

Azure AI Foundry における Azure OpenAI モデル

2025/06/16

Azure OpenAI 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を利用しています。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 OpenAI サービス](#)を参照してください。

[🔗 テーブルを展開する](#)

モデル	説明
codex-mini	o4-mini の微調整されたバージョン。
GPT-4.1 シリーズ	Azure OpenAI からの最新モデル リリース
model-router	特定のプロンプトに応答するために、基になる一連のチャット モデルからインテリジェントに選択するモデル。
コンピューター使用プレビュー	Responses API コンピューター使用ツールで使用するためにトレーニングされた実験モデル。
GPT-4.5 プレビュー	多様なテキストと画像のタスクに優れた最新の GPT モデル。
o シリーズ モデル	高度な問題解決とフォーカスと機能の向上による 推論モデル 。
GPT-4o & GPT-4o ミニ & GPT-4 ターボ	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-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。
画像の生成	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。
オーディオ	音声テキスト変換、翻訳、およびテキスト読み上げのための一連のモデル。GPT-4o オーディオ モデルでは、低待機時間、"音声入力、音声出力" の会話操作またはオーディオ生成がサポートされます。

GPT 4.1 シリーズ

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-4.1 (2025-04-14)	モデルの 表 を参照してください。
gpt-4.1-nano (2025-04-14)	モデルの 表 を参照してください。
gpt-4.1-mini (2025-04-14)	モデルの 表 を参照してください。

資格

[🔗 テーブルを展開する](#)

モデル ID	説明	コンテキスト ウィンドウ	最大出力トークン数	トレーニング データ (最大)
gpt-4.1 (2025-04-14)	- テキストと画像の入力 - テキスト出力 - Chat completions API - レスポンスAPI - ストリーミング - 関数呼び出し 構造化された出力 (チャットの入力候補)	- 1,047,576 - 128,000 (プロビジョニング済みのマネージド デプロイ)	32,768	2024 年 5 月 31 日
gpt-4.1-nano (2025-04-14)	- テキストと画像の入力 - テキスト出力	- 1,047,576 - 128,000 (プロビジョニング済みのマネージド デ	32,768	2024 年 5 月 31 日

モデル ID	説明	コンテキスト ウィンドウ	最大出力トークン数	トレーニング データ (最大)
	- Chat completions API - レスポンスAPI - ストリーミング - 関数呼び出し 構造化された出力 (チャットの入力候補)	プロイ)		
gpt-4.1-mini (2025-04-14)	- テキストと画像の入力 - テキスト出力 - Chat completions API - レスポンスAPI - ストリーミング - 関数呼び出し 構造化された出力 (チャットの入力候補)	- 1,047,576 - 128,000 (プロビジョニング済みのマネージド デプロイ)	32,768	2024 年 5 月 31 日

モデルルーター

特定のプロンプトに応答するために、基になる一連のチャット モデルからインテリジェントに選択するモデル。

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
model-router (2025-05-19)	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)

資格

[🔗 テーブルを展開する](#)

モデル ID	説明	コンテキスト ウィンドウ	最大出力トークン数	トレーニング データ (最大)
model-router (2025-05-19)	特定のプロンプトに応答するために、基になる一連のチャット モデルからインテリジェントに選択するモデル。	200,000*	32768 (GPT 4.1 シリーズ) 100 K (o4-mini)	2024 年 5 月 31 日

*より大きなコンテキストウィンドウは、基になるモデルの 一部 と互換性があります。つまり、より大きなコンテキストを持つ API 呼び出しは、プロンプトが適切なモデルにルーティングされた場合にのみ成功し、それ以外の場合は呼び出しは失敗します。

コンピューター利用プレビュー

[Responses API](#) コンピューター使用ツールで使用するためにトレーニングされた実験モデル。 サード パーティ製ライブラリと組み合わせて使用すると、現在の環境のスクリーンショットからコンテキストを取得しながら、モデルでマウスとキーボードの入力を制御できます。

⊗ **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 バージョンが最新の安定 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。

可用性

[computer-use-preview](#) 登録へのアクセスが必要であり、Microsoft の資格条件に基づいてアクセス権が付与されます。 他の制限付きアクセス モデルにアクセスできるお客様は、引き続きこのモデルへのアクセスを要求する必要があります。

アクセスの要求: [制限付きアクセス モデル アプリケーション](#)[computer-use-preview](#)

アクセス権が付与されたら、モデルのデプロイを作成する必要があります。

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
computer-use-preview	モデルの 表 を参照してください。

資格

🔗 テーブルを展開する

モデル ID	説明	コンテキスト ウィンドウ	最大出力トークン数	トレーニング データ (最大)
computer-use-preview (2025-03-11)	Responses API コンピューター使用ツールで使用するための特殊なモデル - ツール - ストリーミング - Text(入力/出力) - 画像(入力)	8,192	1,024	2023年10月

GPT-4.5 プレビュー

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-4.5-preview	モデルの 表 を参照してください。

資格

🔗 テーブルを展開する

モデル ID	説明	コンテキスト ウィンドウ	最大出力トークン数	トレーニング データ (最大)
gpt-4.5-preview (2025-02-27) GPT-4.5 プレビュー	GPT 4.1 は、このモデルの推奨される代替品です。 多様なテキストタスクと画像タスクに優れています。 - 構造化された出力 - プロンプト キャッシュ - ツール - ストリーミング - テキスト (入力と出力) - 画像(入力)	128,000	16,384	2023年10月

ⓘ 注意

モデルがそれ自体に関する質問に答えられないことが予想される動作です。 モデルのトレーニング データのナレッジ カットオフがいつであるか、またはモデルに関するその他の詳細を知りたい場合は、上記のモデル ドキュメントを参照する必要があります。

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)
codex-mini (2025-05-16)	o4-mini の微調整されたバージョン。 - Responses API - 構造化出力 - テキスト、画像処理 - 関数/ツール 機能の完全な概要	入力: 200,000 出力: 100,000	2024 年 5 月 31 日
o3-pro (2025-06-10)	- Responses API - 構造化出力 - テキスト、画像処理 - 関数/ツール 機能の完全な概要	入力: 200,000 出力: 100,000	2024 年 5 月 31 日
o4-mini (2025-04-16)	- 新しい 推論モデル、 強化された推論能力を提供します。 - チャット完了API - Responses API - 構造化出力 - テキスト、画像処理 - 関数/ツール 機能の完全な概要	入力: 200,000 出力: 100,000	2024 年 5 月 31 日
o3 (2025-04-16)	- 新しい 推論モデル、 強化された推論能力を提供します。 - チャット完了API - Responses API - 構造化出力 - テキスト、画像処理 - 関数/ツール/並列ツールの呼び出し 機能の完全な概要	入力: 200,000 出力: 100,000	2024 年 5 月 31 日
o3-mini (2025-01-31)	- 推論能力の強化。 - 構造化出力 - テキストのみの処理 - 関数/ツール	入力: 200,000 出力: 100,000	2023年10月
o1 (2024-12-17)	- 推論能力の強化。 - 構造化出力 - テキスト、画像処理 - 関数/ツール	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え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

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
codex-mini	米国東部 2 およびスウェーデン中部 (グローバル標準)
o3-pro	米国東部 2 およびスウェーデン中部 (グローバル標準)
o4-mini	モデルの 表 を参照してください。

モデル	リージョン
o3	モデルの 表を参照してください。
o3-mini	モデルの 表を参照してください。
o1	モデルの 表を参照してください。
o1-preview	モデルの 表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 および GPT-4o mini は、 **標準** および **グローバル標準** モデルのデプロイに使用できます。

モデルが使用可能な **サポートされている標準**または**グローバル標準**リージョンで、既存のリソースを **作成**または使用する必要があります。

リソースが作成されたら、GPT-4o モデルを **デプロイ** できます。 プログラムによるデプロイを実行する場合、 **モデル** 名は次のようになります。

- 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

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 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 (オムニ)	最新の大規模 GA モデル - 構造化出力	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)
	<ul style="list-style-type: none">- 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。- クリエイティブな文章力の強化		
gpt-4o (2024-08-06) GPT-4o (オムニ)	<ul style="list-style-type: none">- 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	入力: 128,000 出力: 16,384	2023年10月
gpt-4o-mini (2024-07-18) GPT-4o mini	最新の小型 GA モデル <ul style="list-style-type: none">- 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 (オムニ)	<ul style="list-style-type: none">- 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 <ul style="list-style-type: none">- 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、機能の可用性は、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-4-32k (0613)	古い GA モデル <ul style="list-style-type: none">- 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 <ul style="list-style-type: none">- 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 <ul style="list-style-type: none">- 退職情報	32,768	2021 年 9 月
gpt-4 (0314)	古い GA モデル <ul style="list-style-type: none">- 退職情報	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 バージョンが最新の安定 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-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。従来の GPT-3.5 および GPT-3 **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 <ul style="list-style-type: none">- JSON モード- 並列関数呼び出し	入力: 16,385 出力: 4,096	2021 年 9 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)
	- 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。		
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)	コンプリーションエンドポイントのみ - レガシー補完モデル の代替	4,097	2021 年 9 月

GPT-3.5 Turbo と Chat Completions API を操作する方法の詳細については、[詳細なハウツー](#)を確認してください。

¹ このモデルは、4,096 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

テストでは、OpenAI は、大規模および小規模の第3世代埋め込みモデルが、[MIRACL](#) ベンチマークを使用した平均的な多言語検索性能を向上させ、同時に [MTEB](#) ベンチマークを使用して英語のタスクでの性能を維持していると報告しています。

[🔗 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。

画像生成モデル

画像生成モデルは、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。 GPT-image-1 は、制限付きアクセスパブリック プレビュー段階です。 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。

可用性

gpt-image-1 登録へのアクセスが必要であり、Microsoft の資格条件に基づいてアクセス権が付与されます。 他の制限付きアクセス モデルにアクセスできるお客様は、引き続きこのモデルへのアクセスを要求する必要があります。

アクセスの要求: [制限付きアクセス モデル アプリケーション](#) **gpt-image-1**

アクセス権が付与されたら、モデルのデプロイを作成する必要があります。

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
dall-e-3	米国東部 オーストラリア東部 スウェーデン中部
gpt-image-1	米国西部 3 (グローバル標準) アラブ首長国連邦北部 (グローバル標準)

ビデオ生成モデル

Soraは、テキストからの指示で現実的で想像力豊かなビデオシーンを作成できる、OpenAIによるAIモデルです。 ソラはパブリック プレビュー 段階です。

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
sora	米国東部 2

オーディオ モデル

Azure OpenAI のオーディオ モデルは、 realtime、 completions、 audio API を介して使用できます。

GPT-4o オーディオ モデル

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。

⊗ 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 パ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。

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 およびテキスト生成用のオーディオ 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のためのオーディオ 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 およびテキスト生成用のオーディオ 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のためのオーディオ 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のためのオーディオ モデル 。	入力: 128,000 出力: 4,096	2023年10月

すべてのリージョンでの GPT-4o オーディオ 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。

Audio API

/audio API を介したオーディオ モデルは、音声テキスト変換、翻訳、テキスト読み上げに使用できます。

音声テキスト変換モデル

🔗 テーブルを展開する

モデル ID	説明	最大要求数 (オーディオ ファイル サイズ)
whisper	汎用音声認識モデル。	25 MB
gpt-4o-transcribe	GPT-4o を利用した音声テキスト変換。	25 MB
gpt-4o-mini-transcribe	GPT-4o ミニを搭載した音声テキスト変換。	25 MB

音声翻訳モデル

🔗 テーブルを展開する

モデル ID	説明	最大要求数 (オーディオ ファイル サイズ)
whisper	汎用音声認識モデル。	25 MB

テキスト・トゥ・スピーチモデル (プレビュー)

🔗 テーブルを展開する

モデル ID	説明
tts	音声合成の速度を最適化。
tts-hd	品質向上のために最適化された Text to Speech。
gpt-4o-mini-tts	GPT-4o ミニを搭載したテキスト読み上げモデル。 スタイルやトーンで話すように音声をガイドできます。

詳細については、この記事の [オーディオ モデルリージョンの可用性](#) 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。

- **Standard** 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** にはグローバル デプロイ オプションも用意されており、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 デプロイの種類の詳細については、[デプロイの種類ガイド](#)を参照してください。

グローバル標準
<h3>Global-Standard モデルの提供状況</h3> <p>🔗 テーブルを展開する</p>

地域	o3- pro、 2025- 06-10	codex- mini、 2025- 05-16	model- router、 2025- 05-19	o3、 2025- 04-16	o4- mini、 2025- 04-16	gpt- image- 1、 2025- 04-15	gpt- 4.1、 2025- 04-14	gpt- 4.1- nano、 2025- 04-14	gpt- 4.1- mini、 2025- 04-14	computer- use- preview、 2025-03- 11	gpt-4.5- preview、 2025-02- 27	o3- mini、 2025- 01-31	o1、 2024- 12-17	o1- preview、 2024-09- 12	o1- mini、 2024- 09-12
オーストラリア イースト	-	-	-	-	-	-	✓	✓	✓	-	-	✓	-	-	-
ブラジル南部	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
カナダ東部	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
イーストアス	-	-	-	-	-	-	✓	-	✓	-	-	✓	✓	✓	✓
eastus2	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓
francecentral	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
ドイツ中西部	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
italynorth	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
japaneast	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
コリアセントラ ル	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
ノースセントラ ルUS	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	✓	✓
ノルウェー・イ ースト	-	-	-	-	-	-	✓	-	✓	-	-	✓	✓	-	-
polandcentral	-	-	-	-	-	-	✓	-	✓	-	-	✓	✓	-	-
southafricanorth	-	-	-	-	-	-	✓	-	✓	-	-	✓	✓	-	-
サウスセントラ ル	-	-	-	-	-	-	✓	-	✓	-	-	✓	✓	✓	✓
南インド	-	-	-	-	-	-	-	✓	✓	✓	-	✓	✓	-	-
spaincentral	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
swedencentral	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓
スイスノース	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
uaenorth	-	-	-	-	-	✓	✓	✓	✓	-	-	✓	✓	-	-
ウクサウス	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
西ヨーロッパ	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	-	-
ウェストユーエ ス	-	-	-	-	-	-	✓	✓	✓	-	-	✓	✓	✓	✓
westus3	-	-	-	-	-	✓	✓	✓	✓	-	-	✓	✓	✓	✓

① 注意

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-mini 制限付きアクセス リリースの一部として、o1-preview 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、地域別の利用可能性に関する詳細な情報は含まれていません。この情報については、[微調整のセクション](#) を参照してください。

エンドポイント別の標準デプロイ (リージョン) モデル

チャットの完了

チャット入力候補

🔗 テーブルを展開する

地域	o1-preview、 2024-09-12	o1-mini、 2024-09-12	gpt-4o、2024-05-13	gpt-4o、2024-11-20	gpt-4o、2024-08-06	gpt-4o-mini、2024-07-18	gpt-4-turbo-2024-04-09	gpt-35-turbo、1106	gpt-35-turbo、0125
オーストラリアー ースト	-	-	-	✔	-	-	-	✔	✔
カナダ東部	-	-	-	✔	-	-	-	✔	✔
イーストアス	✔	✔	✔	✔	✔	✔	✔	-	✔
eastus2	✔	✔	✔	✔	✔	✔	✔	-	✔
francecentral	-	-	-	✔	-	-	-	✔	✔
japaneast	-	-	-	✔	-	-	-	-	✔
ノースセントラ ルUS	✔	✔	✔	✔	✔	✔	✔	-	✔
ノルウェー・イ ースト	-	-	-	✔	-	-	-	-	-
サウスセントラ ル	✔	✔	✔	✔	✔	✔	✔	-	✔
南インド	-	-	-	✔	-	-	-	✔	✔
swedencentral	✔	✔	✔	✔	✔	✔	✔	✔	✔
スイスノース	-	-	-	✔	-	-	-	-	✔
ウクサウス	-	-	-	✔	-	-	-	✔	✔
西ヨーロッパ	-	-	-	-	-	-	-	-	✔
ウェストユーエ ス	✔	✔	✔	✔	✔	✔	✔	✔	✔
westus3	✔	✔	✔	✔	✔	✔	✔	-	✔

① 注意

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-mini 制限付きアクセス リリースの一部として、o1-preview 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

モデル

リージョン

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 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

① 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	標準トレーニングリージョン	グローバル トレーニング (プレビュー)	最大要求 (トークン)	トレーニング データ (最大)	モダリティ
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	テキスト間
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	テキスト間
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 トレーニング用コンテキストの例 の長さ: 65,536	2023年10月	テキスト間
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 トレーニング用コンテキストの例 の長さ: 65,536	2023年10月	テキストおよび Vision からテキスト
gpt-4.1 (2025-04-14)	米国中北部 スウェーデン中部	✅	入力: 128,000 出力: 16,384 トレーニング用コンテキストの例 の長さ: 65,536	2024 年 5 月	テキストおよび Vision からテキスト
gpt-4.1-mini (2025-04-14)	米国中北部 スウェーデン中部	✅	入力: 128,000 出力: 16,384 トレーニング用コンテキストの例 の長さ: 65,536	2024 年 5 月	テキスト間
gpt-4.1-nano (2025-04-14)	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: 32,768	2024 年 5 月	テキスト間
o4-mini (2025-04-16)	米国東部 2 スウェーデン中部	-	入力: 128,000 出力: 16,384 トレーニング用コンテキストの例 の長さ: 65,536	2024 年 5 月	テキスト間

① 注意

グローバル トレーニング (パブリック プレビュー) では、トークンごとの [より手頃な価格](#) のトレーニングが提供されますが、[データ所在地](#) は提供されません。現在、次のリージョンの Azure OpenAI リソースで使用でき、近日中にさらに多くのリージョンが提供されます。

- オーストラリア東部
- ブラジル南部
- 米国東部
- 米国東部 2
- フランス中部
- ドイツ中西部
- イタリア北部
- 東日本 (ビジョンサポートなし)
- 韓国中部
- 米国中北部
- ノルウェー東部
- ポーランド中部
- 東南アジア
- 南アフリカ北部
- スペイン中部
- スウェーデン中部
- スイス西部
- スイス北部
- 英国南部
- 米国西部
- 米国西部 3

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、標準のデプロイ用です。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、[プロビジョニングされたスループット](#)に関するページ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。[グローバル標準モデル](#)は、以下に示すリージョンでサポートされている場合に使用できます。

[🔗 テーブルを展開する](#)

地域	gpt-4o,2024-05-13	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
オーストラリア イースト	-	-	-	✔	✔	-	-	✔	✔	✔	✔	✔
イーストアス	✔	✔	✔	-	-	✔	✔	-	✔	-	✔	✔
eastus2	✔	✔	✔	-	✔	-	✔	-	✔	-	✔	✔
francecentral	-	-	-	✔	✔	-	-	✔	✔	✔	-	✔
japaneast	-	-	-	-	-	-	-	-	✔	-	✔	✔
ノルウェー・イースト	-	-	-	-	✔	-	-	-	-	-	-	-
南インド	-	-	-	-	✔	-	-	-	-	✔	✔	-
swedencentral	✔	✔	✔	✔	✔	-	✔	✔	✔	✔	-	✔
ウクサウス	-	-	-	-	✔	✔	-	-	✔	✔	✔	✔
ウェストユーエス	✔	✔	✔	-	✔	-	✔	-	-	✔	✔	-

地域	gpt-4o,2024-05-13	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
westus3	✔	✔	✔	-	✔	-	✔	-	-	-	✔	-

モデルの廃止

モデルの提供終了に関する最新情報については、モデル提供 [終了ガイド](#)を参照してください。

次のステップ

- モデルの廃止と非推奨
- [Azure OpenAI モデルの操作の詳細](#)
- [Azure OpenAI の詳細](#)
- [Azure OpenAI モデルの微調整の詳細を確認する](#)