

Azure OpenAI Service モデル

[アーティクル] • 2023/10/08

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ごとに異なります。2024 年 7 月に廃止される GPT-3 やその他のモデルについては、「[Azure OpenAI Service のレガシ モデル](#)」を参照してください。

モデル	説明
GPT-4	GPT-3.5 を基に改善され、自然言語とコードを理解し、生成できるモデルのセット。
GPT-3.5	GPT-3 を基に改善され、自然言語とコードを理解し、生成できるモデルのセット。
埋め込み	テキストを数値ベクトル形式に変換して、テキストの類似性を促進できるモデルのセット。
DALL-E (プレビュー)	自然言語からオリジナル画像を生成できるモデルのシリーズ (プレビュー)。
ささやき (プレビュー)	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。

GPT-4

GPT-4 は、オープンエイのこれまでのどのモデルよりも高い精度で難問を解くことができます。GPT-3.5 Turbo と同様に、GPT-4 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。Chat Completions API を使用して GPT-4 を使用します。GPT-4 と Chat Completions API の使用方法については、[詳細なハウツー](#)をご覧ください。

- `gpt-4`
- `gpt-4-32k`

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

GPT-3.5

GPT-3.5 モデルは、自然言語とコードを理解および生成できます。GPT-3.5 ファミリで最も能力とコスト効率の高いモデルは GPT-3.5 Turbo です。これはチャット用に最適化

されており、従来の補完タスクでも適切に動作します。GPT-3.5 Turbo は、チャット補完 API で使用できます。GPT-3.5 Turbo Instruct には、チャット入力候補 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。

- `gpt-35-turbo`
- `gpt-35-turbo-16k`
- `gpt-35-turbo-instruct`

`gpt-35-turbo` モデルでは最大 4096 個の入力トークンがサポートされ、`gpt-35-turbo-16k` モデルでは最大 16,384 個のトークンがサポートされます。`gpt-35-turbo-instruct` では、最大 4097 個の入力トークンがサポートされます。

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。

埋め込みモデル

📌 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、ドキュメントの埋め込みとドキュメント検索は同じバージョンのモデルを使用して行う必要があります。

以前の埋め込みモデルは、次の新しい代替モデルに統合されました。

`text-embedding-ada-002`

DALL-E (プレビュー)

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。

ささやき (プレビュー)

現在プレビュー段階のささやきモデルは、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。

モデルの概要テーブルとリージョンの可用性

① 重要

需要が高いため:

- 米国中南部は、新しいリソースとデプロイの作成が一時的に使用できなくなっています。

GPT-4 モデル

GPT-4 と GPT-4-32k は、すべての Azure OpenAI Service のお客様が利用できるようになりました。可用性はリージョンごとに異なります。使用中のリージョンで GPT-4 が表示されない場合は、時間を置いて再度確認してください。

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)
<code>gpt-4</code> ² (0314)	米国東部 ¹ 、フランス中部 ¹	該当なし	8,192	2021 年 9 月
<code>gpt-4-32k</code> ² (0314)	米国東部 ¹ 、フランス中部 ¹	該当なし	32,768	2021 年 9 月
<code>gpt-4</code> (0613)	オーストラリア東部 ¹ 、カナダ東部、米国東部 ¹ 、米国東部 2 ¹ 、フランス中部 ¹ 、東日本 ¹ 、スウェーデン中部、スイス北部、英国南部 ¹	該当なし	8,192	2021 年 9 月
<code>gpt-4-32k</code> (0613)	オーストラリア東部 ¹ 、カナダ東部、米国東部 ¹ 、米国東部 2 ¹ 、フランス中部 ¹ 、東日本 ¹ 、スウェーデン中部、スイス北部、英国南部 ¹	該当なし	32,768	2021 年 9 月

¹ 需要が高いため、リージョンでは可用性が制限されています

² gpt-4 と gpt-4-32k のバージョン 0314 は、2024 年 7 月 5 日より前に廃止されます。モデルのアップグレード動作については、「[モデルの更新](#)」を参照してください。

GPT-3.5 モデル

GPT-3.5 Turbo は、Chat Completions API と共に使用されます。GPT-3.5 Turbo (0301) も、Completions API と共に使用できます。GPT3.5 Turbo (0613) では、Chat Completions API のみがサポートされます。

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>gpt-35-turbo</code> ¹ (0301)	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
<code>gpt-35-turbo</code> (0613)	オーストラリア東部、カナダ東部、米国東部、米国東部 2、フランス中部、東日本、米国中北部、スウェーデン中部、スイス北部、英国南部	該当なし	4,096	2021 年 9 月
<code>gpt-35-turbo-16k</code> (0613)	オーストラリア東部、カナダ東部、米国東部、米国東部 2、フランス中部、東日本、米国中北部、スウェーデン中部、スイス北部、英国南部	該当なし	16,384	2021 年 9 月
<code>gpt-35-turbo-instruct</code> (0914)	米国東部、スウェーデン中部	該当なし	4,097	2021 年 9 月

¹ gpt-35-turbo のバージョン 0301 は、2024 年 7 月 5 日以降に廃止されます。モデルのアップグレード動作については、「[モデルの更新](#)」を参照してください。

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](#)² 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)	出力ディメンション
text-embedding-ada-002 (バージョン 2)	カナダ東部、米国東部、米国東部 2、フランス中部、東日本、米国中北部、米国中南部、スイス北部、英国南部、西ヨーロッパ	該当なし	8,191	2021 年 9 月	1536
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	1536

DALL-E モデル (プレビュー)

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)
dalle2	East US	該当なし	1000	該当なし

ささやきモデル (プレビュー)

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 (オーディオファイル サイズ)	トレーニング データ (最大)
ささやく	米国中北部、西ヨーロッパ	N/A	25 MB	該当なし

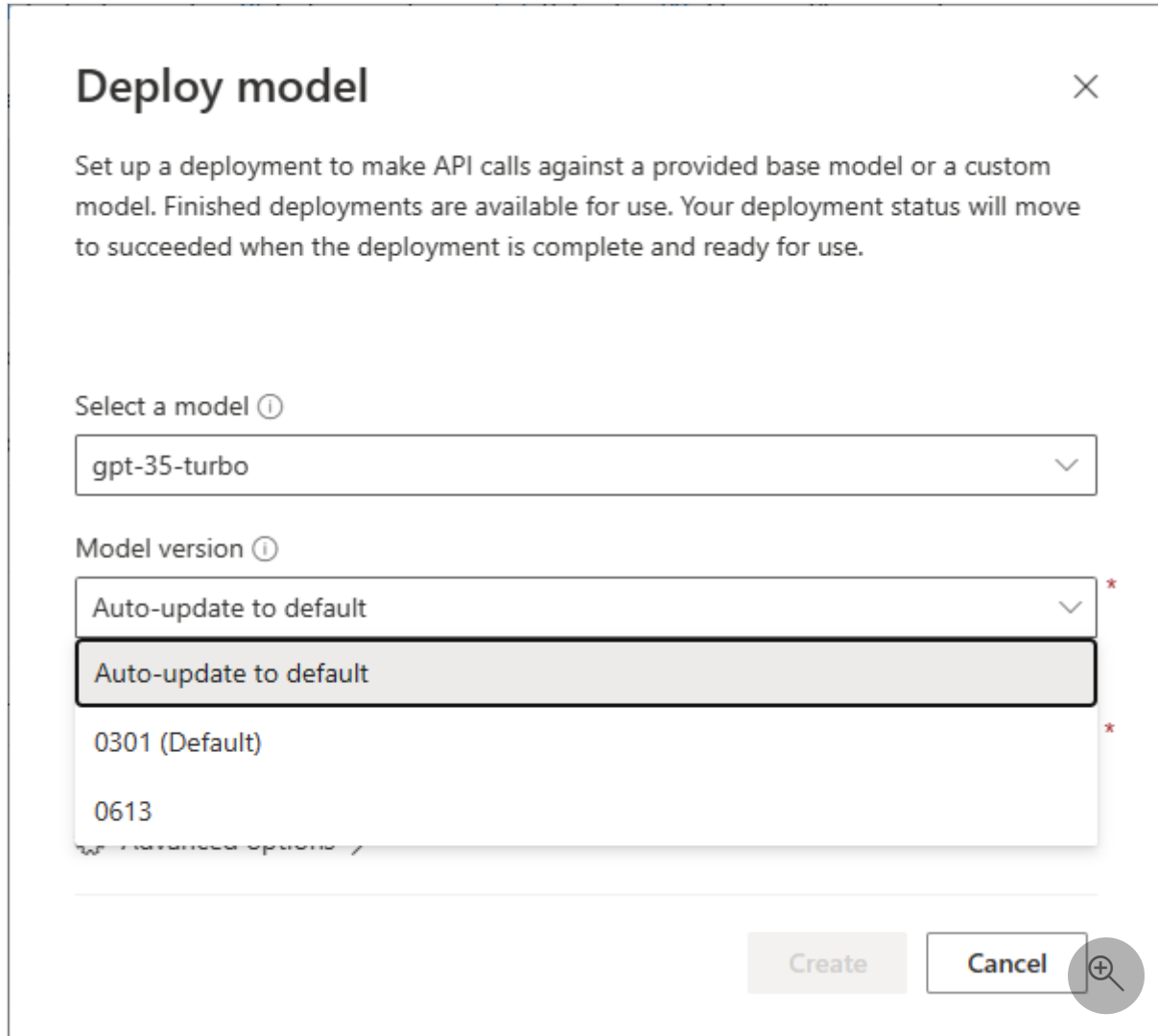
モデルの操作

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 [X]

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 ⓘ

gpt-35-turbo [v]

Model version ⓘ

Auto-update to default [v] *

Auto-update to default

0301 (Default) *

0613

Create Cancel [🔍]

既定値に自動更新

[既定値に自動更新] が選択されている場合、モデル デプロイは、既定のバージョンの変更が行われてから 2 週間以内に自動的に更新されます。

まだ推論モデルの早期テスト フェーズにある場合、**[既定値に自動更新]** が使用可能であれば常にこれを設定してモデルをデプロイすることをお勧めします。

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と、アップグレードする前にモデルのパフォーマンスが引き続きユースケースに対して一貫しているこ

とを初めにテストして検証できるように、モデルの更新を手動で制御することが必要になる場合があります。

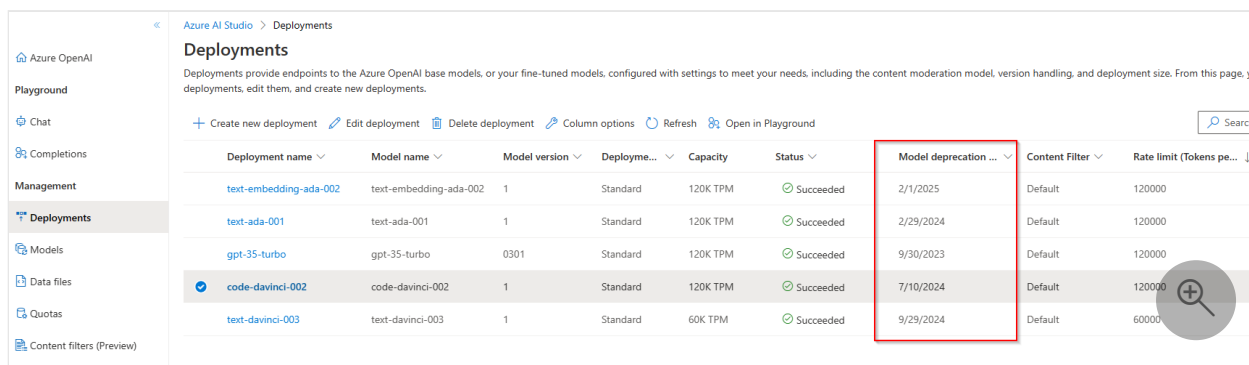
デプロイに特定のモデル バージョンを選択すると、自分で手動更新するか、モデルの提供終了日に達するまで、このバージョンは選択されたままになります。提供終了日に達すると、モデルは提供終了時に既定のバージョンに自動アップグレードされます。

GPT-35-Turbo 0301 および GPT-4 0314 の廃止

`gpt-35-turbo` (0301) モデルと両方 `gpt-4` (0314) モデルは、2024 年 7 月 5 日以降に廃止されます。提供終了時には、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。

Models

Azure OpenAI is powered by models with different capabilities and price points. Deploy one of the provided base models to try it out in [Playground](#) or train a custom model to your specific use case and data for better performance and more accurate results. [Learn more about the different types of base models](#)

Base models

Deploy Create a custom model Column options Refresh

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	9/29/2023 8:00 PM	Succeeded	No	-	9/29/2024 8:00 PM
text-embedding-ada-002	2	4/2/2023 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

モデル デプロイのアップグレード構成

モデル デプロイのアップグレード構成には、次の 3 つの異なるオプションがあります。これらは、REST API を使用して構成できます。

名前	説明
OnceNewDefaultVersionAvailable	新しいバージョンを既定として指定すると、モデル デプロイは、その指定変更が行われてから 2 週間以内に既定のバージョンに自動アップグレードされます。
OnceCurrentVersionExpired	廃止日になると、モデル デプロイは現在の既定のバージョンに自動アップグレードされます。
NoAutoUpgrade	モデル デプロイは自動アップグレードされません。廃止日になると、モデル デプロイは機能しなくなります。有効期限が切れていないモデル デプロイを指すように、そのデプロイを参照するコードを更新する必要があります。

特定のリソースのデプロイ アップグレード構成など、現在のモデル デプロイの設定に対してクエリを実行するには、[Deployments List](#)を使用します

HTTP
<p>GET</p> <p>https://management.azure.com/subscriptions/{subscriptionId}/resourceGroups/{resourceGroupName}/providers/Microsoft.CognitiveServices/accounts/{accountName}/deployments?api-version=2023-05-01</p>

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>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](#)

応答の例

JSON

```
{
  "id": "/subscriptions/{Subscription-GUID}/resourceGroups/{Resource-Group-Name}/providers/Microsoft.CognitiveServices/accounts/{Resource-Name}/deployments/text-davinci-003",
  "type": "Microsoft.CognitiveServices/accounts/deployments",
  "name": "text-davinci-003",
  "sku": {
    "name": "Standard",
    "capacity": 60
  },
  "properties": {
    "model": {
      "format": "OpenAI",
      "name": "text-davinci-003",
      "version": "1"
    },
    "versionUpgradeOption": "OnceNewDefaultVersionAvailable",
    "capabilities": {
      "completion": "true",
      "search": "true"
    },
    "raiPolicyName": "Microsoft.Default",
    "provisioningState": "Succeeded",
    "rateLimits": [
      {
        "key": "request",
        "renewalPeriod": 10,
        "count": 60
      }
    ]
  }
}
```

```
    },
    {
      "key": "token",
      "renewalPeriod": 60,
      "count": 60000
    }
  ]
}
```

次に、デプロイのアップグレード構成を変更する場合は、このリストから設定を取得して、以下で説明するようにモデル更新 REST API 呼び出しを作成できます。

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 リファレンス ドキュメント](#)をご覧ください。

パラメーター	Type	説明
versionUpgradeOption	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: OnceNewDefaultVersionAvailable OnceCurrentVersionExpired NoAutoUpgrade
capacity	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?api-version=2023-05-01 \
-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、[Azure portal](#) 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}\""
}
```

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 レガシ モデル

[アーティクル] • 2023/07/20

Azure OpenAI Service では、さまざまなユースケースに対応する多様なモデルを提供しています。次のモデルは、2023 年 7 月 6 日以降の新しいデプロイでは使用できません。2023 年 7 月 6 日より前に作成されたデプロイは、2024 年 7 月 5 日まで利用できます。2024 年 7 月 5 日の廃止前に、交換モデルに移行することをお勧めします。

GPT-3.5

影響を受ける GPT-3.5 モデルは次のとおりです。GPT-3.5 モデルの代わりとなるのは、そのモデルが使用可能になったときの GPT-3.5 Turbo Instruct です。

- `text-davinci-002`
- `text-davinci-003`
- `code-davinci-002`

GPT-3

影響を受ける GPT-3 モデルは次のとおりです。GPT-3 モデルの代わりとなるのは、そのモデルが使用可能になったときの GPT-3.5 Turbo Instruct です。

- `text-ada-001`
- `text-babbage-001`
- `text-curie-001`
- `text-davinci-001`
- `code-cushman-001`

埋め込みモデル

以下の埋め込みモデルは、2024 年 7 月 5 日に廃止されます。お客様は `text-embedding-ada-002` に移行する必要があります (バージョン 2)。

- [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。