータを選択する \*

## サンプルデータの生成

GPT モデルを使用してサンプルの質問を生成し、回答を評価します

## 既存のデータセットを使用する

確立されたデータセットのコレクションから選択する

## データセットの追加

ファイルのアップロード

## 次に関する質問を生成するトピック \* ⓘ

猫について、日本語で例を生成

生成されたサンプルの質問のプレビュー (合計: 5 行)

## 質問

猫は好奇心旺盛ですか？

## 回答

はい、猫は好奇心旺盛な生き物です。例えば、猫が新しい箱を見つけると、すぐ...

猫は狩りをするのですか？

はい、猫は狩りをすることがあります。例えば、夜になると庭でネズミを捕まえ...

猫は夜よく鳴きますか？

猫は夜中に「ニヤーニヤー」と鳴いて、眠れない夜を過ごさせてくれることがあ...

猫は人間をなめることができますか？

はい、猫は時々人間をなめることができます。例えば、猫は飼い主の手をなめて...

猫は高い所が好きですか？

はい、猫は高い所が好きです。例えば、猫は本棚の上や冷蔵庫の上など、高い場...

再生成する

ⓘ 生成されたデータセットは、評価実行を作成する際、プロジェクトの BLOB ストレージに格納されます。

モデルのベンチマーク

プロンプトカタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント

プレビュー

リアルтай

ムオーディオ

プレビュー

イメージ

入力候補

音声

プレビュー

ツール

&lt;/&gt; コード

プレビュー

プロンプトフロー

評価

プレビュー

微調整

プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルタ

-

猫は狩りをするのですか？

はい、猫は狩りをすることがあります。例えば、夜になると庭でネズミを捕まえ...

猫は夜よく鳴きますか？

猫は夜中に「ニヤーニヤー」と鳴いて、眠れない夜を過ごさせてくれることがあ...

猫は人間をなめることができますか？

はい、猫は時々人間をなめることができます。例えば、猫は飼い主の手をなめて...

猫は高い所が好きですか？

はい、猫は高い所が好きです。例えば、猫は本棚の上や冷蔵庫の上など、高い場...

再生成する

生成されたデータセットは、評価実行を作成する際、プロジェクトの BLOB ストレージに格納されます。

## 評価する内容を選択してください \*



## 言語

メジャーは、言語モデルがスムーズに流れ、自然に読み取り、人間のような言語に似た出力を生成できる程度を評価します。これにより、以下の評価を行います **コピーレンス**、および **流暢性**。



## 安全性

コンテンツの安全性基準を満たしていないコンテンツの量を測定します。これにより、以下の評価を行います **自傷行為に関連するコンテンツ**、**悪意のある不公平なコンテンツ**、**暴力的なコンテンツ**、および **性的な内容**。

## ジャッジとしてモデル デプロイを選択する \*

gpt-35-turbo-0125-ft-travel (version:1)



作成

キャンセル

モードルのベンチマーク  
プロンプトカタログ  
AI サービス

プロジェクトのプレイグラウンド

チャット  
アシスタン  
ト  
リアルタイ  
ムオーディ  
オ  
イメージ  
入力候補  
音声

ツール  
コード  
プロンプトフロー  
評価  
微調整

コンポーネント  
データ  
インデックス  
デプロイ  
コンテンツ フィルタ

## evaluation\_teal\_sugar\_lw82lsvs4n

② Not satisfied with results?

Output Logs Tracing

⟳ 更新

### Evaluation details

Status	Dataset	評価されたモデル	判断者としてのモデル	説明
未開始	tender_parsnip_qkx66x9zgj:1	gpt-35-turbo-0125-ft-travel-2	gpt-35-turbo-0125-ft-travel	<span>編集</span>
Create time	タグ	最大トークン数	温度	Top P
Nov 8, 2024 3:14 AM	タグがありません	800	0.7	0.95
Created by	See all properties	システム メッセージ		
User1-45518660	生 JSON	You are an AI assistant that helps people find information.		

モードルのベンチマークプロンプトカタログAIサービスプロジェクトのプレイグラウンドチャットアシスタンス プレビューリアルタイムオーディオ プレビューイメージ入力候補音声 プレビューツールコード プレビュープロンプトフロー評価 プレビュー微調整 プレビューコンポーネントデータインデックスデプロイコンテンツ フィルタ

## メトリック ダッシュボード

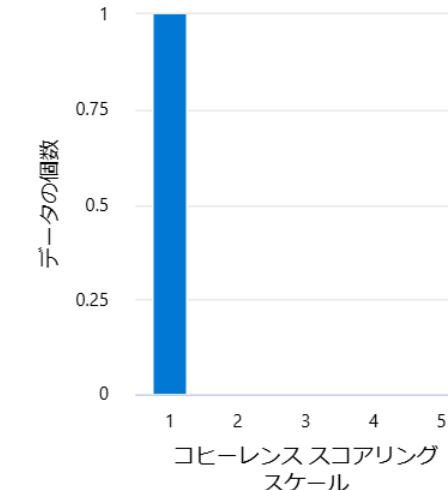
AI品質 (AI支援)

適用されたフィルター

すべてのメトリック

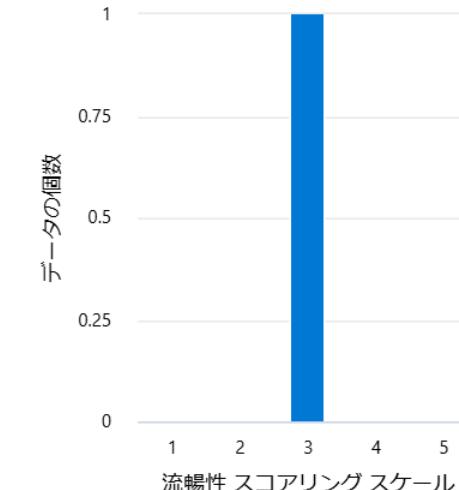
### コヒーレンス ⓘ

平均スコア

**1.00**

### 流暢性 ⓘ

平均スコア

**3.00**

## 詳細なメトリックの結果

検索フィルター列インデックス状態コヒーレンス流暢性queryresponse

0

完了

1

3

猫は肉食動物ですか？耳の動きで何を表現して

猫は肉食動物です。耳の

# モジュール6 パフォーマンス評価

- ・パフォーマンス評価とは？
- ・パフォーマンス評価の観点
- ・パフォーマンス評価の実施例
- ・まとめ

# モジュール6まとめ

- Azure AI Foundryを使用して、モデルのデプロイのパフォーマンスを評価できる
- 評価の観点には「コヒーレンス」（一貫性）と、流暢性（文法・構文・ボキャブラリの適切な使用）などがある
- Azure AI Foundryで「評価」を作成し、サンプルデータをアップロードする（または生成AIを使ってサンプルデータを生成する）
- 評価の実行結果がグラフで表示される

# モジュール7



## Azure AI Foundry での責任ある生成 AI

50 分 • モジュール • 9 ユニット

生成 AI を使用すると、優れたクリエイティブソリューションが可能になりますが、有害なコンテンツ生成のリスクを最小限に抑えるには、責任を持って実装する必要があります。

# モジュール6 責任ある生成AI

- ・生成AI利用時の「潜在的なリスク」とは？
- ・開発者は「潜在的なリスク」にどう対処すればよいのか？
- ・Azure OpenAI モデルにおける責任ある AI プラクティス
- ・コンテンツフィルターとは？
- ・コンテンツフィルターの設定例
- ・まとめ

# 生成AI利用時の「潜在的なリスク」とは？

- ・たとえば、ユーザーが食材を入力すると、その食材を使った料理のレシピを生成するAIアプリを作るとする。
- ・食材の加熱時間が不十分であるレシピが生成されると、そのレシピに沿って料理を作つて食べた人が病気になつたり死亡したりする可能性がある。
- ・このように、生成AIを使ったアプリやシステム（生成AIソリューション）には「**潜在的なリスク**」（隠れたりスク）がある。
  - ・※「潜在的」 = 表に現れていないもの。反対語は「顕在的」 = 表に現れているもの。
  - ・※「リスク」 = 悪いことが起こる可能性。

# モジュール6 責任ある生成AI

- ・生成AI利用時の「潜在的なリスク」とは？
- ・開発者は「潜在的なリスク」にどう対処すればよいのか？
- ・Azure OpenAI モデルにおける責任ある AI プラクティス
- ・コンテンツフィルターとは？
- ・コンテンツフィルターの設定例
- ・まとめ

# 開発者は「潜在的なリスク」にどう対処すればよいのか？

- ・マイクロソフトの専門家が定義した  
**「Azure OpenAI モデルにおける責任ある AI プラクティス」**  
(Responsible AI practices for Azure OpenAI models)  
というドキュメントがあるので、開発者はそれを参考にして、  
**識別・測定・緩和・運用の4ステージ**を実行するとよい。
  - ・※プラクティス = 実践
  - ・「Azure OpenAI モデルにおける責任ある AI プラクティス」は以下のページで確認できる <https://learn.microsoft.com/ja-jp/legal/cognitive-services/openai/overview>

# モジュール6 責任ある生成AI

- ・生成AI利用時の「潜在的なリスク」とは？
- ・開発者は「潜在的なリスク」にどう対処すればよいのか？
- ・Azure OpenAI モデルにおける責任ある AI プラクティス
- ・コンテンツフィルターとは？
- ・コンテンツフィルターの設定例
- ・まとめ

ステージ名	作業名称	作業内容
ステージ 1: 識別	潜在的な害の識別	<p>ソリューションの計画に関連する潜在的な害を特定する。</p> <p>※「害（harm）」=人間に対する身体的・健康的な傷害や、財産や環境に対する損害など。</p>
	害の優先順位付け	特定された害の発生頻度と影響度を評価し、優先順位を設定する。
	害のテストと検証	優先順位付けされた害が実際に発生するかどうかをテストし、条件を確認する。
	害の文書化と共有	特定された害の詳細を文書化し、ステークホルダーと共有する。

ステージ名	作業名称	作業内容
ステージ 2: 測定	入力プロンプトの準備	特定された害を引き起こす可能性のある多様な入力プロンプトを準備する。
	出力の生成	準備したプロンプトをシステムに送信し、生成された出力を取得する。
	有害な結果の測定	事前定義された基準を使用して出力を評価し、有害な結果を分類する。

ステージ名	作業名称	作業内容
ステージ 3: 緩和	モデル層での緩和	適切なモデルの選択や、独自のトレーニングデータでのファインチューニングを行う。
	セーフティシステム層での緩和	<b>コンテンツフィルター</b> の設定や、悪用検出アルゴリズムの実装を行う。※ <b>コンテンツフィルター</b> についてはこの後解説
	メタプロンプトとグラウンド層での緩和	システム入力の定義や、信頼できるデータソースからのコンテキストデータの取り込みを行う。
	ユーザーエクスペリエンス層での緩和	入力と出力の検証や、システムの機能と制限に関する透明性のあるドキュメントの提供を行う。

ステージ名	作業名称	作業内容
	プレリリースレビューの完了	リリース前に、法務、プライバシー、セキュリティ、アクセシビリティなどのコンプライアンスレビューを実施する。
ステージ 4: 運用	ソリューションのリリースと運用	フェーズごとのリリース計画を立て、インシデント対応計画やロールバック計画を作成する。また、ユーザーからのフィードバックを受け取る仕組みや、テレメトリーデータを追跡する仕組みを実装する。

# モジュール6 責任ある生成AI

- ・生成AI利用時の「潜在的なリスク」とは？
- ・開発者は「潜在的なリスク」にどう対処すればよいのか？
- ・Azure OpenAI モデルにおける責任ある AI プラクティス
- ・コンテンツフィルターとは？
- ・コンテンツフィルターの設定例
- ・まとめ

# コンテンツフィルターとは？

- GPTやDALL-E等において、不適切な入力や出力に対するフィルタリングを行うしくみ
- 「**潜在的なリスク**」を緩和する手段の一つ
- 入力のフィルタリング: ユーザーが入力する可能性がある危険なプロンプト、例えば「**爆弾の作り方を教えて下さい**」といったプロンプトがフィルタリングされる。この場合、GPTやDALL-Eにはそのプロンプトは送信されず、回答は生成されない
- 出力のフィルタリング: GPTやDALL-Eが万が一、不適切な回答を生成した場合、それがフィルタリングされる。この場合、GPTやDALL-Eが生成されたコンテンツは送信されない。

# コンテンツフィルターの4カテゴリ

- ・暴力 (violence)
- ・ヘイト (hate)
- ・性的 (sexual)
- ・自傷 (self-harm)
- ・たとえば「爆弾の作り方を教えて」といった入力は  
「暴力」カテゴリの「中程度」といったように判定される。

# オプションで使用できるカテゴリ

- 入力フィルター
  - Prompt shields for jailbreak attacks: 攻撃者がシステムプロンプトの設定を無視したり回避したりしようとする攻撃を検出
  - Prompt shields for indirect attacks: 攻撃者がデータの中にプロンプトを紛れ込ませ、そのプロンプトを実行させようとする攻撃を検出
- 出力フィルター
  - Protected material for text: 歌の歌詞など、著作権のあるテキストの生成を検出
  - Protected material for code: GitHubリポジトリに存在するコードと同じのコードの生成を検出

# モジュール6 責任ある生成AI

- ・生成AI利用時の「潜在的なリスク」とは？
- ・開発者は「潜在的なリスク」にどう対処すればよいのか？
- ・Azure OpenAI モデルにおける責任ある AI プラクティス
- ・コンテンツフィルターとは？
- ・コンテンツフィルターの設定例
- ・まとめ

# コンテンツフィルターの設定例

- Azure AI Foundryで「コンテンツフィルター」を定義する
- GPTやDALL-Eのデプロイに対して「コンテンツフィルター」を適用する

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタンス  
ト プレビューリアルタイ  
ムムオーディ  
オ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプトフロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルター

## 不適切なコンテンツを管理する

コンテンツ フィルターは、コア モデルと共に機能します。プロジェクト内でフィルターを作成し、デプロイに割り当てて、カテゴリ別にコンテンツを管理します。特定の用語を防ぐためのブロックリストを作成します。

[コンテンツ フィルターとブロックリストに関する詳細情報](#)

コンテンツ フィルター ブロックリスト(プレビュー)

コンテンツ フィルターは、コア モデルと共に機能します。プロジェクト内でフィルターを作成し、デプロイに割り当てて、カテゴリ別にコンテンツを管理します。特定の用語を防ぐためのブロックリストを作成します。

[コンテンツ フィルターとブロックリストに関する詳細情報](#)

+ コンテンツ フィルターの作成

編集

削除

更新



コンテンツ フィルターがまだ作成されていません。

開始するには、上の [コンテンツ フィルターの作成] ボタンをクリックします。

モデルのベンチマー  
ク

プロンプト カタログ

AI サービス

プロジェクトのプレイグ  
ラウンド ^

チャット

アシスタン  
ト プレビュ

リアルタイ  
ム

オーディ  
オ プレビュ

イメージ

入力候補

音声 プレビュ

ツール ^

コード プレビュ

プロンプトフロー

評価 プレビュ

微調整 プレビュ

コンポーネント ^

データ

インデックス

デプロイ

コンテンツ フィルタ  
ー

## 特定の種類のコンテンツを許可またはブロックするフィルターを作成する

← コンテンツ フィルターに戻る

### 基本情報の追加

名前 \*

myfilter1

Connection \*

ai-admin1863ai433032826210\_aoai

### 基本情報

入力フィルター

出力フィルター

デプロイ (オプション)

レビュー

次へ

キャンセル

 設定

概要

 モデル カタログ モデルのベンチマーク プロンプト カタログ AI サービスプロジェクトのフレイグ  
ラウンド チャット アシスタン  
ト  プレビュー リアルタイ  
ム ムオーディ  
オ  プレビュー イメージ 入力候補 音声  プレビュー

ツール

 コード プレビュー プロンプト フロー 評価  プレビュー 微調整  プレビュー

コンポーネント

 データ インデックス デプロイ コンテンツ フィルタ  
ー

## 特定の種類のコンテンツを許可またはブロックするフィルターを作成する

[← コンテンツ フィルターに戻る](#) 基本情報 入力フィルター 出力フィルター デプロイ (オプション) レビュー

### 入力フィルターの設定

コンテンツにはカテゴリ別に注釈が付けられ、設定したしきい値に従ってブロックされます。暴力、憎悪、性的、および自傷行為のカテゴリについては、スライダーを調整して重大度が高、中、低のコンテンツをブロックします。

カテゴリ	メディア	アクション	しきい値
Violence	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Hate	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Sexual	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Self-harm	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Prompt shields for jailbreak attacks	Text	Annotate and block	<input checked="" type="checkbox"/> Jailbreak attacks will be blocked  コンテンツに注釈が付けられること
Prompt shields for indirect attacks	Text	Off	 コンテンツに注釈が付けられること

[戻る](#)[次へ](#)[キャンセル](#)



概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

&lt;/&gt; コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

デプロイ

コンテンツ フィルター

## 特定の種類のコンテンツを許可またはブロックするフィルターを作成する

[← コンテンツ フィルターに戻る](#) 基本情報 入力フィルター 出力フィルター デプロイ (オプション) レビュー

Set output filter

これらのカテゴリとは 

コンテンツには各カテゴリ別に注釈が付けられ、設定したしきい値に従ってブロックされます。暴力的なコンテンツ、憎悪表現を含むコンテンツ、性的なコンテンツ、自傷行為に関するコンテンツのカテゴリについては、同程度以上の重大度で有害なコンテンツをブロックできるよう、しきい値を調整してください。

カテゴリ	メディア	アクション	しきい値
Violence	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Hate	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Sexual	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Self-harm	Text Image	Annotate and block	 中 低を許可 / 中および高をブロック
Protected material for text	Text	Annotate and block	 <input checked="" type="checkbox"/> Protected material will be blocked

[戻る](#)[次へ](#)[キャンセル](#)

 設定

概要

 モデル カタログ モデルのベンチマーク プロンプト カタログ AI サービスプロジェクトのプレイグラウンド  チャット アシスタント  リアルタイム  
オーディオ イメージ 入力候補 音声 プレビューツール  コード  プロンプト フロー 評価  微調整 コンポーネント  データ インデックス

## 特定の種類のコンテンツを許可またはブロックするフィルターを作成する

 コンテンツ フィルターに戻る

- 基本情報
- 入力フィルター
- 出力フィルター
- デプロイ (オプション)
- レビュー

デプロイへのフィルターの適用 (省略可能)

接続

ai-admin1863ai433032826210\_aoai

デプロイ

<input type="checkbox"/> 名前	モデル名	モデル バージョン	コンテンツ フィルター	変更日
<input checked="" type="checkbox"/> gpt-4o	gpt-4o	2024-05-13	 Microsoft.DefaultV2	Nov 7, 2024 10:23 AM
<input type="checkbox"/> text-embedding-ada...	text-embedding-ada...	2	 Microsoft.DefaultV2	Nov 8, 2024 8:36 AM

戻る

次へ

キャンセル

設定

概要

モデル カタログ

モデルのベンチマーク

プロンプト カタログ

AI サービス

プロジェクトのプレイグラウンド

チャット

アシスタント プレビュー

リアルタイム

ムオーディオ プレビュー

イメージ

入力候補

音声 プレビュー

ツール

コード プレビュー

プロンプト フロー

評価 プレビュー

微調整 プレビュー

コンポーネント

データ

インデックス

## 特定の種類のコンテンツを許可またはブロックするフィルターを作成する

← コンテンツ フィルターに戻る

- 基本情報
- 入力フィルター
- 出力フィルター
- デプロイ (オプション)
- レビュー

コンテンツ フィルターの構成を確認してください

## 基本情報

  
**コンテンツ フィルター名** **接続**

myfilter1 ai-admin1863ai433032826210\_aoai

## 入力フィルター

## 注釈付けとブロック

カテゴリ	メディアの種類	アクション	ブロック
Violence	Text, Image	Annotate and block	Medium and High
Hate	Text, Image	Annotate and block	Medium and High
Sexual	Text, Image	Annotate and block	Medium and High
Self-harm	Text, Image	Annotate and block	Medium and High
Prompt shields for jailbreak attacks	Text	Annotate and block	Jailbreak attacks will be blocked
Prompt shields for indirect attacks	Text	Off	-

## 入力ブロックリスト

戻るフィルターの作成キャンセル

# モジュール6 責任ある生成AI

- ・生成AI利用時の「潜在的なリスク」とは？
- ・開発者は「潜在的なリスク」にどう対処すればよいのか？
- ・Azure OpenAI モデルにおける責任ある AI プラクティス
- ・コンテンツフィルターとは？
- ・コンテンツフィルターの設定例
- ・まとめ

# モジュール7まとめ

- ・生成AIアプリには「**潜在的なリスク**」（気づきにくいリスク）が含まれる可能性がある。これにより、思わぬ事故が発生する可能性がある
- ・生成AIアプリの開発者は、マイクロソフトの専門家が開発した「**Azure OpenAIモデルにおける責任あるAIプラクティス**」の4ステージ（識別・測定・緩和・運用）を実践することで「**潜在的なリスク**」に対処できる
- ・**コンテンツフィルター**は、Azure OpenAI Serviceに組み込まれたフィルタリングの仕組み。GPTやDALL-Eにおいて、不適切な入力や出力をフィルタリングできる（不適切な入力や出力が行われないようにできる）。デプロイに対して適用する

# 全体のまとめ: Azure AI Foundryとは？

- 生成AIアプリの開発に役立つプラットフォーム
- Azure OpenAI Serviceが提供するGPTなどの生成AIモデル、従来からAzureで提供してきた言語・音声・視覚などのAIサービス、Azure Machine Learningの「プロンプトフロー」など、Azureの主要なAIサービスやツールをこの一つのプラットフォームで扱うことができる

# 全体のまとめ: Azure AI Foundryの主な機能

- ・**プレイグラウンド**での生成AIモデルの動作確認（モジュール1で解説）
- ・Azure AI Servicesの視覚・音声・言語などの**AIサービス**の動作確認（モジュール1で解説）
- ・**モデルカタログ**でのモデルの検索とデプロイ（モジュール2で解説）
- ・**プロンプトフロー**を使った、モデルの制御フローの構築とデプロイ（モジュール3で解説）
- ・**データの追加**（RAGアーキテクチャの実装）（モジュール4で解説）
- ・**微調整**によるモデルのファインチューニング（モジュール5で解説）
- ・モデルの**パフォーマンス評価**（モジュール6で解説）
- ・**コンテンツフィルター**による、モデルの入出力データのフィルタリング（モジュール7で解説）