

Azure OpenAI Service とは

[アーティクル] • 2023/06/28

Azure OpenAI Service では、GPT-3、Codex、Embeddings モデル シリーズなど OpenAI の強力な言語モデルを REST API として使用できます。さらに、新しい GPT-4 と ChatGPT (gpt-35-turbo) モデルシリーズが一般提供されました。これらのモデルは、特定のタスクに合わせて簡単に調整できます。たとえば、コンテンツの生成、まとめ、セマンティック検索、自然言語からコードへの翻訳などです。ユーザーは、REST API、Python SDK、または Azure OpenAI Studio の Web ベースのインターフェイスを介してサービスにアクセスできます。

機能の概要

機能	Azure OpenAI
使用できるモデル	新しい GPT-4 シリーズ GPT-3 ベース シリーズ 新しい ChatGPT (gpt-35-turbo) Codex シリーズ 埋め込みシリーズ 詳細については、 モデル に関するページを参照してください。
微調整	Ada Babbage Curie Cushman Davinci 現在、新規のお客様はファインチューニングを利用できません。
Price	こちらで入手可能 
仮想ネットワークのサポート & プライベート リンクのサポート	はい (独自のデータに基づく Azure OpenAI を使用しない限り)。
マネージド ID	はい、Azure Active Directory 経由
UI エクスペリエンス	アカウントとリソースの管理には Azure Portal 、 モデルの探索と微調整には Azure OpenAI Service Studio
FPGA のリージョン別の提供状況	モデルの可用性
コンテンツのフィルター処理	プロンプトと入力候補は、自動システムを使ってコンテンツ ポリシーに対して評価されます。重大度の高いコンテンツはフィルターで除外されます。

責任ある AI

Microsoft は、人を第一に考える原則に基づいて、AI の発展に取り組んでいます。Azure OpenAI で使用できる生成モデルには、かなりの潜在的利益がありますが、慎重な設計と熟考した軽減策がない場合、そのようなモデルによって、正しくない、または有害なコンテンツが生成される可能性があります。Microsoft は、悪用や意図しない損害から保護するために多大な投資を行っています。たとえば、明確に定義したユースケースを示すことを申請者の要件とする、[責任ある AI 使用に関する Microsoft の原則](#)を取り入れる、顧客をサポートするコンテンツ フィルターを構築する、オンボードされた顧客に対して責任ある AI 実装のガイダンスを提供するなどです。

Azure OpenAI にアクセスするにはどうすればよいですか？

Azure OpenAI にアクセスするにはどうすればよいですか？

高い需要、今後の製品の機能強化、[Microsoft の責任ある AI へのコミットメント](#)を考慮し、現在、アクセスは制限されています。現在のところ、Microsoft と既存のパートナーシップ関係があるお客様、リスクの低いユース ケース、軽減策の取り入れに取り組んでいるお客様を対象としています。

より具体的な情報は、申請フォームに記載されています。Azure OpenAI に対するアクセスを拡大できるよう、責任を持って取り組んでいますので、しばらくお待ちください。

アクセスはこちらからお申し込みください。

[\[今すぐ適用する\]](#)

Azure OpenAI と OpenAI の比較

Azure OpenAI Service では、OpenAI GPT-4、GPT-3、Codex、DALL-E モデルを使用した高度な言語 AI を顧客に提供し、Azure のセキュリティとエンタープライズの約束を実現します。Azure OpenAI は OpenAI と共に API を共同開発し、互換性を確保し、一方から他方へのスムーズな移行を保証します。

Azure OpenAI を使用すると、顧客は OpenAI と同じモデルを実行しながら、Microsoft Azure のセキュリティ機能を使用できます。Azure OpenAI では、プライベート ネットワーク、リージョンの可用性、責任ある AI コンテンツのフィルター処理が提供されます。

主要な概念

プロンプトと入力候補

入力候補エンドポイントは、API サービスのコア コンポーネントです。この API は、モデルのテキストイン、テキストアウト インターフェイスへのアクセスを提供します。ユーザーは、英語のテキスト コマンドを含む入力**プロンプト**を入力するだけで、モデルによってテキスト**入力候補**が生成されます。

単純なプロンプトと入力候補の例を次に示します。

プロンプト: `"" count to 5 in a for loop ""`

入力候補: `for i in range(1, 6): print(i)`

トークン

Azure OpenAI では、テキストをトークンに分割して処理します。トークンには、単語または文字のチャンクのみを指定できます。たとえば、"hamburger" という単語はトークン "ham"、"bur"、"ger" に分割されますが、"pear" のような短くて一般的な単語は 1 つのトークンです。多くのトークンは、"hello" や "bye" などの空白で始まります。

所与の要求で処理されるトークンの合計数は、入力、出力、および要求パラメーターの長さによって異なります。処理されるトークンの量は、モデルの応答待機時間とスループットにも影響します。

リソース

Azure OpenAI は、Azure の新しい製品オファリングです。Azure OpenAI は、他の Azure 製品と同じように、Azure サブスクリプションにこのサービス用の[リソースまたはインスタンスを作成](#)して使用を開始できます。Azure の[リソース管理設計](#)について詳しくご覧いただけます。

デプロイメント

Azure OpenAI リソースを作成したら、API 呼び出しを開始してテキストを生成する前に、モデルをデプロイする必要があります。このアクションは、Deployment API を使用して実行できます。これらの API を使用すると、使用するモデルを指定できます。

プロンプト エンジニアリング


OpenAI の GPT-3、GPT-3.5、GPT-4 モデルは、プロンプト ベースです。プロンプト ベースのモデルでは、ユーザーはテキスト プロンプトを入力してモデルと対話し、モデルはテキスト入力候補でそれに応答します。この入力候補は、入力テキストに対してモデルが続けたものです。

これらのモデルは非常に強力ですが、その動作もプロンプトに対して非常に敏感です。このため、[プロンプトエンジニアリング](#)が開発のための重要なスキルになります。

プロンプトの構築は難しい場合があります。実際には、プロンプトは目的のタスクを完了するためにモデルの重みを構成するように機能しますが、これは科学というより芸術であり、多くの場合、成功するプロンプトを作成するには経験と直感が必要になります。

モデル

このサービスでは、ユーザーはいくつかのモデルにアクセスできます。各モデルには、異なる機能と価格ポイントが用意されています。

GPT-4 モデルは、利用可能な最新のモデルです。このモデル シリーズへのアクセスの需要がとて多いため、現在はリクエストによってのみ使用できます。アクセスをリクエストする場合、既存の Azure OpenAI のお客様は[こちらのフォームに入力して申請](#)  できます

GPT-3 ベース モデルは、Davinci、Curie、Babbage、Ada と呼ばれます (機能では降順、速度では昇順)。

Codex シリーズのモデルは GPT-3 の後継であり、自然言語とコードの両方でトレーニングされ、自然言語からコードへのユース ケースに役立ちます。各モデルの詳細については、[モデルの概念に関するページ](#)を参照してください。

DALL-E モデルは、現在プレビュー段階にあり、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。

次の手順

[Azure OpenAI をサポートする基となるモデル](#)に関する記事を確認します。

Azure OpenAI Service のクォータと制限

[アーティクル] • 2023/06/27

この記事には、Azure Cognitive Services 内の Azure OpenAI のクォータと制限に関するクイック リファレンスおよび詳細な説明が記載されています。

クォータと制限のリファレンス

以降のセクションでは、Azure OpenAI に適用されるデフォルトのクォータと制限のクイック ガイドを提供します。

制限名	制限値
各 Azure サブスクリプションのリージョンあたりの OpenAI リソース数	30
モデルとリージョンあたりのデフォルトのクォータ (1 分あたりのトークン数) ¹	Text-Davinci-003: 120 K GPT-4: 20 K GPT-4-32K: 60 K その他すべて: 240 K
既定の DALL-E クォータ制限	2 同時要求
要求あたりの最大プロンプト トークン数	モデルごとに異なります。詳細については、「 Azure OpenAI Service モデル 」を参照してください。
微調整されたモデル デプロイの最大数	2
リソースあたりのトレーニング ジョブの合計数	100
リソースあたりの同時実行トレーニング ジョブの最大数	1
キューに入ったトレーニング ジョブの最大数	20
リソースあたりの最大ファイル数	30
リソースあたりのすべてのファイルの合計サイズ	1 GB
トレーニングジョブの最大時間 (超過した場合、ジョブは失敗します)	720 時間

制限名	制限値
トレーニング ジョブの最大サイズ (トレーニング ファイル内のトークン) x (エポックの数)	20 億

¹ デフォルトのクォータ制限は変更される可能性があります。

レート制限内に収まるようにするための一般的なベストプラクティス

レート制限に関連する問題を最小限に抑えるには、次の手法を使用することをお勧めします。

- アプリケーションで再試行ロジックを実装します。
- ワークロードが急激に変化しないようにします。ワークロードは徐々に増やします。
- さまざまな負荷増加パターンをテストします。
- デプロイに割り当てられているクォータを増やします。必要に応じて、別のデプロイからクォータを移動します。

既定のクォータと制限の引き上げを要求する方法

クォータの増加要求は、Azure OpenAI Studio の [\[クォータ\]](#) ページから送信できます。現在、需要が多いため、新しいクォータ増加要求は承認されません。要求は、後で処理できるようになるまでキューに入れられます。

その他のレート制限については、[サービス リクエストを送信](#)してください。

次のステップ

Azure OpenAI デプロイの [クォータを管理](#)する方法を確認してください。 [Azure OpenAI をサポートする基となるモデル](#)に関する記事を確認します。

Azure OpenAI Service モデル

[アーティクル] • 2023/06/21

Azure OpenAI を使うと、ファミリーや機能ごとにグループ化されたさまざまなモデルにアクセスできます。モデル ファミリは、通常、目的のタスクによってモデルを関連付けます。次の表は、Azure OpenAI で現在使用できるモデル ファミリについて説明したものです。現在、すべてのリージョンですべてのモデルを使用できるわけではありません。詳細については、この記事の[モデル機能の表](#)を参照してください。

モデルファミリー	説明
GPT-4	GPT-3.5 を基に改善され、自然言語とコードを生成するだけでなく、理解できるモデルのセット。
GPT-3	自然言語を理解し、生成できるモデルのシリーズ。これには、新しい ChatGPT モデル が含まれます。
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。
Codex	自然言語のコードへの変換を含め、コードを理解し、生成できるモデルのシリーズ。
埋め込み	埋め込みを理解し、使用できるモデルのセット。埋め込みは、機械学習モデルとアルゴリズムで簡単に利用できる特別な形式のデータ表現です。埋め込みは、テキストの意味論的意味の情報密度の高い表現です。現在、異なる機能に対応する埋め込みモデルの 3 つのファミリー (類似性、テキスト検索、コード検索) を提供しています。

モデル機能

各モデル ファミリには、さらに機能によって区別されるモデルのシリーズがあります。通常、これらの機能は名前で識別されます。また、これらの名前のアルファベット順は、一般的に、特定のモデル ファミリ内のそのモデルの相対的な機能とコストを示します。たとえば GPT-3 モデルの場合、Ada、Babbage、Curie、Davinci という名前を使って相対的な機能とコストを示しています。Davinci は Curie よりも高性能で高額、Curie は Babbage よりも高性能で高額、などと続きます。

ⓘ 注意

Ada のような低い能力のモデルで実行できるタスクは、Curie や Davinci のような高い能力のモデルでも実行できます。

名称に関する規則

通常、Azure OpenAI のモデル名は、次のような標準的な名前付け規則に対応しています。

```
{capability}-{family}[-{input-type}]-{identifier}
```

要素	説明
{capability}	モデルのモデル機能。たとえば、 GPT-3 モデル の場合は <code>text</code> を使い、 Codex モデル の場合は <code>code</code> を使います。
{family}	モデルの相対ファミリー。たとえば、GPT-3 モデルには <code>ada</code> 、 <code>babbage</code> 、 <code>curie</code> 、 <code>davinci</code> などがあります。
{input-type}	(埋め込みモデル のみ) モデルがサポートする埋め込みの入力の種類。たとえば、テキスト検索の埋め込みモデルは <code>doc</code> と <code>query</code> をサポートしています。
{identifier}	モデルのバージョン識別子。

たとえば、Microsoft の最も強力な GPT-3 モデルは `text-davinci-003` と名前であり、最も強力な Codex モデルは `code-davinci-002` という名前です。

`ada`、`babbage`、`curie`、`davinci` という旧バージョンの GPT-3 モデルは標準的な名前付け規則に従っておらず、主に微調整を目的としています。詳細については、「[アプリケーションのモデルをカスタマイズする方法について](#)」を参照してください。

使用可能なモデルの検索

[Models List API](#) を使用して、Azure OpenAI リソースによる推論と微調整の両方に使用できるモデルの一覧を取得できます。

モデルの更新

Azure OpenAI では、選択したモデル デプロイの自動更新がサポートされるようになりました。自動更新サポートが利用可能なモデルでは、Azure OpenAI Studio の **[新しいデプロイを作成する]** と **[デプロイの編集]** にモデル バージョンのドロップダウンが表示されます。

Deploy model

×

Set up a deployment to make API calls against a provided base model or a custom model. Finished deployments are available for use. Your deployment status will move to succeeded when the deployment is complete and ready for use.

Select a model ⓘ

text-embedding-ada-002

Model version ⓘ

Auto-update to latest

Auto-update to latest

2

1

CreateCancel

最新に自動更新する

[Auto-update to latest] (**最新に自動更新**) が選択されている場合、モデル デプロイは、新しいバージョンがリリースされてから 2 週間以内に自動的に更新されます。

まだ Completion と Chat Completion に基づくモデルの早期テスト フェーズである場合は、使用可能な場合は常に、**最新に自動更新**を設定してモデルをデプロイすることをお勧めします。埋め込みモデルの場合は、最新バージョンのモデルを使用することをお勧めしますが、以前のモデル バージョンで生成された埋め込みは新しいバージョンと互換性がないため、アップグレードするタイミングを選択する必要があります。

特定のモデル バージョン

Azure OpenAI の使用が進み、アプリケーションの構築と統合が始まる中で、モデルの更新を手動で制御すると、モデルのパフォーマンスがユース ケースに対して一貫していることを最初にテストおよび検証してから、アップグレードできるようになります。

デプロイに特定のモデル バージョンを選択すると、自分で手動更新するか、モデルの有効期限に達するまで、このバージョンは選択されたままになります。非推奨となる

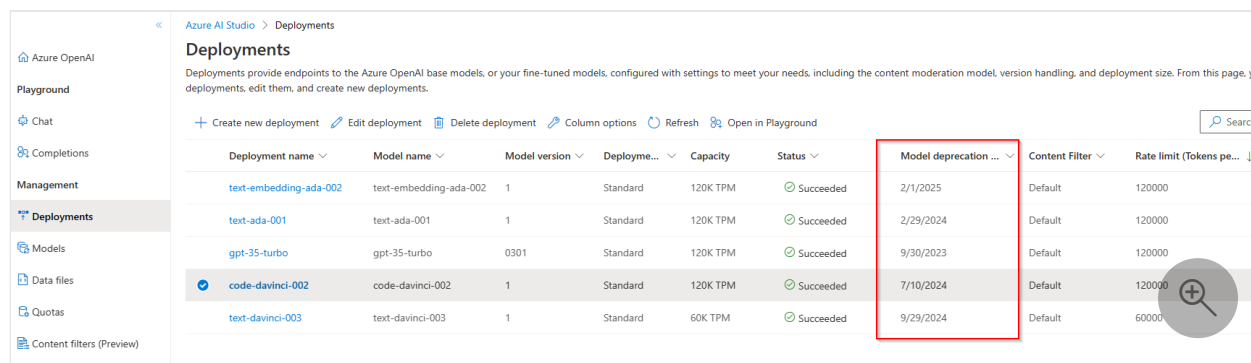
日/有効期限に達すると、モデルは利用可能な最新バージョンに自動アップグレードされます。

GPT-35-Turbo 0301 および GPT-4 0314 の有効期限

元の gpt-35-turbo (0301) モデルと両方の gpt-4 (0314) モデルは、2023 年 9 月 30 日以降に期限切れになります。有効期限が切れると、デプロイはその時点で既定のバージョンに自動的にアップグレードされます。アップグレードではなく完了要求の受け入れを停止するようにデプロイする場合は、API を使用してモデルのアップグレードオプションを期限切れに設定できます。これに関するガイドラインは 9 月 1 日までに公開されます。

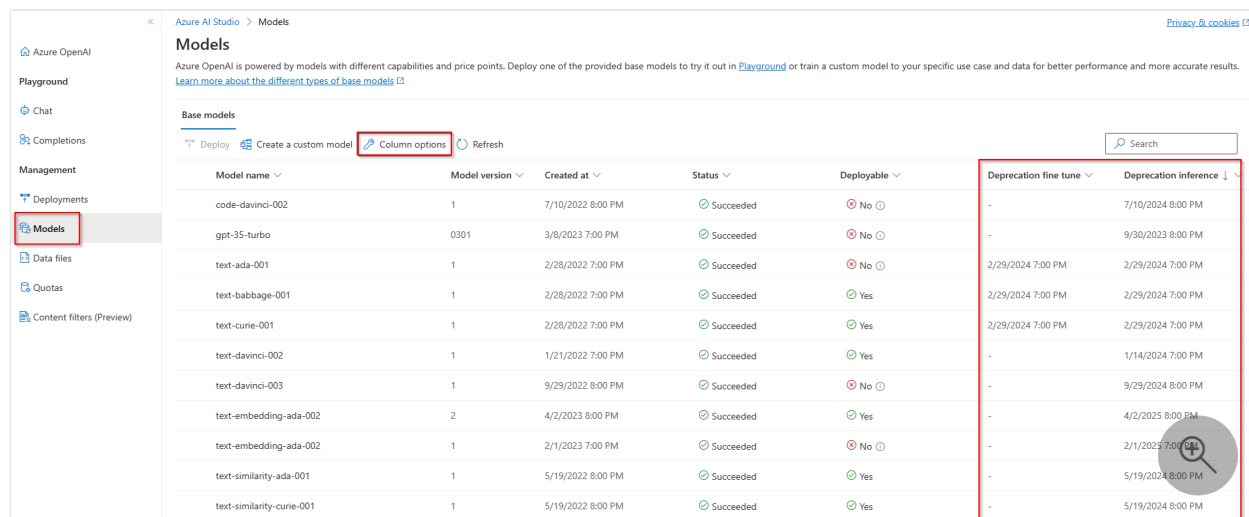
非推奨になる日付の表示

現在デプロイされているモデルの場合は、Azure OpenAI Studio から [デプロイ] を選択します。



Deployment name	Model name	Model version	Deploye...	Capacity	Status	Model deprecation ...	Content Filter	Rate limit (Tokens pe...
text-embedding-ada-002	text-embedding-ada-002	1	Standard	120K TPM	Succeeded	2/1/2025	Default	120000
text-ada-001	text-ada-001	1	Standard	120K TPM	Succeeded	2/29/2024	Default	120000
gpt-35-turbo	gpt-35-turbo	0301	Standard	120K TPM	Succeeded	9/30/2023	Default	120000
code-davinci-002	code-davinci-002	1	Standard	120K TPM	Succeeded	7/10/2024	Default	120000
text-davinci-003	text-davinci-003	1	Standard	60K TPM	Succeeded	9/29/2024	Default	60000

Azure OpenAI Studio から特定のリージョンで使用可能なすべてのモデルの非推奨となる日/有効期限を表示するには、[モデル]>[列のオプション]>[Deprecation fine tune] (非推奨の微調整) と [Deprecation inference] (非推奨の推定) を選択します。



Model name	Model version	Created at	Status	Deployable	Deprecation fine tune	Deprecation inference
code-davinci-002	1	7/10/2022 8:00 PM	Succeeded	No	-	7/10/2024 8:00 PM
gpt-35-turbo	0301	3/8/2023 7:00 PM	Succeeded	No	-	9/30/2023 8:00 PM
text-ada-001	1	2/28/2022 7:00 PM	Succeeded	No	2/29/2024 7:00 PM	2/29/2024 7:00 PM
text-babbage-001	1	2/28/2022 7:00 PM	Succeeded	Yes	2/29/2024 7:00 PM	2/29/2024 7:00 PM
text-curie-001	1	2/28/2022 7:00 PM	Succeeded	Yes	2/29/2024 7:00 PM	2/29/2024 7:00 PM
text-davinci-002	1	1/21/2022 7:00 PM	Succeeded	Yes	-	1/14/2024 7:00 PM
text-davinci-003	1	9/29/2022 8:00 PM	Succeeded	No	-	9/29/2024 8:00 PM
text-embedding-ada-002	2	4/2/2022 8:00 PM	Succeeded	Yes	-	4/2/2025 8:00 PM
text-embedding-ada-002	1	2/1/2023 7:00 PM	Succeeded	No	-	2/1/2025 7:00 PM
text-similarity-ada-001	1	5/19/2022 8:00 PM	Succeeded	Yes	-	5/19/2024 8:00 PM
text-similarity-curie-001	1	5/19/2022 8:00 PM	Succeeded	Yes	-	5/19/2024 8:00 PM

API を使用してモデルを更新しデプロイする

HTTP

PUT

```
https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.CognitiveServices/accounts/{accountName}/deployments/{deploymentName}?api-version=2023-05-01
```

パス パラメーター

パラメーター	Type	必須	説明
<code>accountname</code>	string	必須	Azure OpenAI リソースの名前。
<code>deploymentName</code>	string	必須	既存のモデルをデプロイしたときに選択したデプロイ名、または新しいモデル デプロイに使用する名前。
<code>resourceGroupName</code>	string	必須	このモデル デプロイに関連付けられているリソース グループの名前。
<code>subscriptionId</code>	string	必須	関連付けられているサブスクリプションの ID。
<code>api-version</code>	string	必須	この操作に使用する API バージョン。これは、YYYY-MM-DD 形式に従います。

サポートされているバージョン

- 2023-05-01 [Swagger の仕様](#)

要求本文

これは、使用可能な要求本文パラメーターのサブセットにすぎません。すべてのパラメーターの一覧については、[REST API 仕様](#)を参照してください。

パラメーター	Type	説明
<code>versionUpgradeOption</code>	String	デプロイ モデル バージョンのアップグレード オプション: <code>OnceNewDefaultVersionAvailable</code> <code>OnceCurrentVersionExpired</code> <code>NoAutoUpgrade</code>
<code>capacity</code>	整数 (integer)	このデプロイに割り当てる クォータ の量を表します。値 1 は、1 分あたり 1,000 トークン (TPM) に相当します

要求の例

Bash

```
curl -X PUT https://management.azure.com/subscriptions/00000000-0000-0000-0000-000000000000/resourceGroups/resource-group-temp/providers/Microsoft.CognitiveServices/accounts/docs-openai-test-001/deployments/text-embedding-ada-002-test-1" \  
-H "Content-Type: application/json" \  
-H 'Authorization: Bearer YOUR_AUTH_TOKEN' \  
-d '{"sku":{"name":"Standard","capacity":1},"properties":{"model":{"format": "OpenAI","name": "text-embedding-ada-002","version": "2"},"versionUpgradeOption":"OnceCurrentVersionExpired"}}'
```

ⓘ 注意

認証トークンを生成するには、複数の方法があります。初期テストの最も簡単な方法は、<https://portal.azure.com> から Cloud Shell を起動することです。次に、`az account get-access-token` を実行します。このトークンは、API テストの一時的な認証トークンとして使用できます。

応答の例

JSON

```
{  
  "id": "/subscriptions/{subscription-id}/resourceGroups/resource-group-temp/providers/Microsoft.CognitiveServices/accounts/docs-openai-test-001/deployments/text-embedding-ada-002-test-1",  
  "type": "Microsoft.CognitiveServices/accounts/deployments",  
  "name": "text-embedding-ada-002-test-1",  
  "sku": {  
    "name": "Standard",  
    "capacity": 1  
  },  
  "properties": {  
    "model": {  
      "format": "OpenAI",  
      "name": "text-embedding-ada-002",  
      "version": "2"  
    },  
    "versionUpgradeOption": "OnceCurrentVersionExpired",  
    "capabilities": {  
      "embeddings": "true",  
      "embeddingsMaxInputs": "1"  
    },  
    "provisioningState": "Succeeded",  
    "ratelimits": [  

```

```
{
  "key": "request",
  "renewalPeriod": 10,
  "count": 2
},
{
  "key": "token",
  "renewalPeriod": 60,
  "count": 1000
}
]
},
"systemData": {
  "createdBy": "docs@contoso.com",
  "createdByType": "User",
  "createdAt": "2023-06-13T00:12:38.885937Z",
  "lastModifiedBy": "docs@contoso.com",
  "lastModifiedByType": "User",
  "lastModifiedAt": "2023-06-13T02:41:04.8410965Z"
},
"etag": "\"{GUID}\""
}
```

適切なモデルを見つける

モデル ファミリで最も機能性の高いモデルから始めて、モデルの機能が要件を満たしているかどうかを確認することをお勧めします。その後、そのモデルを使い続けることも、より低い機能とコストのモデルに移行してそのモデルの機能を中心に最適化することもできます。

GPT-4 モデル

GPT-4 は、OpenAI の以前のどのモデルよりも高い精度で難しい問題を解決することができます。gpt-35-turbo と同様に、GPT-4 はチャット用に最適化されていますが、従来の補完タスクでも適切に動作します。

このモデル シリーズは、高いアクセス需要により、現在はリクエストによってのみ使用できます。アクセスをリクエストする場合、既存の Azure OpenAI のお客様は[こちらのフォームに入力して申請](#)してください。

- gpt-4
- gpt-4-32k

gpt-4 では最大 8192 個の入力トークンがサポートされ、gpt-4-32k では最大 32,768 個のトークンがサポートされます。

GPT-3 モデル

GPT-3 モデルは、自然言語を理解および生成できます。このサービスには、さまざまなタスクに応じて能力と速度のレベルが異なる 4 つのモデル機能が用意されています。Davinci が最も能力の高いモデルであり、Ada が最速です。機能性の高いものから低い順に、モデルは以下のとおりです。

- `text-davinci-003`
- `text-curie-001`
- `text-babbage-001`
- `text-ada-001`

Davinci は最も能力が高い一方で、他のモデルには大幅な速度の利点があります。実験時には Davinci から始めることをお勧めします。なぜなら、最も良い結果を出し、Azure OpenAI が提供できる価値がわかるからです。プロトタイプが動作したら、アプリケーションの待機時間とパフォーマンスの最適なバランスでモデルの選択を最適化できます。

Davinci

Davinci は最も能力の高いモデルであり、他のモデルが実行できるすべてのタスクを実行でき、多くの場合、より少ない命令で実行できます。特定の対象ユーザー向けの要約やクリエイティブなコンテンツ生成など、コンテンツを深く理解する必要があるアプリケーションの場合、Davinci は最適な結果を生成します。Davinci が提供する強力な機能には、より多くのコンピューティング リソースが必要になるため、Davinci のコストは高くなり、他のモデルほど高速ではありません。

Davinci が優れているもう 1 つの領域は、テキストの意図を理解することです。Davinci は、多くの種類のロジックの問題を解決し、文字の動機を説明することに優れています。Davinci は、因果関係に関連する最も困難な AI の問題のいくつかを解決しています。

用途: 複雑な意図、因果関係、対象者向けの要約

Curie

Curie is は強力かつ高速です。複雑なテキストの分析に関しては Davinci の方が強力ですが、Curie は感情分類や要約などの多くの微妙なタスクに対して優れた能力を発揮します。Curie はさらに、一般的なサービス チャットボットとして質問に答えたり、Q&A を実行したりすることにも優れています。

用途: 言語翻訳、複雑な分類、テキスト センチメント、要約

Babbage

Babbage は、単純な分類などの簡単なタスクを実行できます。また、ドキュメントが検索クエリとどの程度一致しているかを示すセマンティック検索のランク付けにも対応できます。

用途: 中程度の分類、セマンティック検索分類

Ada

Ada は通常、最速のモデルであり、テキストの解析、住所変更、過度のニュアンスを必要としない特定の種類の分類などのタスクを実行できます。Ada のパフォーマンスは、多くの場合、より多くのコンテキストを提供することで改善できます。

用途: テキストの解析、単純な分類、住所変更、キーワード

ChatGPT (gpt-35-turbo)

ChatGPT モデル (gpt-35-turbo) は会話型インターフェイス用に設計された言語モデルであり、モデルの動作は以前の GPT-3 モデルとは異なります。以前のモデルはテキストインとテキストアウトでした。つまり、プロンプト文字列を受け入れ、プロンプトに追加する入力候補を返しました。しかし、ChatGPT モデルはカンバセーションインとメッセージアウトです。このモデルでは、チャットのような特定のトランスクリプト形式でフォーマットされたプロンプト文字列が必要であり、チャット内のモデルで記述されたメッセージを表す入力候補を返します。

ChatGPT モデルの詳細と Chat API の操作方法については、[詳細なハウツー](#)をご覧ください。

DALL-E モデル

DALL-E モデルは、現在プレビュー段階にあり、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。

Codex モデル

Codex モデルは、コードを理解して生成できる基本 GPT-3 モデルの子孫です。トレーニングデータには、自然言語と、GitHub からの数十億行のパブリックコードの両方が含まれています。

これらは Python で最も能力を発揮し、C#、JavaScript、Go、Perl、PHP、Ruby、Swift、TypeScript、SQL、シェルなど、12 以上の言語に精通しています。機能性の高いものから低い順に、Codex モデルは以下のとおりです。

- `code-davinci-002`
- `code-cushman-001`

Davinci

GPT-3 と同様に、Davinci は最も能力の高い Codex モデルであり、他のモデルが実行できるすべてのタスクを実行でき、多くの場合、より少ない命令で実行できます。コンテンツの深い理解が必要なアプリケーションの場合、Davinci を使うと、最も良い結果を出せます。強力な機能にはより多くのコンピューティング リソースが必要になるため、Davinci のコストは高くなり、他のモデルほど高速ではありません。

Cushman

Cushman は強力かつ高速です。複雑なタスクを分析する場合には Davinci の方が強力ですが、Cushman は多くのコード生成タスクに対応できる高い能力のモデルです。通常、Cushman は Davinci よりも実行が高速で低コストです。

埋め込みモデル

① 重要

`text-embedding-ada-002 (Version 2)` を使用することを強くお勧めします。このモデル/バージョンでは、OpenAI の `text-embedding-ada-002` と同等の機能が提供されます。このモデルによって提供される機能強化の詳細については、[OpenAI のブログ記事](#) を参照してください。現在バージョン 1 を使用している場合でも、最新の重みや更新されたトークン制限を利用するには、バージョン 2 に移行する必要があります。バージョン 1 とバージョン 2 は互換性がないため、同じバージョンのモデルを使用してドキュメントの埋め込みとドキュメント検索を行う必要があります。

現在、異なる機能を持つ 3 つの埋め込みモデル ファミリを提供しています。

- [Similarity](#)
- [テキスト検索](#)
- [コード検索](#)

各ファミリには、さまざまな機能のモデルが含まれています。次の一覧は、サービスから返される数値ベクトルの長さを、モデルの機能に基づいて示したものです。

基本モデル	モデル	Dimensions
Ada	-001 で終わるモデル (バージョン 1)	1024
Ada	text-embedding-ada-002 (バージョン 2)	1536
Babbage		2048
Curie		4096
Davinci		12288

Davinci は最も能力が高いですが、他のモデルよりも遅く、高価です。Ada は最も能力が低いですが、高速かつ安価です。

類似性埋め込み

これらのモデルは、2 つ以上のテキスト間のセマンティック類似性を捉えることに適しています。

ユース ケース	モデル
クラスタリング、回帰、異常検出、視覚化	<code>text-similarity-ada-001</code> <code>text-similarity-babbage-001</code> <code>text-similarity-curie-001</code> <code>text-similarity-davinci-001</code>

テキスト検索埋め込み

これらのモデルは、長いドキュメントが短い検索クエリに関連しているかどうかを測定するのに役立ちます。このファミリがサポートする入力の種類は 2 つあります。取得するドキュメントを埋め込むための `doc` と、検索クエリを埋め込むための `query` です。

ユース ケース	モデル
---------	-----

ユースケース	モデル
検索、コンテキストの関連性、情報の取得	text-search-ada-doc-001 text-search-ada-query-001 text-search-babbage-doc-001 text-search-babbage-query-001 text-search-curie-doc-001 text-search-curie-query-001 text-search-davinci-doc-001 text-search-davinci-query-001

コード検索埋め込み

テキスト検索埋め込みモデルと同様に、このファミリがサポートする入力の種類は 2 つあります。取得するコード スニペットを埋め込むための `code` と、自然言語検索クエリを埋め込むための `text` です。

ユースケース	モデル
コード検索と関連性	code-search-ada-code-001 code-search-ada-text-001 code-search-babbage-code-001 code-search-babbage-text-001

埋め込みモデルを使用する場合は、制限事項とリスクに留意してください。

モデルの概要テーブルとリージョンの可用性

📌 重要

米国中南部は、高い需要により、新しいリソースの作成で一時的に使用できなくなっています。

GPT-3 モデル

これらのモデルは Completion API 要求で使用できます。 `gpt-35-turbo` は、Completion API 要求と Chat Completion API の両方で使用できる唯一のモデルです。

モデル ID	基本モデルのリージョン	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)
		整		

モデル ID	基本モデルのリージョン	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)
ADA	該当なし	該当なし	2,049	2019 年 10 月
text-ada-001	米国東部、米国中南部、西ヨーロッパ	該当なし	2,049	2019 年 10 月
Babbage	該当なし	該当なし	2,049	2019 年 10 月
text-babbage-001	米国東部、米国中南部、西ヨーロッパ	該当なし	2,049	2019 年 10 月
Curie	該当なし	該当なし	2,049	2019 年 10 月
text-curie-001	米国東部、米国中南部、西ヨーロッパ	該当なし	2,049	2019 年 10 月
davinci	該当なし	該当なし	2,049	2019 年 10 月
text-davinci-001	米国中南部、西ヨーロッパ	該当なし		
text-davinci-002	米国東部、米国中南部、西ヨーロッパ	該当なし	4,097	2021 年 6 月
text-davinci-003	米国東部、西ヨーロッパ	該当なし	4,097	2021 年 6 月
text-davinci-fine-tune-002	該当なし	該当なし		
gpt-35-turbo ¹ (ChatGPT)	米国東部、フランス中部、米国中南部、英国南部、西ヨーロッパ	該当なし	4,096	2021 年 9 月

¹ 現在、このモデルのバージョン **0301** のみが使用できます。

GPT-4 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 ID	基本モデルのリージョン	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)
gpt-4 ^{1,2}	米国東部、フランス中部	該当なし	8,192	2021 年 9 月

モデル ID	基本モデルのリージョン	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>gpt-4-32k</code> ^{1,2}	米国東部、フランス中部	該当なし	32,768	2021 年 9 月

¹ このモデルは[リクエストによってのみ使用できます](#)。

² 現在、このモデルのバージョン `0314` のみが使用できます。

Dall-E モデル

モデル ID	基本モデルのリージョン	リージョンの微調整	最大要求数 (文字)	トレーニング データ (最大)
<code>dalle2</code>	East US	該当なし	1000	該当なし

Codex モデル

これらのモデルは Completion API 要求でのみ使用できます。

モデル ID	基本モデルのリージョン	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>code-cushman-001</code> ¹	米国中南部、西ヨーロッパ	現在は利用不可	2,048	
<code>code-davinci-002</code>	米国東部、西ヨーロッパ	該当なし	8,001	2021 年 6 月

¹ このモデルの微調整は、リクエストに応じて提供されます。現在、このモデルの微調整に関する新しいリクエストは受け付けていません。

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-ada-002` (Version 2) を使用することを強くお勧めします。このモデル/バージョンでは、OpenAI の `text-embedding-ada-002` と同等の機能が提供されます。このモデルによって提供される機能強化の詳細については、[OpenAI のブログ記事](#) を参照してください。現在バージョン 1 を使用している場合でも、最新の重みや更新されたトークン制限を利用するには、バージョン 2 に移行する

必要があります。バージョン 1 とバージョン 2 は互換性がないため、同じバージョンのモデルを使用してドキュメントの埋め込みとドキュメント検索を行う必要があります。

モデル ID	基本モデルのリージョン	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)
text-embedding-ada-002 (バージョン 2)	米国東部、米国中南部	該当なし	8,191	2021 年 9 月
text-embedding-ada-002 (バージョン 1)	米国東部、米国中南部、西ヨーロッパ	該当なし	2,046	2021 年 9 月
text-similarity-ada-001	米国東部、米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-similarity-babbage-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-similarity-curie-001	米国東部、米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-similarity-davinci-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-ada-doc-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-ada-query-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-babbage-doc-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-babbage-query-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-curie-doc-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-curie-query-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-davinci-doc-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
text-search-davinci-query-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月

モデル ID	基本モデルのリージョン	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-search-ada-code-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
code-search-ada-text-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
code-search-babbage-code-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月
code-search-babbage-text-001	米国中南部、西ヨーロッパ	該当なし	2,046	2020 年 8 月

次のステップ

- [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](#)

Azure OpenAI Service の新機能

[アーティクル] • 2023/06/19

2023 年 6 月

英国南部

- Azure OpenAI が米国南部リージョンで使えるようになりました。各リージョンでのモデル提供状況の最新情報は、[モデルのページ](#)をご確認ください。

コンテンツのフィルターと注釈 (プレビュー)

- Azure OpenAI Service で[コンテンツ フィルター](#)を構成する方法
- [注釈を有効に](#)して、GPT ベースの Completion 呼び出しと Chat Completion 呼び出しの一部としてコンテンツ フィルター カテゴリと重大度情報を表示します。

Quota

- クォータを使用すると、サブスクリプション内の[デプロイ全体で、レート制限の割り当てを柔軟に管理](#)できます。

2023 年 5 月

Java & JavaScript SDK のサポート

- [JavaScript](#) と [Java](#) のサポートを提供する新しい Azure OpenAI プレビュー SDK。

Azure OpenAI Chat Completion の一般提供 (GA)

- 一般提供サポート:
 - Chat Completion API バージョン `2023-05-15`。
 - GPT-35-Turbo モデル。
 - GPT-4 モデル シリーズ。このモデル シリーズは需要が高いため、現在はリクエストがあった場合にのみ利用できます。アクセスをリクエストする場合、既存の Azure OpenAI のお客様は、[このフォームに入力することで申請](#)できます

現在 2023-03-15-preview API を使用している場合は、GA 2023-05-15 API に移行することをお勧めします。現在 API バージョン 2022-12-01 を使用している場合、この API は GA のままですが、最新の Chat Completion 機能は含まれません。

① 重要

補完エンドポイントでの GPT-35-Turbo モデルの現在のバージョンの使用は、プレビュー段階のままです。

フランス中部

- Azure OpenAI がフランス中部リージョンで使えるようになりました。各リージョンでのモデル提供状況の最新情報は、[モデルのページ](#)をご確認ください。

2023 年 4 月

- **DALL-E 2 パブリックプレビュー。** Azure OpenAI Service では、OpenAI の DALL-E 2 モデルを利用したイメージ生成 API がサポートされるようになりました。指定した説明テキストに基づいて、AI によって生成されたイメージを取得します。詳細については、[クイックスタート](#)を参照してください。アクセスをリクエストする場合、既存の Azure OpenAI のお客様は、[このフォームに入力することで申請](#) できます。
- **カスタマイズされたモデルの非アクティブなデプロイは、15 日後に削除されます。モデルは引き続き再デプロイに使用できます。** カスタマイズされた (微調整された) モデルが 15 日間を超えてデプロイされ、候補呼び出しやチャット候補呼び出しが行われなかった場合、デプロイは自動的に削除されます (そのデプロイに対するホスティング料金は発生しません)。基になるカスタマイズされたモデルは引き続き使用でき、いつでも再デプロイできます。詳しくは、[操作方法に関する記事](#)をご覧ください。

2023 年 3 月

- **GPT-4 シリーズ モデルは、Azure OpenAI でプレビューで利用できるようになりました。** アクセスをリクエストする場合、既存の Azure OpenAI のお客様は、[このフォームに入力することで申請](#) できます。これらのモデルは現在、米国東部と米国中南部のリージョンで使用できます。
- **3 月 21 日にプレビューでリリースされた、ChatGPT および GPT-4 モデル用の新しいチャット補完 API。** 詳細については、[更新されたクイックスタートと操作方](#)

[法に関する記事](#)を参照してください。

- **ChatGPT (gpt-35-turbo) プレビュー**。詳細については、[操作方法に関する記事](#)を確認してください。
- 微調整のためにトレーニング制限を増加: トレーニング ジョブの最大サイズ (トレーニング ファイル内のトークン) x (エポック数) は、すべてのモデルに対して 20 億トークンになりました。また、最大トレーニング ジョブを 120 時間から 720 時間に増やしました。
- 既存のアクセス権へのユース ケースの追加。以前は、新しいユース ケースを追加するプロセスで、お客様がサービスに再適用する必要がありました。現在、サービスの使用に新しいユース ケースを迅速に追加できる、新しいプロセスをリリースしています。このプロセスは、Azure Cognitive Services 内で確立されている制限付きアクセス プロセスに従っています。[既存のお客様は、こちらからすべての新しいユース ケースを証明できます](#)。これは、最初に申請しなかった新しいユース ケースでサービスを使用するときに必ず必要になるので注意してください。

2023 年 2 月

新機能

- .NET SDK (推論) の[プレビュー リリース](#) | [サンプル](#)
- Azure OpenAI 管理操作をサポートするための [Terraform SDK の更新](#)。
- `suffix` パラメーターを使用して入力候補の末尾にテキストを挿入できるようになりました。

更新プログラム

- コンテンツのフィルター処理が既定でオンになっています。

次に関する新しい記事:

- [Azure OpenAI Service を監視する](#)
- [Azure OpenAI のコストを計画および管理する](#)

新しいトレーニング コース:

- [Azure OpenAI の概要](#)

2023 年 1 月

新機能

- **サービス GA。** Azure OpenAI Service が一般提供になりました。
- **新しいモデル:** 最新のテキスト モデル text-davinci-003 (米国東部、西ヨーロッパ)、text-ada-embeddings-002 (米国東部、米国中南部、西ヨーロッパ) の追加

2022 年 12 月

新機能

- **OpenAI の最新モデル。** Azure OpenAI を使うと、GPT-3.5 シリーズを含むすべての最新モデルにアクセスできます。
- **新しい API バージョン (2022-12-01)。** この更新プログラムには、リクエストをいただいていた機能強化がいくつか含まれています。たとえば、API 応答でのトークン使用情報、ファイルのエラー メッセージの改善、作成データ構造の微調整に関する OpenAI との整合、微調整されたジョブのカスタム名前付けを可能にする suffix パラメーターのサポートなどです。
- **1 秒あたりの要求数の上限を引き上げました。** 非 Davinci モデルの場合は 50。Davinci モデルの場合は 20。
- **デプロイの微調整を高速化しました。** Ada と Curie の微調整されたモデルを 10 分未満でデプロイできます。
- **トレーニング上限を引き上げました:** Ada、Babbage、Curie の場合は 40M トレーニング トークン。Davinci の場合は 10M。
- **データ ログと人間によるレビューの不正使用と誤用に対する変更要求のプロセス。** 現在、このサービスでは、これらの強力なモデルが不正使用されないように、不正使用と誤用を検出する目的で要求と応答のデータをログしています。ただし、多くのお客様はデータのプライバシーとセキュリティの要件が厳格なので、データをより細かく管理する必要があります。このようなユース ケースをサポートするために、お客様がコンテンツ フィルター処理ポリシーを変更することや、低リスクのユース ケースで不正使用ログをオフにすることができる新しいプロセスをリリースしています。このプロセスは、Azure Cognitive Services 内で確立されている制限付きアクセス プロセスに従っているため、[既存の OpenAI のお客様はこちらからお申し込みいただけます](#)。

- **カスタマー マネージド キー (CMK) の暗号化。** CMK にはトレーニング データとカスタマイズされたモデルの格納に使われる独自の暗号化キーがあるので、お客様は Azure OpenAI のデータ管理をより細かく制御できます。カスタマー マネージド キー (CMK、Bring Your Own Key (BYOK) と呼ばれます) を使用すると、アクセス制御の作成、ローテーション、無効化、取り消しを、いっそう柔軟に行うことができます。また、データを保護するために使われる暗号化キーを監査することもできます。 [詳細については、保存時の暗号化ドキュメントを参照してください。](#)
- **ロックボックスのサポート**
- **SOC-2 への準拠**
- Azure Resource Health、コスト分析、メトリックと診断の設定を使った**ログと診断**。
- **Studio の機能強化。** 微調整されたモデルの作成とデプロイにチーム内の誰がアクセスできるかを制御するための Azure AD ロール サポートを含め、Studio ワークフローのさまざまな点を使いやすくしました。

変更 (破壊的)

微調整: OpenAI のスキーマに合わせて、作成 API 要求が更新されました。

プレビュー API のバージョン:

JSON

```
{
  "training_file": "file-XGinujblHPwGLSztz8cPS8XY",
  "hyperparams": {
    "batch_size": 4,
    "learning_rate_multiplier": 0.1,
    "n_epochs": 4,
    "prompt_loss_weight": 0.1,
  }
}
```

API バージョン 2022-12-01:

JSON

```
{
  "training_file": "file-XGinujblHPwGLSztz8cPS8XY",
  "batch_size": 4,
  "learning_rate_multiplier": 0.1,
```

```
"n_epochs": 4,  
"prompt_loss_weight": 0.1,  
}
```

既定で**コンテンツのフィルター処理は一時的にオフ**です。 Azure コンテンツ モデレーションは、OpenAI とは異なる方法で動作します。 Azure OpenAI を使うと、生成呼び出し時にコンテンツ フィルターを実行し、有害な、または不正使用のコンテンツとフィルターを検出し、応答から除外することができます。 [詳細情報](#)

これらのモデルは 2023 年第 1 四半期に再び有効になり、既定でオンになります。

お客様のアクション

- お使いのサブスクリプションでこれらを有効にする場合は、[Azure サポートにお問い合わせください](#)。
- 無効のままにする場合は、[フィルター処理の変更をお申し込みください](#) (このオプションは低リスクのユース ケースに限定されます)。

次の手順

[Azure OpenAI をサポートする基となるモデル](#)に関する記事を確認します。

Azure OpenAI Service に関してよく寄せられる質問

よく寄せられる質問

このドキュメントで質問に対する回答が見つからず、さらにサポートが必要な場合は、[Cognitive Services のサポート オプションのガイド](#)を確認してください。 Azure OpenAI は Azure Cognitive Services の一部です。

データとプライバシー

モデルのトレーニングには自社のデータが使用されますか？

Azure OpenAI では、モデルの再トレーニングに顧客データは使用されません。 詳細については、[Azure OpenAI のデータ、プライバシー、セキュリティのガイド](#)を参照してください。

全般

Azure OpenAI では GPT-4 がサポートされていますか？

Azure OpenAI では、最新の GPT-4 モデルがサポートされています。 現在、これらのモデルはリクエストによってのみ使用できます。 既存の Azure OpenAI のお客様は、[このフォームに入力してアクセスを申請](#) できます。

Azure OpenAI の機能を OpenAI と比較するとどうですか？

Azure OpenAI Service では、OpenAI GPT-3、Codex、DALL-E モデルを使用した高度な言語 AI を顧客に提供し、Azure のセキュリティとエンタープライズの約束を実現します。 Azure OpenAI は OpenAI と共に API を共同開発し、互換性を確保し、一方から他方へのスムーズな移行を保証します。

Azure OpenAI を使用すると、顧客は OpenAI と同じモデルを実行しながら、Microsoft Azure のセキュリティ機能を使用できます。