

# Azure OpenAI Service モデル

[アーティクル] • 2024/10/04

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
<a href="#">o1-preview</a> と <a href="#">o1-mini</a>	推論と問題解決のタスクに取り組むために特別に設計され、絞られたフォーカスと高い能力を持つ、制限付きアクセス モデルです。
<a href="#">GPT-4o</a> 、 <a href="#">GPT-4o mini</a> 、 <a href="#">GPT-4 Turbo</a>	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。
<a href="#">GPT-4o audio</a>	低待機時間の "音声入力、音声出力" 会話をサポートする GPT-4o モデル。
<a href="#">GPT-4</a>	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。
<a href="#">GPT-3.5</a>	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。
<a href="#">埋め込み</a>	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。
<a href="#">DALL-E</a>	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。
<a href="#">Whisper</a>	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。
<a href="#">テキスト読み上げ</a> (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。

## o1-preview と o1-mini モデルの制限付きアクセス

Azure OpenAI の [o1-preview](#) と [o1-mini](#) 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)
<a href="#">o1-preview</a> (2024-09-12)	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。	入力: 128,000 出力: 32,768	2023年10月
<a href="#">o1-mini</a> (2024-09-12)	o1 シリーズの中でのより速く、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。	入力: 128,000 出力: 65,536	2023年10月

## 可用性

[o1-preview](#) および [o1-mini](#) モデルで API アクセスとモデル デプロイが利用可能になりました。登録が必要であり、Microsoft の資格条件に基づいてアクセス権が付与されます。

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。

## API のサポート

o1 シリーズ モデルのサポートが API バージョン [2024-09-01-preview](#) に追加されました。

[max\\_tokens](#) パラメーターは非推奨となり、新しい [max\\_completion\\_tokens](#) パラメーターに置き換えられました。o1 シリーズ モデルは、[max\\_completions\\_tokens](#) パラメーターでしか機能しません。

# 利用可能なリージョン

承認された利用者が、米国東部 2 とスウェーデン中部における標準およびグローバル標準デプロイで利用することができます。

## GPT-4o audio

`gpt-4o-realtime-preview` モデルは GPT-4o モデル ファミリの一部であり、低待機時間の "音声入力、音声出力" の会話をサポートします。GPT-4o 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 audio が利用できるのは、米国東部 2 (`eastus2`) とスウェーデン中部 (`swedencentral`) のリージョンにおいてです。GPT-4o audio を使用するには、サポートされているリージョンのいずれかの中のリソースを**作成する**か、既存のリソースを使用する必要があります。

リソースが作成されたら、GPT-4o audio 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 `gpt-4o-realtime-preview` となります。GPT-4o audio の使用方法の詳細については、[GPT-4o audio のドキュメント](#)を参照してください。

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)
<code>gpt-4o-realtime-preview</code> (2024-10-01-preview) <code>GPT-4o audio</code>	リアルタイム オーディオ処理のための <b>オーディオ モデル</b>	入力: 128,000 出力: 4,096	2023年10月

## 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 `2024-08-06`
- `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。

## 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。

## 利用可能なリージョン

リージョン別のモデルの提供状況については、[標準とプロビジョニングされたデプロイのモデル マトリックス](#)を参照してください。

## GPT-4 Turbo with Vision GA のデプロイ

Studio UI 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)
<code>gpt-4o (2024-08-06)</code> <code>GPT-4o (Omni)</code>	<b>最新の大きい GA モデル</b> <ul style="list-style-type: none"><li>- 構造化出力</li><li>- テキスト、画像処理</li><li>- JSON モード</li><li>- 並列関数呼び出し</li><li>- 精度と応答性の向上</li><li>- 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</li></ul>	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)
	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		
<code>gpt-4o-mini</code> (2024-07-18) GPT-4o mini	<b>最新の小さい GA モデル</b> - 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
<code>gpt-4o</code> (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
<code>gpt-4</code> (turbo-2024-04-09) GPT-4 Turbo with Vision	<b>新しい GA モデル</b> - 以前のすべての GPT-4 プレビュー モデル ( <code>vision-preview</code> 、 <code>1106-Preview</code> 、 <code>0125-Preview</code> ) についての代替モデル。 現在、 <b>機能の使用可否</b> 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
<code>gpt-4</code> (0125-Preview)* GPT-4 Turbo プレビュー	<b>プレビュー モデル</b> - 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
<code>gpt-4</code> (vision-preview) GPT-4 Turbo with Vision Preview	<b>プレビュー モデル</b> - 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
<code>gpt-4</code> (1106-Preview) GPT-4 Turbo プレビュー	<b>プレビュー モデル</b> - 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
<code>gpt-4-32k</code> (0613)	<b>古い GA モデル</b> - ツールによる基本的な関数呼び出し	32,768	2021 年 9 月
<code>gpt-4</code> (0613)	<b>古い GA モデル</b> - ツールによる基本的な関数呼び出し	8,192	2021 年 9 月
<code>gpt-4-32k</code> (0314)	<b>古い GA モデル</b> - <a href="#">廃止に関する情報</a>	32,768	2021 年 9 月
<code>gpt-4</code> (0314)	<b>古い GA モデル</b> - <a href="#">廃止に関する情報</a>	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。

## ① 重要

GPT-4 (gpt-4) バージョン 1106-Preview、0125-Preview、vision-preview は、将来的に安定バージョンの gpt-4 でアップグレードされます。

- "既定値に自動更新する" と "期限切れになったときにアップグレードする" に設定された gpt-4 のバージョン 1106-Preview、0125-Preview、vision-preview のデプロイに対しては、安定バージョンがリリースされるとアップグレードが開始されます。デプロイごとに、API 呼び出しのサービスを中断せず、モデル バージョンのアップグレードが行われます。アップグレードはリージョン別にステージングされ、完全なアップグレード プロセスには 2 週間かかると予想されます。
- "自動更新なし" に設定された gpt-4 のバージョン 1106-Preview、0125-Preview、vision-preview のデプロイはアップグレードされず、リージョン内でプレビュー バージョンがアップグレードされると動作を停止します。アップグレードのタイミングについては、[Azure OpenAI モデルの提供終了と非推奨](#)に関する記事を参照してください。

# 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)
<code>gpt-35-turbo</code> (0125) <b>新規</b>	<b>最新の GA モデル</b> <ul style="list-style-type: none"><li>- JSON モード</li><li>- 並列関数呼び出し</li><li>- 再現可能な出力 (プレビュー)</li><li>- 要求された形式での応答精度の向上。</li><li>- 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。</li></ul>	入力: 16,385 出力: 4,096	2021 年 9 月
<code>gpt-35-turbo</code> (1106)	<b>古い GA モデル</b> <ul style="list-style-type: none"><li>- JSON モード</li><li>- 並列関数呼び出し</li><li>- 再現可能な出力 (プレビュー)</li></ul>	入力: 16,385 出力: 4,096	2021 年 9 月
<code>gpt-35-turbo-instruct</code> (0914)	<b>入力候補エンドポイントのみ</b> <ul style="list-style-type: none"><li>- <b>レガシ補完モデル</b>の置き換え</li></ul>	4,097	2021 年 9 月
<code>gpt-35-turbo-16k</code> (0613)	<b>古い GA モデル</b> <ul style="list-style-type: none"><li>- ツールによる基本的な関数呼び出し</li></ul>	16,384	2021 年 9 月
<code>gpt-35-turbo</code> (0613)	<b>古い GA モデル</b> <ul style="list-style-type: none"><li>- ツールによる基本的な関数呼び出し</li></ul>	4,096	2021 年 9 月
<code>gpt-35-turbo</code> <sup>1</sup> (0301)	<b>古い GA モデル</b> <ul style="list-style-type: none"><li>- <b>廃止に関する情報</b></li></ul>	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。

<sup>1</sup> 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	<code>text-embedding-ada-002</code>	<code>text-embedding-3-small</code>	<code>text-embedding-3-large</code>
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

### ⓘ 注意

この記事では、展開の種類を **[標準]** とする Azure OpenAI のすべてのお客様に適用されるモデルとリージョンの可用性について説明します。一部のお客様は、以下に統合されている表には記載されていないモデルとリージョンの組み合わせにアクセスできます。プロビジョニング済みデプロイに関する詳細については、[プロビジョニング済みに関するガイダンス](#)を参照してください。

### 標準の展開モデルの可用性

[🔗 テーブルを展開する](#)

リージョン	<code>gpt-4o-0613</code>	<code>gpt-4o-1106-Preview</code>	<code>gpt-4o-0125-Preview</code>	<code>gpt-4o-vision-preview</code>	<code>gpt-4o-turbo-2024-04-09</code>	<code>gpt-4o-2024-年 5 月 13 日</code>	<code>gpt-4o-2024-08-06</code>	<code>gpt-4o-mini-2024-07-18</code>	<code>gpt-35-turbo-0301</code>	<code>gpt-35-turbo-0613</code>	<code>gpt-35-turbo-1106</code>	<code>gpt-35-turbo-0125</code>	<code>gpt-35-turbo-16k-0613</code>	<code>gpt-35-turbo-instruct-0914</code>
australiaeast	✓	✓	-	✓	-	-	-	✓	-	✓	✓	-	✓	-

リージョン	gpt-4、0613	gpt-4、1106-Preview	gpt-4、0125-Preview	gpt-4、vision-preview	gpt-4、turbo-2024-04-09	gpt-4o、2024年5月13日	gpt-4o、2024-08-06	gpt-4o-mini、2024-07-18	gpt-4、32k、0613	gpt-35-turbo、0301	gpt-35-turbo、0613	gpt-35-turbo、1106	gpt-35-turbo、0125	gpt-35-turbo-16k、0613	gpt-35-turbo-instruct、0914
brazilsouth	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
canadaeast	✓	✓	-	-	-	-	-	-	✓	-	✓	✓	✓	✓	-
eastus	-	-	✓	-	✓	✓	✓	✓	-	✓	✓	-	✓	✓	✓
eastus2	-	✓	-	-	✓	✓	✓	✓	-	-	✓	-	✓	✓	-
francecentral	✓	✓	-	-	-	-	-	-	✓	✓	✓	✓	-	✓	-
japaneast	-	-	-	✓	-	-	-	-	-	-	✓	-	-	✓	-
northcentralus	-	-	✓	-	✓	✓	✓	✓	-	-	✓	-	✓	✓	-
norwayeast	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
southafricanorth	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
southcentralus	-	-	✓	-	✓	✓	✓	-	-	✓	-	-	✓	-	-
southindia	-	✓	-	-	-	-	-	-	-	-	-	✓	-	-	-
swedencentral	✓	✓	-	✓	✓	✓	✓	✓	✓	-	✓	✓	-	✓	✓
switzerlandnorth	✓	-	-	✓	-	-	-	-	✓	-	✓	-	-	✓	-
uksouth	-	✓	✓	-	-	-	-	-	-	✓	✓	✓	-	✓	-
westeurope	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-
westus	-	✓	-	✓	✓	✓	✓	✓	-	-	-	✓	✓	-	-
westus3	-	✓	-	-	✓	✓	✓	✓	-	-	-	-	✓	-	-

この表には、GPT-4o の [グローバル標準](#) モデルの展開のリージョン別の提供状況や、微調整のリージョン別の提供状況に関する情報は含まれていません。この情報については、専用の [グローバル標準デプロイに関するセクション](#) と、 [微調整に関するセクション](#) をご覧ください。


## 標準とグローバル標準の展開モデルのクォータ

[🔗 テーブルを展開する](#)

リージョン	GPT-4	GPT-4-32K	GPT-4-Turbo	GPT-4-Turbo-V	gpt-4o	gpt-4o-mini	GPT-35-Turbo	GPT-35-Turbo-Instruct	gpt-4o - GlobalStandard	gpt-4o-mini - GlobalStandard	GPT-4-Turbo - GlobalStandard	GPT-4o - Global-Batch	GPT-4o-mini - GlobalBatch
australiaeast	40 K	80 K	80 K	30 K	-	-	300 K	-	30 M	-	2 M	-	-
brazilsouth	-	-	-	-	-	-	-	-	30 M	-	2 M	-	-
canadaeast	40 K	80 K	80 K	-	-	-	300 K	-	30 M	-	2 M	-	-
eastus	-	-	80 K	-	1 M	2 M	240 K	240 K	30 M	50 M	2 M	5 B	5 B
eastus2	-	-	80 K	-	1 M	2 M	300 K	-	30 M	50 M	2 M	-	-
francecentral	20 K	60 K	80 K	-	-	-	240 K	-	30 M	-	2 M	-	-
germanywestcentral	-	-	-	-	-	-	-	-	30 M	-	2 M	-	-
japaneast	-	-	-	30 K	-	-	300 K	-	30 M	-	2 M	-	-
koreacentral	-	-	-	-	-	-	-	-	30 M	-	2 M	-	-
northcentralus	-	-	80 K	-	1 M	2 M	300 K	-	30 M	50 M	2 M	-	-
norwayeast	-	-	150 K	-	-	-	-	-	30 M	-	2 M	-	-

リージョン	GPT-4	GPT-4-32K	GPT-4-Turbo	GPT-4-Turbo-V	gpt-4o	gpt-4o-mini	GPT-35-Turbo	GPT-35-Turbo-Instruct	gpt-4o - GlobalStandard	gpt-4o-mini - GlobalStandard	GPT-4-Turbo - GlobalStandard	GPT-4o - Global-Batch	GPT-4o-mini - Global Batch
polandcentral	-	-	-	-	-	-	-	-	30 M	-	2 M	-	-
southafricanorth	-	-	-	-	-	-	-	-	30 M	-	2 M	-	-
southcentralus	-	-	80 K	-	1 M	-	240 K	-	30 M	-	2 M	-	-
southindia	-	-	150 K	-	-	-	300 K	-	30 M	-	2 M	-	-
spaincentral	-	-	-	-	-	-	-	-	30 M	-	2 M	-	-
swedencentral	40 K	80 K	150 K	30 K	1 M	2 M	300 K	240 K	30 M	50 M	2 M	5 B	5 B
switzerlandnorth	40 K	80 K	-	30 K	-	-	300 K	-	30 M	50 M	2 M	-	-
switzerlandwest	-	-	-	-	-	-	-	-	-	-	-	-	-
uksouth	-	-	80 K	-	-	-	240 K	-	30 M	-	2 M	-	-
westeurope	-	-	-	-	-	-	240 K	-	30 M	50 M	2 M	-	-
westus	-	-	80 K	30 K	1 M	2 M	300 K	-	30 M	50 M	2 M	5 B	5 B
westus3	-	-	80 K	-	1 M	2 M	300 K	-	30 M	50 M	2 M	-	-

## プロビジョニング済みデプロイ モデルの可用性

 テーブルを展開する

リージョン	gpt-4、0613	gpt-4、1106-Preview	gpt-4、0125-Preview	gpt-4、turbo-2024-04-09	gpt-4o、2024 年 5 月 13 日	gpt-4o-mini、2024-07-18	gpt-4-32k、0613	gpt-35-turbo、1106	gpt-35-turbo、0125
australiaeast	✓	✓	✓	✓	✓	-	✓	✓	✓
brazilsouth	✓	✓	✓	-	✓	-	✓	✓	-
canadacentral	✓	-	-	-	-	-	✓	-	✓
canadaeast	✓	✓	-	✓	✓	✓	-	✓	-
eastus	✓	✓	✓	✓	✓	✓	✓	✓	✓
eastus2	✓	✓	✓	✓	✓	✓	✓	✓	✓
francecentral	✓	✓	✓	-	✓	-	✓	-	✓
germanywestcentral	✓	✓	✓	✓	✓	-	✓	✓	-
japaneast	-	✓	✓	✓	✓	-	-	-	✓
koreacentral	✓	-	-	✓	✓	-	✓	✓	-
northcentralus	✓	✓	✓	✓	✓	✓	✓	✓	✓
norwayeast	✓	-	✓	-	-	-	✓	-	-
polandcentral	✓	✓	✓	-	-	-	✓	✓	✓
southafricanorth	✓	✓	-	✓	-	-	✓	✓	-
southcentralus	✓	✓	✓	✓	✓	-	✓	✓	✓
southindia	✓	✓	✓	-	✓	-	✓	✓	✓
swedencentral	✓	✓	✓	✓	✓	✓	✓	✓	✓
switzerlandnorth	✓	✓	✓	✓	✓	-	✓	✓	✓
switzerlandwest	-	-	-	-	-	-	-	-	✓
uksouth	✓	✓	✓	✓	✓	-	✓	✓	✓



リージョン	gpt-4、 0613	gpt-4、 1106- Preview	gpt-4、 0125- Preview	gpt-4、 turbo-2024- 04-09	gpt-4o、 2024 年 5 月 13 日	gpt-4o- mini、2024- 07-18	gpt-4- 32k、 0613	gpt-35- turbo、1106	gpt-35- turbo、0125
westus	✓	✓	✓	✓	✓	-	✓	✓	✓
westus3	✓	✓	✓	✓	✓	-	✓	✓	✓

ⓘ 注意

**gpt-4 バージョン:** turbo-2024-04-09 のプロビジョニングされたバージョンは、現在、テキストのみに制限されています。

## プロビジョニング スループットにアクセスするにはどうすればよいですか？

プロビジョニング スループットを取得するには、Microsoft の営業/アカウント チームに問い合わせる必要があります。営業/アカウント チームがない場合、残念ながら現時点ではプロビジョニング スループットを購入することはできません。

プロビジョニング済みデプロイに関する詳細については、[プロビジョニング済みに関するガイダンス](#)を参照してください。

## Global-Standard モデルの提供状況

**gpt-4o バージョン** 2024-08-06

サポートされているリージョン:

- eastus
- eastus2
- northcentralus
- southcentralus
- swedencentral
- westus
- westus3

**gpt-4o バージョン** 2024-05-13

サポートされているリージョン:

- australiaeast
- brazilsouth
- canadaeast
- eastus
- eastus2
- francecentral
- germanywestcentral
- japaneast
- koreacentral
- northcentralus
- norwayeast
- polandcentral
- spaincentral
- southafricanorth
- southcentralus
- southindia
- swedencentral
- switzerlandnorth
- uksouth
- westeurope
- westus
- westus3

**gpt-4o-mini バージョン** 2024-07-18

サポートされているリージョン:

- eastus
- eastus2
- northcentralus
- swedencentral
- switzerlandnorth
- westus
- westus3

gpt-4 バージョン: turbo-2024-04-09

- australiaeast
- brazilsouth
- canadaeast
- eastus
- eastus2
- francecentral
- germanywestcentral
- japaneast
- koreacentral
- northcentralus
- norwayeast
- polandcentral
- spaincentral
- southafricanorth
- southcentralus
- southindia
- swedencentral
- switzerlandnorth
- uksouth
- westeurope
- westus
- westus3

## グローバル バッチ モデルの可用性

### リージョンとモデルのサポート

以下のモデルがグローバル バッチをサポートしています。

[🔗 テーブルを展開する](#)

モデル	バージョン	入力形式
gpt-4o	2024-08-06	テキストと画像
gpt-4o-mini	2024-07-18	テキストと画像
gpt-4o	2024-05-13	テキストと画像
gpt-4	turbo-2024-04-09	text
gpt-4	0613	text
gpt-35-turbo	0125	text
gpt-35-turbo	1106	text
gpt-35-turbo	0613	text

グローバル バッチが現在サポートされているのは以下のリージョンです。

- 米国東部
- 米国西部
- スウェーデン中部

# GPT-4 および GPT-4 Turbo モデルの可用性

## パブリック クラウド リージョン

 テーブルを展開する

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


## 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

 **重要**

新しい `gpt-35-turbo (0125)` モデルはさまざまな機能強化が組み込まれました。たとえば、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などです。

GPT-3.5 Turbo は、Chat Completions API と共に使用されます。GPT-3.5 Turbo バージョン 0301 は Completions API でも使用できますが、これは推奨されません。GPT-3.5 Turbo バージョン 0613 および 1106 では、Chat Completions API のみがサポートされます。

GPT-3.5 Turbo バージョン 0301 は、リリースされたモデルの最初のバージョンです。バージョン 0613 は、モデルの 2 番目のバージョンであり、関数呼び出しのサポートが追加されます。

[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-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。

## GPT-3.5-Turbo モデルの可用性

### パブリック クラウド リージョン

[🔗 テーブルを展開する](#)

リージョン	gpt-35-turbo、 0301	gpt-35-turbo、 0613	gpt-35-turbo、 1106	gpt-35-turbo、 0125	gpt-35-turbo-16k、 0613	gpt-35-turbo-instruct、 0914
australiaeast	-	✓	✓	-	✓	-
canadaeast	-	✓	✓	✓	✓	-
eastus	✓	✓	-	✓	✓	✓
eastus2	-	✓	-	✓	✓	-
francecentral	✓	✓	✓	-	✓	-
japaneast	-	✓	-	-	✓	-
northcentralus	-	✓	-	✓	✓	-
southcentralus	✓	-	-	✓	-	-
southindia	-	-	✓	-	-	-
swedencentral	-	✓	✓	-	✓	✓
switzerlandnorth	-	✓	-	-	✓	-
uksouth	✓	✓	✓	-	✓	-
westeurope	✓	-	-	-	-	-
westus	-	-	✓	✓	-	-
westus3	-	-	-	✓	-	-

### 埋め込みモデル

これらのモデルは埋め込み API 要求でのみ使用できます。

#### ⓘ 注意

`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。

[🔗 テーブルを展開する](#)

モデル ID	最大要求 (トークン)	出力ディメンション	トレーニングデータ (最大)
<code>text-embedding-ada-002</code> (バージョン 2)	8,191	1,536	2021 年 9 月
<code>text-embedding-ada-002</code> (バージョン 1)	2,046	1,536	2021 年 9 月
<code>text-embedding-3-large</code>	8,191	3,072	2021 年 9 月
<code>text-embedding-3-small</code>	8,191	1,536	2021 年 9 月

#### ⓘ 注意

埋め込み用に入力の配列を送信する場合、埋め込みエンドポイントへの呼び出しにつき、配列内での入力項目の最大数は 2048 です。

## パブリック クラウド リージョン

[🔗 テーブルを展開する](#)

リージョン	text-embedding-ada-002、1	text-embedding-ada-002、2	text-embedding-3-small、1	text-embedding-3-large、1
australiaeast	-	✓	-	-
brazilsouth	-	✓	-	-
canadaeast	-	✓	✓	✓
eastus	✓	✓	✓	✓
eastus2	-	✓	✓	✓
francecentral	-	✓	-	✓
japaneast	-	✓	-	✓
northcentralus	-	✓	-	-
norwayeast	-	✓	-	✓
southafricanorth	-	✓	-	-
southcentralus	✓	✓	-	-
southindia	-	✓	-	✓
swedencentral	-	✓	-	✓
switzerlandnorth	-	✓	-	-
uksouth	-	✓	-	✓
westeurope	-	✓	-	-
westus	-	✓	-	-
westus3	-	✓	-	✓

## DALL-E モデル

[🔗 テーブルを展開する](#)

モデル ID	機能の可用性	最大要求数 (文字)
dalle2 (プレビュー)	米国東部	1,000
dall-e-3	米国東部、オーストラリア東部、スウェーデン中部	4,000

## モデルの微調整

`babbage-002` と `davinci-002` は、指示に従ってトレーニングされていません。これらの基本モデルのクエリは、トレーニングの進行状況を評価するために、微調整されたバージョンへの参照ポイントとしてのみ実行する必要があります。

`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。

[🔗 テーブルを展開する](#)

モデル ID	リージョンの微調整	最大要求 (トークン)	トレーニング データ (最大)
<code>babbage-002</code>	米国中北部 スウェーデン中部 スイス西部	16,384	2021 年 9 月

モデル ID	リージョンの微調整	最大要求 (トークン)	トレーニング データ (最大)
davinci-002	米国中北部 スウェーデン中部 スイス西部	16,384	2021 年 9 月
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) <sup>1</sup>	米国中北部 スウェーデン中部	8192	2021 年 9 月
gpt-4o-mini <sup>1</sup> (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 <sup>1</sup> (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

<sup>1</sup> GPT-4、GPT-4o、GPT-4o mini のファインチューニングは現在パブリック プレビュー段階です。詳細については、「[GPT-4、GPT-4o、GPT-4o mini のファインチューニングの安全性評価ガイドンス](#)」を参照してください。

## Whisper モデル

[🔗 テーブルを展開する](#)

モデル ID	モデルの可用性	最大要求数 (オーディオ ファイル サイズ)
whisper	米国東部 2 米国中北部 ノルウェー東部 インド南部 スウェーデン中部 西ヨーロッパ	25 MB

## テキスト読み上げモデル (プレビュー)

[🔗 テーブルを展開する](#)

モデル ID	モデルの可用性
tts-1	米国中北部 スウェーデン中部
tts-1-hd	米国中北部 スウェーデン中部

## 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。 Assistants API、SDK、Azure AI Studio、Azure OpenAI Studio では、次のモデルを使用できます。 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	<code>gpt-35-turbo</code> (0613)	<code>gpt-35-turbo</code> (1106)	<code>fine tuned gpt-3.5-turbo-0125</code>	<code>gpt-4</code> (0613)	<code>gpt-4</code> (1106)	<code>gpt-4</code> (0125)	<code>gpt-4o</code> (2024-05-13)	<code>gpt-4o-mini</code> (2024-07-18)
オーストラリア東部	✓	✓		✓	✓			
米国東部	✓					✓	✓	✓
米国東部 2	✓		✓	✓	✓		✓	
フランス中部	✓	✓		✓	✓			
東日本	✓							
ノルウェー東部					✓			
スウェーデン中部	✓	✓	✓	✓	✓		✓	
英国南部	✓	✓			✓	✓		
米国西部		✓			✓		✓	
米国西部 3					✓		✓	

## 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