Azure OpenAI Service モデル

[アーティクル] • 2025/02/28

Azure OpenAI Service では、さまざまな機能と価格ポイントを備えた多様なモデルセットが利用されています。 モデルの可用性はリージョンとクラウドごとに異なります。 Azure Government モデルの可用性については、Azure Government の OpenAI Service に関するセクションを参照してください。

テーブルを展開する

モデル	説明
o シリーズ モデル	高度な問題解決、増強された集中力と能力を備えた推論モデル。
GPT-4o、GPT-4o mini、GPT-4 Turbo	最新の最も能力の高い Azure OpenAI モデルであり、テキストと画像の両方を入力として受け入れることができるマルチモーダル バージョンを備えています。
GPT-4o audio	低遅延、"音声入力、音声出力" の会話のやり取り、またはオーディオ生成をサポートする GPT-4o audio モデル。
GPT-4	GPT-3.5 を基に改善され、自然言語とコードを理解し、生成できるモデルのセット。
GPT-3.5	GPT-3 を基に改善され、自然言語とコードを理解し、生成できるモデルのセット。
埋め込み	テキストを数値ベクトル形式に変換して、テキストの類似性を促進できるモデルのセット。
DALL-E	自然言語からオリジナルの画像を生成できるモデルのシリーズ。
Whisper	音声を文字起こしして音声テキスト変換を翻訳できる一連のモデル。
テキスト読み上げ (プレビュー)	テキストを音声に合成できるプレビュー段階の一連のモデル。

o シリーズ モデル

Azure OpenAI の o * シリーズ モデルは、集中と能力を高めて推論と問題解決のタスクに取り組むために特に設計されています。 これらのモデルは、ユーザーの要求の処理と理解により多くの時間を費やし、これまでのイテレーションと比較して、科学、コーディング、数学などの分野で非常に強力になっています。

テーブルを展開する

モデル ID	説明	最大要求 (トークン)	トレーニング データ (最大)
`o3-mini` (2025-01-31)	最新の推論モデルであり、推論能力が強化されています。 - 構造化出力 - テキストのみの処理 - 機能/ツール **アクセスの要求**: **制限付きアクセス モデルの申請**	入力: 200,000 出力: 100,000	2023年10月
`o1` (2024-12-17)	o1 シリーズの中で最も能力の高いモデルで、推論能力が強化されています。 - 構造化出力 - テキスト、画像処理 - 機能/ツール **アクセスの要求**: **制限付きアクセス モデルの申請**	入力: 200,000 出力: 100,000	2023年10月
`o1-preview` (2024-09-12)	以前のプレビュー バージョン	入力: 128,000 出力: 32,768	2023年10月
`o1-mini` (2024-09-12)	o1 シリーズの中のより速く、よりコスト効率の高いオプションであり、速度を必要としリソース消費を削減する必要があるコーディング タスクに最適です。 グローバル標準デプロイが既定で使用できるようになりました。 現在、標準 (リージョン) のデプロイは、`o1-preview` の制限付きアクセス リリースの一部としてアクセス権を付与されたお客様のみが利用できます。	入力: 128,000 出力: 65,536	2023年10月

可用性

o3-mini と o1 にアクセスするには、登録が必要であり、Microsoft の適格性条件に基づいてアクセスが許可されます。 以前に o1-preview または o1 へのアクセスを申請して受け取ったお客様は、o シリーズの最新モデルの待機リストに自動的に追加されるため、再申請する必要はありません。

アクセスの要求: 制限付きアクセス モデルの申請

アクセス権が付与されたら、モデルごとにデプロイを作成する必要があります。

高度な o-series モデルの詳細については、「推論モデルの概要」を参照してください。

利用可能なリージョン

テーブルを展開する

モデル	リージョン
o3-mini	「モデル テーブル」を参照してください。
o1	「モデル テーブル」を参照してください。
o1-preview	「モデル テーブル」を参照してください。 このモデルを使用できるのは、元の制限付きアクセスの一部としてアクセス権を付与されたお客様に限られます
o1-mini	「モデル テーブル」を参照してください。

GPT-4o audio

GPT 4o audio モデルは GPT-4o モデル ファミリの一部であり、低遅延の "音声入力、音声出力" の会話のやり取りまたはオーディオ生成のいずれかをサポートします。

- GPT-4o real-time audio は、リアルタイムで低待機時間の会話を処理するように設計されており、サポート エージェント、アシスタント、翻訳者、およびユーザーとの応答性の高いやり取りを必要とするその他のユース ケースに最適です。 GPT-4o real-time audio の使用方法の詳細については、GPT-4o real-time audio のクイックスタートおよび GPT-4o audio の使用方法を参照してください。
- GPT-4o audio completion は、オーディオまたはテキスト プロンプトからオーディオを生成するように設計されており、オーディオ ブックやオーディオ コンテンツの生成、およびオーディオ生成を必要とするその他のユース ケースに最適です。 GPT-4o audio completion モデルでは、既存の /chat/completions API にオーディオ モダリティが導入されています。 GPT-4o audio completion の使用方法の詳細については、 audio 生成のクイックスタートを参照してください。

GPT-4o audio を使用するには、いずれかのサポートされているリージョンの Azure OpenAI リソースが必要です。

リソースが作成されたら、GPT-4o audio モデルをデプロイできます。

次の表では、最大要求トークン数とトレーニング データに関する詳細を確認できます。

テーブルを展開する

モデル ID	説明	最大要求 (トークン)	トレーニング データ (最大)
gpt-4o-mini-audio-preview (2024-12-17) **GPT-4o audio**	オーディオとテキスト生成向けの**オーディオ モデル**。	入力: 128,000 出力: 4,096	2023年10月
gpt-4o-mini-realtime-preview (2024-12-17) **GPT-4o audio**	リアルタイム オーディオ処理向けの**オーディオ モデル**。	入力: 128,000 出力: 4,096	2023年10月
gpt-4o-audio-preview (2024-12-17) **GPT-4o audio**	オーディオとテキスト生成向けの**オーディオ モデル**。	入力: 128,000 出力: 4,096	2023年10月
gpt-4o-realtime-preview (2024-12-17) **GPT-4o audio**	リアルタイム オーディオ処理向けの**オーディオ モデル**。	入力: 128,000 出力: 4,096	2023年10月
gpt-4o-realtime-preview (2024-10-01) **GPT-4o audio**	リアルタイム オーディオ処理向けの**オーディオ モデル**。	入力: 128,000 出力: 4,096	2023年10月

利用可能なリージョン

テーブルを展開する

モデル	リージョン
gpt-4o-mini-audio-preview	米国東部 2 (グローバル標準)
gpt-4o-mini-realtime-preview	米国東部 2 (グローバル標準) スウェーデン中部 (グローバル標準)
gpt-4o-audio-preview	米国東部 2 (グローバル標準) スウェーデン中部 (グローバル標準)
gpt-4o-realtime-preview	米国東部 2 (グローバル標準) スウェーデン中部 (グローバル標準)

すべてのリージョンで GPT-4o audio モデルの可用性を比較するには、モデルの表を参照してください。

GPT-4o および GPT-4 Turbo

GPT-4o は、テキストと画像を 1 つのモデルに統合し、複数のデータ型を同時に処理できるようにします。 このマルチモーダル アプローチにより、人間とコンピューターの対話における精度と応答性が向上します。 GPT-4o は、英語以外の言語とビジョン タスクで優れたパフォーマンスを提供しながら、英語のテキストとコーディング タスクにおいて GPT-4 Turbo に匹敵し、AI 機能の新しいベンチマークを設定します。

GPT-4o と GPT-4o mini のモデルにアクセスする方法

GPT-4o と GPT-4o mini は、**Standard** と **Global-Standard** のモデル デプロイで利用できます。

このモデルを利用できる サポート対象の標準リージョンまたはグローバル標準リージョンに、新しいリソースを作成するか既存のリソースを使用する必要があります。

リソースの作成が済んだ後、GPT-4o モデルをデプロイできます。 プログラムでデプロイを実行する場合、**モデル**の名前は次のとおりです。

- `gpt-4o` **バージョン** `2024-11-20`
- `gpt-4o` **バージョン** `2024-08-06`
- `gpt-4o` **バージョン** `2024-05-13`
- `gpt-4o-mini` **バージョン** `2024-07-18`

GPT-4 Turbo

GPT-4 Turbo は、大規模なマルチモーダル モデル (テキストまたは画像の入力を受け入れ、テキストを生成します) であり、OpenAI の以前のモデルよりも高い精度で困難な問題を解決できます。 GPT-3.5 Turbo や以前の GPT-4 モデルと同様に、GPT-4 Turbo はチャット用に最適化されており、従来の入力候補タスクでも適切に動作します。

GPT-4 Turbo の最新の GA リリースは次のとおりです。

- `gpt-4` **バージョン** `turbo-2024-04-09`

これは、次のプレビュー モデルに代わるものです。

- `gpt-4` **バージョン** `1106-Preview`
- `gpt-4` **バージョン** `0125-Preview`
- `gpt-4` **バージョン** `vision-preview`

OpenAI と Azure OpenAI GPT-4 Turbo GA モデルの違い

- OpenAI の最新の `0409` ターボ モデル バージョンでは、すべての推論要求に対して JSON モードと関数呼び出しがサポートされています。
- Azure OpenAI の最新の `turbo-2024-04-09` バージョンでは、現在、画像 (ビジョン) 入力による推論要求を行う場合、JSON モードと関数呼び出しの使用はサポートされていません。 テキスト ベース入力の要求 (`image_url` とインライン イメージがない要求) では、JSON モードと関数呼び出しがサポートされています。

gpt-4 vision-preview との違い

- Azure AI 固有の Vision 拡張機能と GPT-4 Turbo with Vision の統合は、 `gpt-4` **バージョン:** `turbo-2024-04-09` ではサポートされません。 これには、光学式文字認識 (OCR)、オブジェクト グラウンディング、ビデオ プロンプト、画像を含むデータの処理の改善が含まれます。

> **重要**
>
> 光学式文字認識 (OCR)、オブジェクト グラウンディング、ビデオ プロンプトなどのビジョン拡張機能のプレビュー機能は廃止され、 `gpt-4` バージョン: `vision-preview` が `turbo-2024-04-09` にアップグレードされると使用できなくなります。 現在これらのプレビュー機能のいずれかに依存している場合、このモデルの自動アップグレードは破壊的変更になります。

GPT-4 Turbo のプロビジョニングされたマネージド可用性

- `gpt-4` **バージョン** `turbo-2024-04-09` は、標準デプロイとプロビジョニングされたデプロイの両方で使用できます。 現在、このモデルのプロビジョニングされたバージョンでは、イメージ/ビジョン推論要求は**サポートされていません**。 このモデルのプロビジョニングされたデプロイでは、テキスト入力のみ受け入れます。 標準のモデル デプロイでは、テキストと画像/ビジョンの両方の推論要求を受け入れます。

GPT-4 Turbo with Vision GA のデプロイ

Azure AI Foundry ポータルから GA モデルをデプロイするには、 `GPT-4` を選択し、ドロップダウン メニューから `turbo-2024-04-09` バージョンを選択します。 `gpt-4-turbo-2024-04-09` モデルの既定のクォータは、GPT-4-Turbo の現在のクォータと同じになります。 リージョン別のクォータ制限を参照してください。

GPT-4

GPT-4 は、GPT-4 Turbo の前身です。 GPT-4 と GPT-4 Turbo のどちらのモデルも、基本モデル名は `gpt-4` です。 モデルのバージョンを調べると、GPT-4 モデルと Turbo モデルを区別できます。

- `gpt-4` **バージョン** `0314`
- `gpt-4` **バージョン** `0613`
- `gpt-4-32k` **バージョン** `0613`

各モデルでサポートされているトークン コンテキストの長さは、モデルの概要テーブルで確認できます。

GPT-4 モデルと GPT-4 Turbo モデル

- これらのモデルは Chat Completion API でのみ使用できます。

モデル バージョンを参照して、Azure OpenAI Service がモデル バージョンのアップグレードを処理する方法と、モデルを使用して GPT-4 デプロイのモデル バージョン設定を表示および構成する方法について説明します。

テーブルを展開する

モデル ID	説明	最大要求 (トークン)	トレーニング データ (最大)
`gpt-4o` (2024-11-20) **GPT-4o (Omni)**	**最新の大きい GA モデル** - 構造化出力 - テキスト、画像処理 - JSON モード - 並列関数呼び出し - 精度と応答性の向上 - GPT-4 Turbo with Vision と比較した英語のテキストおよびコーディング タスクの同等性 - 英語以外の言語とビジョン タスクでの優れたパフォーマンス。 - **クリエイティブ ライティング能力の向上**	入力: 128,000 出力: 16,384	2023年10月
`gpt-4o` (2024-08-06) **GPT-4o (Omni)**	- 構造化出力 - テキスト、画像処理 - JSON モード - 並列関数呼び出し	入力: 128,000 出力: 16,384	2023年10月

モデル ID	説明	最大要求 (トークン)	トレーニング データ (最大)
	- 精度と応答性の向上 - GPT-4 Turbo with Vision と比較した英語のテキストおよびコーディング タスクの同等性 - 英語以外の言語とビジョン タスクでの優れたパフォーマンス		
`gpt-4o-mini` (2024-07-18) **GPT-4o mini**	**最新の小さい GA モデル** - GPT-3.5 Turbo シリーズのモデルを置き換えるのに最適な、高速で安価で高機能のモデル。 - テキスト、画像処理 - JSON モード - 並列関数呼び出し	入力: 128,000 出力: 16,384	2023年10月
`gpt-4o` (2024-05-13) **GPT-4o (Omni)**	テキスト、画像処理 - JSON モード - 並列関数呼び出し - 精度と応答性の向上 - GPT-4 Turbo with Vision と比較した英語のテキストおよびコーディング タスクの同等性 - 英語以外の言語とビジョン タスクでの優れたパフォーマンス	入力: 128,000 出力: 4,096	2023年10月
`gpt-4` (turbo-2024-04-09) **GPT-4 Turbo with Vision**	**新しい GA モデル** - 以前のすべての GPT-4 プレビュー モデル (`vision-preview`、 `1106-Preview`、 `0125-Preview`) についての代替モデル。 現在、- **機能の使用の可否**は、入力方法とデプロイの種類によって異なります。	入力: 128,000 出力: 4,096	2023年12月
`gpt-4` (0125-Preview)* **GPT-4 Turbo プレビュー**	**プレビュー モデル** - 1106-Preview に代わるものです - コード生成パフォーマンスが向上 - モデルがタスクを完了しないケースを減らします。 - JSON モード - 並列関数呼び出し - 再現可能な出力 (プレビュー)	入力: 128,000 出力: 4,096	2023年12月
`gpt-4` (vision-preview) **GPT-4 Turbo with Vision Preview**	**プレビュー モデル** - テキストと画像の入力を受け入れます。 - 機能強化に対応します - JSON モード - 並列関数呼び出し - 再現可能な出力 (プレビュー)	入力: 128,000 出力: 4,096	2023 年 4 月
`gpt-4` (1106-Preview) **GPT-4 Turbo プレビュー**	**プレビュー モデル** - JSON モード - 並列関数呼び出し - 再現可能な出力 (プレビュー)	入力: 128,000 出力: 4,096	2023 年 4 月
`gpt-4-32k` (0613)	**古い GA モデル** - ツールによる基本的な関数呼び出し	32,768	2021 年 9 月
`gpt-4` (0613)	**古い GA モデル** - ツールによる基本的な関数呼び出し	8,192	2021 年 9 月
`gpt-4-32k` (0314)	**古い GA モデル** - 廃止に関する情報	32,768	2021 年 9 月
`gpt-4` (0314)	**古い GA モデル** - 廃止に関する情報	8,192	2021 年 9 月

⊗ 注意事項

運用環境でプレビュー モデルを使用することはおすすめしません。 プレビュー モデルのすべてのデプロイは、将来のプレビュー バージョンか最新の安定 GA バージョンにアップグレードされます。 プレビューに指定されたモデルは、標準の Azure OpenAI モデルのライフサイクルに従っていません。

- GPT-4 バージョン 0125-preview は、以前にバージョン 1106-preview としてリリースされた GPT-4 Turbo プレビューの更新バージョンです。
- GPT-4 バージョン 0125-preview は、gpt-4-1106-preview と比較して、コード生成などのタスクをより完全に完了します。 このため、タスクによっては、GPT-4-0125-preview が gpt-4-1106-preview と比較してより多くの出力を生成することがあります。 お客様には、新しいモデルの出力を比較することをお勧めします。 GPT-4-0125-preview では、英語以外の言語の UTF-8 処理に関する gpt-4-1106-preview のバグにも対処しています。

- GPT-4 バージョン `turbo-2024-04-09` は最新の GA リリースであり、`0125-Preview`、`1106-preview`、`vision-preview` に代わるものです。

GPT-3.5

GPT-3.5 モデルは、自然言語とコードを理解および生成できます。 GPT-3.5 ファミリで最も能力とコスト効率の高いモデルは GPT-3.5 Turbo です。これはチャット用に最適化されており、従来の補完タスクでも適切に動作します。 GPT-3.5 Turbo は、Chat Completions API で使用できます。 GPT-3.5 Turbo Instruct には、Chat Completions API の代わりに Completions API を使用する `text-davinci-003` のと同様の機能があります。 GPT-3.5 および GPT-3 のレガシ モデルよりも GPT-3.5 Turbo および GPT-3.5 Turbo Instruct を使用することをお勧めします。

テーブルを展開する

モデル ID	説明	最大要求 (トークン)	トレーニング データ (最大)
`gpt-35-turbo` (0125) **新規**	**最新の GA モデル** - JSON モード - 並列関数呼び出し - 再現可能な出力 (プレビュー) - 要求された形式での応答精度の向上。 - 英語以外の言語の関数呼び出しに対してテキスト エンコードの問題が発生していたバグの修正。	入力: 16,385 出力: 4,096	2021 年 9 月
`gpt-35-turbo` (1106)	**古い GA モデル** - JSON モード - 並列関数呼び出し - 再現可能な出力 (プレビュー)	入力: 16,385 出力: 4,096	2021 年 9 月
`gpt-35-turbo-instruct` (0914)	**入力候補エンドポイントのみ** - レガシ補完モデルの置き換え	4,097	2021 年 9 月
`gpt-35-turbo-16k` (0613)	**古い GA モデル** - ツールによる基本的な関数呼び出し	16,384	2021 年 9 月
`gpt-35-turbo` (0613)	**古い GA モデル** - ツールによる基本的な関数呼び出し	4,096	2021 年 9 月
`gpt-35-turbo` [1] (0301)	**古い GA モデル** - 廃止に関する情報	4,096	2021 年 9 月

GPT-3.5 Turbo と Chat Completions API の使用方法について詳しくは、詳細なハウツーをご覧ください。

[1] このモデルは、> 4096 個のトークン要求を受け入れます。 モデルの新しいバージョンは 4,096 個のトークンに制限されるため、4,096 個の入力トークンの制限を超えないようにすることをお勧めします。 このモデルで 4,096 個の入力トークンを超えたときに問題が発生した場合、この構成は公式にはサポートされていません。

埋め込み

`text-embedding-3-large` は、最新かつ最も高性能の埋め込みモデルです。 埋め込みモデル間でアップグレードすることはできません。 `text-embedding-ada-002` の使用から `text-embedding-3-large` の使用に移行するには、新しい埋め込みを生成する必要があります。

- `text-embedding-3-large`
- `text-embedding-3-small`
- `text-embedding-ada-002`

OpenAI の報告によると、テストでは、大規模と小規模の第 3 世代埋め込みモデルのいずれも、MIRACL ベンチマークで多言語検索の平均パフォーマンスが向上しており、さらに MTEB ベンチマークで英語タスクのパフォーマンスを維持しています。

テーブルを展開する

評価ベンチマーク	`text-embedding-ada-002`	`text-embedding-3-small`	`text-embedding-3-large`
MIRACL 平均	31.4	44.0	54.9
MTEB 平均	61.0	62.3	64.6

第 3 世代の埋め込みモデルは、新しい `dimensions` パラメーターを使った埋め込みのサイズ削減をサポートしています。 通常、埋め込みが大きくなると、コンピューティング、メモリ、ストレージの観点からコストが高くなります。 ディメンション数を調整できるので、全体的なコストとパフォーマンスをより詳細に制御できます。 `dimensions` パラメーターは OpenAI 1.x Python ライブラリのすべてのバージョンでサポートされているわけではありません。このパラメーターを利用するには、最新バージョンの `pip install openai --upgrade` にアップグレードすることをお勧めします。

OpenAI の MTEB ベンチマーク テストにより、第 3 世代モデルのディメンションは、`text-embeddings-ada-002` 1,536 ディメンション未満に減らした場合でも、パフォーマンスはわずかに優れていることがわかりました。

DALL-E

DALL-E モデルは、ユーザーが提供するテキスト プロンプトから画像を生成します。 DALL-E 3 は、REST API との併用で一般提供されています。 クライアント SDK を使用する DALL-E 2 と DALL-E 3 は、プレビュー段階です。

Whisper

Whisper モデルは、音声テキスト変換に使用できます。

Azure AI Speech バッチ文字起こし API を使用して、ささやきモデルを使用することもできます。 Azure AI 音声と Azure OpenAI Service の使い分けの詳細については、「Whisper モデルとは」を参照してください。

テキスト読み上げ (プレビュー)

現在プレビュー段階にある OpenAI テキスト読み上げモデルを使って、テキストを音声に合成できます。

Azure AI 音声経由で OpenAI テキスト読み上げの音声を使うこともできます。 詳細については、Azure OpenAI Service または Azure AI 音声経由の OpenAI テキスト読み上げ音声のガイドを参照してください。

モデルの概要テーブルとリージョンの可用性

デプロイの種類別モデル

Azure OpenAI では、お客様はビジネスと使用のパターンに合ったホスティング構造を選択できます。 このサービスで提供されるデプロイの 2 つの主要な種類は、以下のとおりです。

- **標準**にはグローバル デプロイ オプションが用意されており、トラフィックをグローバルにルーティングしてスループットを向上させます。
- **プロビジョニング済み**はグローバル デプロイ オプションでも提供されており、お客様はプロビジョニングされたスループット ユニットを購入して Azure グローバル インフラストラクチャ全体にデプロイできます。

実行される推論操作はどのデプロイもまったく同じですが、課金、スケール、パフォーマンスは大きく異なります。 Azure OpenAI のデプロイの種類の詳細については、デプロイの種類に関するガイドを参照してください。

グローバル標準

Global-Standard モデルの提供状況

テーブルを展開する

リージョン	o3-mini、2025-01-31	o1、2024-12-17	o1-preview、2024-09-12	o1-mini、2024-09-12	gpt-4o、2024 年 5 月 13 日	gpt-4o、2024-08-06	gpt-4o、2024-11-20	gpt-4o-mini、2024-07-18	gpt-4o-realtime-preview、2024-12-17	gpt-4o-realtime-preview、2024-10-01	gpt-4o-audio-preview、2024-12-17	gpt-4o-mini-realtime-preview、2024-12-17	gpt-4o-mini-audio-preview、2024-12-17	t 2 (
australiaeast	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
brazilsouth	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	

リージョン	o3-mini、2025-01-31	o1、2024-12-17	o1-preview、2024-09-12	o1-mini、2024-09-12	gpt-4o、2024 年 5 月 13 日	gpt-4o、2024-08-06	gpt-4o、2024-11-20	gpt-4o-mini、2024-07-18	gpt-4o-realtime-preview、2024-12-17	gpt-4o-realtime-preview、2024-10-01	gpt-4o-audio-preview、2024-12-17	gpt-4o-mini-realtime-preview、2024-12-17	gpt-4o-mini-audio-preview、2024-12-17	t 2 (
canadaeast	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
eastus	-	-	✅	✅	✅	✅	✅	✅	-	-	-	-	-	
eastus2	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	
francecentral	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
germanywestcentral	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
japaneast	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
koreacentral	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
northcentralus	-	-	✅	✅	✅	✅	✅	✅	-	-	-	-	-	
norwayeast	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
polandcentral	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
southafricanorth	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
southcentralus	-	-	✅	✅	✅	✅	✅	✅	-	-	-	-	-	
southindia	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
spaincentral	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
swedencentral	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	-	
switzerlandnorth	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
uaenorth	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
uksouth	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
westeurope	-	-	-	-	✅	✅	-	✅	-	-	-	-	-	
westus	-	-	✅	✅	✅	✅	✅	✅	-	-	-	-	-	
westus3	-	-	✅	✅	✅	✅	✅	✅	-	-	-	-	-	

注意

ほとんどの o シリーズモデルは、制限付きアクセスです。 制限付きアクセス モデルの申請 のページから、アクセスを申請してください。 現在、`o1-mini` は、グローバル標準デプロイのすべてのお客様が利用できます。

一部のお客様には、`o1-preview` 制限付きアクセス リリースの一部として、`o1-mini` への標準 (リージョン) デプロイ アクセスが付与されています。 現時点で、`o1-mini` 標準 (リージョン) デプロイへのアクセスは拡大されていません。

この表には、微調整のリージョン別の提供状況は含まれていません。 この情報については、微調整についてのセクションをご覧ください。

エンドポイント別の標準モデル

チャット入力候補

チャット入力候補

テーブルを展開する

リージョン	o1-preview、2024-09-12	o1-mini、2024-09-12	gpt-4o、2024 年 5 月 13 日	gpt-4o、2024-08-06	gpt-4o-mini、2024-07-18	gpt-4、0613	gpt-4、1106-Preview	gpt-4、0125-Preview	gpt-4、vision-preview	gpt-4、turbo-2024-04-09	gpt-4-32k、0613	gpt-35-turbo、0301	gpt-35-turbo、0613	gpt-35-turbo、1106	gpt-35-turbo、0125
australiaeast	-	-	-	-	-	✅	✅	-	✅	-	✅	-	✅	✅	✅
canadaeast	-	-	-	-	-	✅	✅	-	-	-	✅	-	✅	✅	✅
eastus	✅	✅	✅	✅	✅	✅	-	✅	-	✅	-	✅	✅	-	✅
eastus2	✅	✅	✅	✅	✅	✅	✅	-	-	✅	-	-	✅	-	✅
francecentral	-	-	-	-	-	✅	✅	-	-	-	✅	✅	✅	✅	-
japaneast	-	-	-	-	-	-	-	-	✅	-	-	-	✅	-	✅
northcentralus	✅	✅	✅	✅	✅	✅	-	✅	-	✅	-	-	✅	-	✅
norwayeast	-	-	-	-	-	-	✅	-	-	-	-	-	-	-	-
southcentralus	✅	✅	✅	✅	✅	-	-	✅	-	✅	-	✅	-	-	✅
southindia	-	-	-	-	-	-	✅	-	-	-	-	-	-	✅	✅
swedencentral	✅	✅	✅	✅	✅	✅	✅	-	✅	✅	✅	-	✅	✅	-
switzerlandnorth	-	-	-	-	-	✅	-	-	✅	-	✅	-	✅	-	✅
uksouth	-	-	-	-	-	-	✅	✅	-	-	-	✅	✅	✅	✅
westeurope	-	-	-	-	-	-	-	-	-	-	-	✅	-	-	-
westus	✅	✅	✅	✅	✅	-	✅	-	✅	✅	-	-	-	✅	✅
westus3	✅	✅	✅	✅	✅	-	✅	-	-	✅	-	-	-	-	✅

注意

ほとんどの o シリーズモデルは、制限付きアクセスです。 制限付きアクセス モデルの申請 のページから、アクセスを申請してください。 現在、`o1-mini` は、グローバル標準デプロイのすべてのお客様が利用できます。

一部のお客様には、`o1-preview` 制限付きアクセス リリースの一部として、`o1-mini` への標準 (リージョン) デプロイ アクセスが付与されています。 現時点で、`o1-mini` 標準 (リージョン) デプロイへのアクセスは拡大されていません。

GPT-4 および GPT-4 Turbo モデルの可用性

お客様のアクセスを選択する

Azure OpenAI のすべてのお客様が利用できる上記のリージョンに加え、一部の既存のお客様には、その他のリージョンでの GPT-4 のバージョンへのアクセスが許可されています。

テーブルを展開する

モデル	リージョン
`gpt-4` (0314) `gpt-4-32k` (0314)	米国東部 フランス中部 米国中南部 英国南部
`gpt-4` (0613) `gpt-4-32k` (0613)	米国東部 米国東部 2 東日本 英国南部

GPT-3.5 モデル

モデル バージョンを参照して、Azure OpenAI Service がモデル バージョンのアップグレードを処理する方法と、モデルを使用して GPT-3.5 Turbo デプロイのモデル バージョン設定を表示および構成する方法について説明します。

モデルの微調整

注意

`gpt-35-turbo` - このモデルの微調整はリージョンのサブセットに限定され、基本モデルが使用可能なすべてのリージョンで使用できるわけではありません。

Azure OpenAI モデルを Azure AI Foundry プロジェクトで使用するか、プロジェクトの外部で使用するかによって、微調整をサポートするリージョンは異なります。

テーブルを展開する

モデル ID	微調整リージョン	最大要求 (トークン)	トレーニング データ (最大)
`gpt-35-turbo` (0613)	米国東部 2 米国中北部 スウェーデン中部 スイス西部	4,096	2021 年 9 月
`gpt-35-turbo` (1106)	米国東部 2 米国中北部 スウェーデン中部 スイス西部	入力: 16,385 出力: 4,096	2021 年 9 月
`gpt-35-turbo` (0125)	米国東部 2 米国中北部 スウェーデン中部 スイス西部	16,385	2021 年 9 月
`gpt-4` (0613) [1]	米国中北部 スウェーデン中部	8192	2021 年 9 月
`gpt-4o-mini` (2024-07-18)	米国中北部 スウェーデン中部	入力: 128,000 出力: 16,384 トレーニング例のコンテキスト長: 64,536	2023年10月
`gpt-4o` (2024-08-06)	米国東部 2 米国中北部 スウェーデン中部	入力: 128,000 出力: 16,384 トレーニング例のコンテキスト長: 64,536	2023年10月

[1] GPT-4 は現在パブリック プレビューの段階です。

アシスタント (プレビュー)

アシスタントの場合は、サポートされているモデルとサポートされているリージョンの組み合わせが必要です。 特定のツールと機能には最新モデルが必要です。 次のモデルは、Assistants API、SDK、Azure AI Foundry で使用できます。 次の表は、従量課金制に関するものです。 プロビジョニング済みスループット ユニット (PTU) の可用性については、プロビジョニング済みスループットに関する記事を参照してください。 一覧で示されているモデルとリージョンは、Assistants v1 と v2 の両方で使用できます。 以下に示すリージョンでサポートされている場合に、グローバル標準モデルを使用できます。

テーブルを展開する

リージョン	gpt-4o、2024 年 5 月 13 日	gpt-4o、2024-08-06	gpt-4o-mini、2024-07-18	gpt-4、0613	gpt-4、1106-Preview	gpt-4、0125-Preview	gpt-4、turbo-2024-04-09	gpt-4-32k、0613	gpt-35-turbo、0613	gpt-35-turbo、1106	gpt-35-turbo、0125	gpt-35-turbo-16k、0613
australiaeast	-	-	-	✅	✅	-	-	✅	✅	✅	✅	✅
eastus	✅	✅	✅	-	-	✅	✅	-	✅	-	✅	✅
eastus2	✅	✅	✅	-	✅	-	✅	-	✅	-	✅	✅

リージョン	gpt-4o、2024 年 5 月 13 日	gpt-4o、2024-08-06	gpt-4o-mini、2024-07-18	gpt-4、0613	gpt-4、1106-Preview	gpt-4、0125-Preview	gpt-4、turbo-2024-04-09	gpt-4-32k、0613	gpt-35-turbo、0613	gpt-35-turbo、1106	gpt-35-turbo、0125	gpt-35-turbo-16k、0613
francecentral	-	-	-	✅	✅	-	-	✅	✅	✅	-	✅
japaneast	-	-	-	-	-	-	-	-	✅	-	✅	✅
norwayeast	-	-	-	-	✅	-	-	-	-	-	-	-
southindia	-	-	-	-	✅	-	-	-	-	✅	✅	-
swedencentral	✅	✅	✅	✅	✅	-	✅	✅	✅	✅	-	✅
uksouth	-	-	-	-	✅	✅	-	-	✅	✅	✅	✅
westus	✅	✅	✅	-	✅	-	✅	-	-	✅	✅	-
westus3	✅	✅	✅	-	✅	-	✅	-	-	-	✅	-

モデルの廃止

モデルの廃止に関する最新情報については、モデル廃止ガイドに関する記事をご覧ください。

次のステップ

- モデルの廃止と非推奨
- Azure OpenAI モデルの操作に関する詳細を確認する
- Azure OpenAI の詳細についてご覧ください
- Azure OpenAI モデルの微調整に関する詳細を確認する

フィードバック

このページはお役に立ちましたか? Yes いいえ

製品フィードバックの提供 | Microsoft Q&A でヘルプを表示する