Azure によって直接販売される Foundry Models

> **Note**
>
> このドキュメントでは、 Microsoft Foundry (クラシック) ポータルを参照します。

この記事では、Azure によって直接販売される Microsoft Foundry モデルの一覧と、その機能、 デプロイの種類、および可用性のリージョン (非推奨モデルとレガシ モデルを除く) を示します。 Foundry Agent Service でサポートされている Azure OpenAI モデルの一覧については、「 エージェント サービスでサポートされるモデル」を参照してください。

Azure によって直接販売されるモデルには、すべての Azure OpenAI モデルと、上位プロバイダーから選択された特定のモデルが含まれます。

Microsoft Foundry で使用する プロジェクトの種類 に応じて、さまざまなモデルの選択が表示されます。 具体的には、Foundry リソースに基づいて構築された Foundry プロジェクトを使用すると、Foundry リソースへの標準デプロイに使用できるモデルが表示されます。 または、Foundry ハブによってホストされているハブ ベースのプロジェクトを使用する場合は、マネージド コンピューティング API とサーバーレス API へのデプロイに使用できるモデルが表示されます。 多くのモデルで複数のデプロイ オプションがサポートされているため、これらのモデルの選択肢は重複することがよくあります。

Azure によって直接販売される Foundry モデルの属性の詳細については、「Foundry モデルの探索」を参照してください。

> **Note**
>
> Azure によって直接販売される Foundry Models には、次の上位モデル プロバイダーから選択したモデルも含まれます。
>
> - ブラック フォレスト ラボ: `FLUX.1-Kontext-pro`、 `FLUX-1.1-pro`
> - DeepSeek: `DeepSeek-V3.1`、 `DeepSeek-V3-0324`、 `DeepSeek-R1-0528`、 `DeepSeek-R1`
> - メタ: `Llama-4-Maverick-17B-128E-Instruct-FP8`、 `Llama-3.3-70B-Instruct`
> - Microsoft: `MAI-DS-R1`
> - ミストラル： `mistral-document-ai-2505`
> - xAI: `grok-code-fast-1`、 `grok-3`、 `grok-3-mini`、 `grok-4-fast-reasoning`、 `grok-4-fast-non-reasoning`、 `grok-4`
>
> これらのモデルの詳細については、この記事の上部にある **他のモデル コレクション** に切り替えます。

Microsoft Foundry モデルの Azure OpenAI

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 の Azure OpenAI」を参照してください。

テーブルを展開する

Models	Description
GPT-5.1 シリーズ	**NEW** `gpt-5.1`、 `gpt-5.1-chat`、 `gpt-5.1-codex`、 `gpt-5.1-codex-mini`
そら	**NEW** sora-2
GPT-5 シリーズ	gpt-5、gpt-5-mini、gpt-5-nano、gpt-5-chat
gpt-oss	オープンウェイト推論モデル
codex-mini	o4-mini の微調整されたバージョン。
GPT-4.1 シリーズ	gpt-4.1、gpt-4.1-mini、gpt-4.1-nano
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。
computer-use-preview	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。
o シリーズ モデル	高度な問題解決とフォーカスと能力の向上を備えた推論モデル。
GPT-4o、GPT-4o mini、GPT-4 Turbo	マルチモーダル バージョンの対応 Azure OpenAI モデル。テキストと画像の両方を入力として受け入れることが可能です。
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。

Models	Description
Embeddings	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。
イメージの生成	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。
Video generation	テキスト命令から元のビデオ シーンを生成できるモデル。
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 *の音声イン、音声アウト* 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-5.1

リージョンの可用性

テーブルを展開する

モデル	リージョン
`gpt-5.1`	米国東部 2 およびスウェーデン中部 (グローバル標準および DataZone 標準)
`gpt-5.1-chat`	米国東部 2 およびスウェーデン中部 (グローバル標準)
`gpt-5.1-codex`	米国東部 2 およびスウェーデン中部 (グローバル標準)
`gpt-5.1-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)

- **gpt-5.1 および gpt-5.1-codex へのアクセスには登録が必要です** 。

アクセスは、Microsoft の資格条件に基づいて付与されます。 以前に制限付きアクセス モデルへのアクセスを適用して受け取ったお客様は、承認されたサブスクリプションがモデルのリリース時に自動的にアクセス権を付与されるため、再適用する必要はありません。

テーブルを展開する

モデル ID	Description	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-5.1` (2025-11-13)	- 推論 - Chat Completions API。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要。	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 9 月 30 日
`gpt-5.1-chat` (2025-11-13)	- 推論 - Chat Completions API。 - 応答 API。 - 構造化された出力 - 関数、ツール、および並列ツール呼び出し。	128,000 入力: 111,616 出力: 16,384	16,384	2024 年 9 月 30 日
`gpt-5.1-codex` (2025-11-13)	- 応答 API のみ。 - テキストと画像処理 - 構造化出力 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要 - Codex CLI および Codex VS Code 拡張機能用に最適化	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 9 月 30 日
`gpt-5.1-codex-mini` (2025-11-13)	- 応答 API のみ。 - テキストと画像処理 - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要 - Codex CLI および Codex VS Code 拡張機能用に最適化	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 9 月 30 日

重要

- `gpt-5.1` `reasoning_effort` は既定で `none` に設定されます。 以前の推論モデルから `gpt-5.1` にアップグレードする場合は、推論を実行する場合は、 `reasoning_effort` レベルを明示的に渡すようにコードを更新する必要がある場合があることに注意してください。
- `gpt-5.1-chat` は、組み込みの推論機能を追加します。 他の **推論モデル** と同様に、 `temperature` などのパラメーターはサポートされません。 `gpt-5-chat` (推論モデルではない) を使用して `gpt-5.1-chat` アップグレードする場合は、推論モデルでサポートされていない `temperature` などのカスタム パラメーターをコードから削除してください。

GPT-5

リージョンの可用性

テーブルを展開する

モデル	リージョン
`gpt-5` (2025-08-07)	「モデル テーブル」を参照してください。
`gpt-5-mini` (2025-08-07)	「モデル テーブル」を参照してください。
`gpt-5-nano` (2025-08-07)	「モデル テーブル」を参照してください。
`gpt-5-chat` (2025-08-07)	「モデル テーブル」を参照してください。
`gpt-5-chat` (2025-10-03)	米国東部 2 (グローバル標準) とスウェーデン中部 (グローバル標準)
`gpt-5-codex` (2025-09-11)	米国東部 2 (グローバル標準) とスウェーデン中部 (グローバル標準)
`gpt-5-pro` (2025-10-06)	米国東部 2 (グローバル標準) とスウェーデン中部 (グローバル標準)

- **gpt-5-pro、gpt-5、および gpt-5-codex モデルへのアクセスには登録が必要です** 。
- `gpt-5-mini` 、 `gpt-5-nano` 、 `gpt-5-chat` は登録を必要としません。

アクセスは、Microsoft の資格条件に基づいて付与されます。 以前に o3 へのアクセスを適用して受け取ったお客様は、承認されたサブスクリプションにモデルのリリース時に自動的にアクセス権が付与されるため、再適用する必要はありません。

テーブルを展開する

モデル ID	Description	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-5` (2025-08-07)	- 推論 - Chat Completions API。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要。	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 9 月 30 日
`gpt-5-mini` (2025-08-07)	- 推論 - Chat Completions API。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要。	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 5 月 31 日
`gpt-5-nano` (2025-08-07)	- 推論 - Chat Completions API。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要。	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 5 月 31 日
`gpt-5-chat` (2025-08-07) **プレビュー**	- Chat Completions API。 - 応答 API。 - **入力**: テキスト/画像 - **出力**: テキストのみ	128,000	16,384	2024 年 9 月 30 日

モデル ID	Description	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-5-chat` (2025-10-03) **プレビュー**[1]	- Chat Completions API。 - 応答 API。 - **入力**: テキスト/画像 - **出力**: テキストのみ	128,000	16,384	2024 年 9 月 30 日
`gpt-5-codex` (2025-09-11)	- 応答 API のみ。 - **入力**: テキスト/画像 - **出力**: テキストのみ - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 - 機能の完全な概要 - Codex CLI および Codex VS Code 拡張機能用に最適化	400,000 入力: 272,000 出力: 128,000	128,000	-
`gpt-5-pro` (2025-10-06)	- 推論 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数とツール - 機能の完全な概要。	400,000 入力: 272,000 出力: 128,000	128,000	2024 年 9 月 30 日

> **Note**
>
> [1] `gpt-5-chat` バージョン `2025-10-03` は、感情的な知性と精神的健康能力に焦点を当てた重要な強化を導入します。 このアップグレードでは、特殊なデータセットと調整された応答戦略が統合され、モデルの次の機能が向上します。
>
> - 感情コンテキストをより正確に**理解して解釈**し、微妙で共感的な相互作用を可能にします。
> - メンタルヘルスに関連する会話では、**支援的で責任ある対応を行い**、感受性と最高の実践に従うことを確保します。
>
> これらの改善により、GPT-5 チャットは、感情的なトーンと幸福に関する考慮事項が重要なシナリオで、コンテキストに対応し、人間中心で信頼性が高くなります。

gpt-oss

リージョンの可用性

テーブルを展開する

モデル	リージョン
`gpt-oss-120b`	すべての Azure OpenAI リージョン

能力

テーブルを展開する

モデル ID	Description	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-oss-120b` (プレビュー)	- テキストイン/テキストアウトのみ - チャット完了API -ストリーミング - 関数呼び出し - 構造化された出力 - 推論 - デプロイ[1] およびマネージド コンピューティングを使用して使用できます	131,072	131,072	2024 年 5 月 31 日
`gpt-oss-20b` (プレビュー)	- テキストイン/テキストアウトのみ - チャット完了API -ストリーミング - 関数呼び出し - 構造化された出力	131,072	131,072	2024 年 5 月 31 日

モデル ID	Description	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)
	- 推論 - マネージド コンピューティングと Foundry Local を介して使用可能			

[1] 他の Azure OpenAI モデルとは異なり、 `gpt-oss-120b` モデルをデプロイするには Foundry プロジェクト が必要です。

コードを使用してデプロイする

cli

```
az cognitiveservices account deployment create \
  --name "Foundry-project-resource" \
  --resource-group "test-rg" \
  --deployment-name "gpt-oss-120b" \
  --model-name "gpt-oss-120b" \
  --model-version "1" \
  --model-format "OpenAI-OSS" \
  --sku-capacity 10 \
  --sku-name "GlobalStandard"
```

GPT-4.1 シリーズ

リージョンの可用性

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 テーブル」を参照してください。
`gpt-4.1-nano` (2025-04-14)	「モデル テーブル」を参照してください。
`gpt-4.1-mini` (2025-04-14)	「モデル テーブル」を参照してください。

能力

重要

既知の問題は、すべての GPT 4.1 シリーズ モデルに影響します。 300,000 トークンを超える大規模なツールまたは関数呼び出しの定義では、モデルの 100 万個のトークン コンテキスト制限に達しなかった場合でも、エラーが発生します。

エラーは、API 呼び出しと基になるペイロードの特性によって異なる場合があります。

Chat Completions API のエラー メッセージを次に示します:

- `Error code: 400 - {'error': {'message': "This model's maximum context length is 300000 tokens. However, your messages resulted in 350564 tokens (100 in the messages, 350464 in the functions). Please reduce the length of the messages or functions.", 'type': 'invalid_request_error', 'param': 'messages', 'code': 'context_length_exceeded'}}`
- `Error code: 400 - {'error': {'message': "Invalid 'tools[0].function.description': string too long. Expected a string with maximum length 1048576, but got a string with length 2778531 instead.", 'type': 'invalid_request_error', 'param': 'tools[0].function.description', 'code': 'string_above_max_length'}}`

Responses API のエラー メッセージを次に示します:

- `Error code: 500 - {'error': {'message': 'The server had an error processing your request. Sorry about that! You can retry your request, or contact us through an Azure support request at: https://go.microsoft.com/fwlink/?linkid=2213926 if you keep seeing this error. (Please include the request ID d2008353-291d-428f-adc1-defb5d9fb109 in your email.)', 'type': 'server_error', 'param': None, 'code': None}}`

テーブルを展開する

モデル ID	Description	コンテキスト ウィンドウ	最大出力トークン	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 - チャット完了 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 (チャット補完)	- 1,047,576 - 128,000 (プロビジョニングされたマネージド デプロイ) - 300,000 (バッチ 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 - チャット完了 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 (チャット補完)	- 1,047,576 - 128,000 (プロビジョニングされたマネージド デプロイ) - 300,000 (バッチ デプロイ)	32,768	2024 年 5 月 31 日
gpt-4.1-mini (2025-04-14)	- テキストと画像の入力 - テキスト出力 - チャット完了 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 (チャット補完)	- 1,047,576 - 128,000 (プロビジョニングされたマネージド デプロイ) - 300,000 (バッチ デプロイ)	32,768	2024 年 5 月 31 日

computer-use-preview

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。

サーとパーティ製ライブラリと共に使用すると、モデルは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、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を、将来のプレビュー バージョンまたは最新の安定した一般公開バージョンにアップグレードします。 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。

`computer-use-preview` にアクセスするには登録が必要です。 アクセスは、Microsoft の資格条件に基づいて付与されます。 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。

アクセスを要求するには、computer-use-preview 制限付きアクセス モデル アプリケーション にアクセスしてください アクセスが付与されたら、モデルのデプロイを作成する必要があります。

リージョンの可用性

テーブルを展開する

モデル	リージョン
computer-use-preview	「モデル テーブル」を参照してください。

能力

テーブルを展開する

モデル ID	Description	コンテキスト ウィンドウ	最大出力トークン	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 - テキスト (入力と出力) - 画像 (入力)	8,192	1,024	2023 年 10 月

o シリーズ モデル

Azure OpenAI o シリーズ モデルは、集中力と能力を高め、推論と問題解決のタスクに取り組むため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、これまでの反復と比較して、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	Description	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` の微調整化バージョン。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数とツール。 機能の完全な要約。	入力: 200,000 出力: 100,000	2024 年 5 月 31 日
`o3-pro` (2025-06-10)	- 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数とツール。 機能の完全な要約。	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。 - Chat Completions API。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数とツール。 機能の完全な要約。	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。 - Chat Completions API。 - 応答 API。 - 構造化出力 - テキストと画像処理。 - 関数、ツール、および並列ツール呼び出し。 機能の完全な要約。	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 - 関数とツール。	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 - テキストと画像処理。 - 関数とツール。	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シリーズのより速く、よりコスト効率の高いオプションで、速度とリソース消費の削減を必要とするタスクのコーディングに最適です。 - グローバル標準の展開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が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

高度な o シリーズ 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。

リージョンの可用性

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 テーブル」を参照してください。
o3	「モデル テーブル」を参照してください。
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。 このモデルは、元の制限付きアクセスの一部としてアクセス権が付与されたお客様のみが使用できます。
o1-mini	「モデル テーブル」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、英語以外の言語のタスクやビジョン タスクで優れたパフォーマンスを提供しながら、英語テキストおよびコーディング タスクの GPT-4 Turbo と一致し、AI 機能の新しいベンチマークを設定します。

GPT-4 モデルと GPT-4 Turbo モデル

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

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については、「モデル バージョン」を参照してください。 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は、「モデルの操作」を参照してください。

テーブルを展開する

モデル ID	Description	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)	- 構造化出力 - テキストと画像処理。 - 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 タスクを含む Parity。 - 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 (オムニ)	- 構造化出力 - テキストと画像処理。 - 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 タスクを含む Parity。 - 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	- 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。 - テキストと画像処理。 - 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)	- テキストと画像処理。 - JSON モード。 - 並列関数呼び出し。	入力: 128,000 出力: 4,096	2023 年 10 月

モデル ID	Description	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 タスクを含む Parity。 - 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-4` (turbo-2024-04-09) GPT-4 ターボ ウィズ ビジョン	新しい一般提供モデル。 - 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。 - 機能の可用性は、現在、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

⊗ 注意事項

運用環境ではプレビュー モデルを使用しないことをお勧めします。 プレビュー モデルのすべてのデプロイを、将来のプレビュー バージョンまたは最新の安定した一般公開バージョンにアップグレードします。 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 ファミリで最も優れたコスト効率の高いモデルは GPT-3.5 Turbo であり、チャット用に最適化されており、従来の完了タスクにも適しています。 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 を使用する場合に `text-davinci-003` する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	Description	最大要求 (トークン)	トレーニング データ (最大)
`gpt-35-turbo` (0125) *new*	- JSON モード。 - 並列関数呼び出し。 - 再現可能な出力 (プレビュー)。 - 要求された形式で応答するときの精度が高くなります。 - 英語以外の関数呼び出しでテキスト エンコードの問題を引き起こしたバグの修正プログラムが含まれています。	入力: 16,385 出力: 4,096	2021 年 9 月
`gpt-35-turbo` (1106)	以前の一般提供モデル。 - 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)	*補完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-3.5 Turbo と Chat Completions API を操作する方法の詳細については、詳細なハウツー記事を参照してください。

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` を使用して `text-embedding-3-large` 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 レポートでは、テストでは、大規模および小規模の第 3 世代埋め込みモデルの両方が MIRACL ベンチマークを使用して平均多言語取得パフォーマンスを向上することを示しています。 MTEB ベンチマークを使用して、英語のタスクのパフォーマンスを引き続き維持します。

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

評価ベンチマーク	text-embedding-ada-002	text-embedding-3-small	text-embedding-3-large
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。 ディメンションの数を調整できる場合は、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 `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 テストでは、第 3 世代モデルのディメンションが `text-embeddings-ada-002` の 1,536 ディメンションより小さい場合でも、パフォーマンスは依然として若干良いことが判りました。

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 シリーズ モデルは、制限付きアクセス 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。

`gpt-image-1` または `gpt-image-1-mini` にアクセスするには、登録が必要です。 アクセスは、Microsoft の資格条件に基づいて付与されます。 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。

アクセスを要求するには、gpt-image-1 制限付きアクセス モデル アプリケーション にアクセスしてください アクセスが付与されたら、モデルのデプロイを作成する必要があります。

リージョンの可用性

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) 米国東部地域 2 (グローバル標準) アラブ首長国連邦北部 (グローバル標準) ポーランド中部 (グローバル標準) スウェーデン中部 (グローバル標準)
`gpt-image-1-mini`	米国西部 3 (グローバル標準) 米国東部地域 2 (グローバル標準) アラブ首長国連邦北部 (グローバル標準) ポーランド中部 (グローバル標準) スウェーデン中部 (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。 Sora はプレビュー段階です。

リージョンの可用性

テーブルを展開する

モデル	リージョン
`sora`	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)
`sora-2`	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)

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を、将来のプレビュー バージョンまたは最新の安定した一般公開バージョンにアップグレードします。 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	Description	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 オーディオ	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: 16,384	2023 年 9 月
`gpt-4o-audio-preview` (2024-12-17) GPT-4o オーディオ	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: 16,384	2023 年 9 月
`gpt-4o-realtime-preview` (2025-06-03) GPT-4o オーディオ	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` (2024-12-17) GPT-4o オーディオ	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-mini-realtime-preview` (2024-12-17) GPT-4o オーディオ	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-realtime` (2025年08月28日) (GA) `gpt-realtime-mini` (2025-10-06) `gpt-audio` (2025年8月28日) `gpt-audio-mini` (2025-10-06)	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: 28,672 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	Description	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 を搭載した音声テキスト変換モデル。	25 MB
`gpt-4o-mini-transcribe`	GPT-4o mini を搭載した音声テキスト変換モデル。	25 MB
`gpt-4o-transcribe-diarize`	自動音声認識を使用した音声テキスト変換モデル。	25 MB

音声翻訳モデル

テーブルを展開する

モデル ID	Description	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	Description
tts	速度に合わせて最適化されたテキスト読み上げモデル。
tts-hd	品質に最適化されたテキスト読み上げモデル。
gpt-4o-mini-tts	GPT-4o mini を搭載したテキスト読み上げモデル。 特定のスタイルまたは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、この記事で後述する「オーディオ モデルの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。

- **標準**: グローバル展開オプションがあり、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**: また、グローバル デプロイ オプションを使用して、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。

すべてのデプロイでまったく同じ推論操作を実行できますが、請求、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

グローバル標準モデルの提供状況

テーブルを展開する

リージョン	gpt-5、2025-08-07	gpt-5-mini、2025-08-07	gpt-5-nano、2025-08-07	gpt-5-chat, 2025-08-07	o3-pro、2025-06-10	codex-mini、2025-05-16	sora、2025-05-02	model-router, 2025-08-07	model-router, 2025-05-19	o3、2025-04-16	o4-mini、2025-04-16	gpt-image-1、2025-04-15	gpt-image-1-mini、2025-10-06	gpt-4.1、2025-04-14	gpt-4.1-nano、2025-04-14	gpt-4.1-mini、2025-04-14
オーストラリア イースト	✅	✅	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-
ブラジルサウス	-	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-
カナダ東部	✅	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-
イーストアス	✅	✅	-	-	-	-	-	-	-	✅	✅	-	✅	-	✅	-
eastus2	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅	✅
francecentral	-	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-
ドイツ中西部	-	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-
italynorth	-	-	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-
japaneast	✅	✅	-	-	-	-	-	-	-	✅	✅	-	-	✅	-	-

リージョン	gpt-5、2025-08-07	gpt-5-mini、2025-08-07	gpt-5-nano、2025-08-07	gpt-5-chat, 2025-08-07	o3-pro、2025-06-10	codex-mini、2025-05-16	sora、2025-05-02	model-router, 2025-08-07	model-router, 2025-05-19	o3、2025-04-16	o4-mini、2025-04-16	gpt-image-1、2025-04-15	gpt-image-1-mini、2025-10-06	gpt-4.1、2025-04-14	gpt-4.1-nano、2025-04-14	gpt-4.1-mini、2025-04-14
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	-	-	-	-	-	-	-	-	-	✅	✅	✅	-	✅	-	-

> **Note**
>
> `o3-deep-research` は現在、Foundry Agent Service でのみ使用できます。 詳細については、 **ディープ リサーチ ツールのガイダンス**を参照してください。

この表には、リージョンごとの提供状況の微調整に関する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

チャットの完了

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、2024-11-20	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
オーストラリアイースト	-	-	-	✅	-	-	-	✅	✅
カナダ東部	-	-	-	✅	-	-	-	✅	✅
イーストアス	✅	✅	✅	✅	✅	✅	✅	-	✅

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、2024-11-20	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
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	✅	✅	✅	✅	✅	✅	✅	-	✅

> **Note**
>
> 現在、 `o1-mini` は、グローバル標準の展開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。 現時点では、 `o1-mini` 標準 (リージョン) デプロイへのアクセスは拡張されていません。

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については、「モデル バージョン」を参照してください。 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は、「モデルを使用した作業」を参照してください。

モデルの微調整

> **Note**
>
> `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。
>
> 微調整でサポートされるリージョンは、Microsoft Foundry プロジェクトで Azure OpenAI モデルを使用する場合とプロジェクト外で使用する場合は異なる場合があります。

テーブルを展開する

モデル ID	標準のトレーニング リージョン	グローバル トレーニング	最大要求 (トークン)	トレーニング データ (最大)	Modality
`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

モデル ID	標準のトレーニング リージョン	グローバル トレーニング	最大要求 (トークン)	トレーニング データ (最大)	Modality
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	テキストと視覚テキスト
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	テキストと視覚テキスト
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

Note

グローバル トレーニングでは、トークンごとに **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ます。

- オーストラリア東部
- ブラジル南部
- カナダ中部
- カナダ東部
- 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 *(4.1 ナノサポートなし)*
- 東南アジア
- 南アフリカ北部
- 米国中南部
- インド南部
- 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
- 西ヨーロッパ
- 米国西部

- 米国西部 3

アシスタント (プレビュー)

アシスタントについては、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、Foundry では、次のモデルを使用できます。 次の表は、標準のデプロイ用です。 プロビジョニングされたスループット ユニットの可用性の詳細については、「プロビジョニングされたスループット」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、次の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

リージョン	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	✅	✅	✅	-	✅	-	✅	-	-	✅	✅	-
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。

関連コンテンツ

- フォンドリーのモデル：パートナーおよびコミュニティから
- 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

ⓘ **注:** 作成者は AI の支援の下、この記事を作成しました。 詳細情報

Last updated on 2025/11/13