Azure OpenAI Service とは

[アーティクル] • 2024/02/15

Azure OpenAI Service では、GPT-4、GPT-4 Turbo with Vision、GPT-3.5-Turbo、埋め込みモデル シリーズなど OpenAI の強力な言語モデルに、REST API でのアクセスを提供します。 また、新しい GPT-4 と GPT-3.5-Turbo モデルシリーズは一般提供になりました。 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、要約、画像の解釈、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

テーブルを展開する

機能	Azure OpenAI
使用できるモデル	**GPT-4 シリーズ** (GPT-4 Turbo with Vision **を含む**) **GPT-3.5-Turbo 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 (プレビュー)	`GPT-3.5-Turbo` (0613)
	`babbage-002`
	`davinci-002`

Price	こちらで入手可能 GPT-4 Turbo with Vision について詳しくは、特別価格情報を参照してください。
仮想ネットワークのサポートとプライベート 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。Microsoft Entra ID を使用
UI エクスペリエンス	**Azure portal** (アカウントとリソースの管理)、 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、Whisper、テキスト読み上げの各モデルを使用した高度な言語 AI を、Azure のセキュリティとエンタープライズの約束と共にお客様に提供します。 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。

重要な概念

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

テキスト 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。

画像トークン (GPT-4 Turbo with Vision)

入力画像のトークン コストは、画像のサイズと、各画像に使用される詳細設定 (低または高) の 2 つの主な要因によって異なります。 仕組みの概要を次に示します。

- **詳細**: **低解像度モード**
  - 低詳細度を使用すると、API ではより高速な応答を返し、高詳細度を必要としないユース ケースに使用する入力トークンを減らすことができます。
  - これらの画像のコストは、画像サイズに関係なくそれぞれ 85 トークンです。
  - **例**: **4096 x 8192 の画像 (低詳細度)**: コストは固定の 85 トークンです。これは低詳細度の画像であり、このモードではサイズがコストに影響しないためです。
- **詳細**: **高解像度モード**

- 高詳細度を使用すると、API では画像をより小さな正方形にトリミングすることでより詳細に表示できます。 それぞれの正方形では、テキストを生成するためにより多くのトークンを使用します。
- トークン コストは、一連のスケーリング手順によって計算されます。

    1. 画像は最初に、縦横比を維持しながら、2048 x 2048 の正方形内に収まるようにスケーリングされます。
    2. その後、最も短い辺が 768 ピクセル長になるように、画像がスケールダウンされます。
    3. 画像は 512 ピクセルの正方形タイルに分割され、これらのタイルの数 (部分的なタイルでは切り上げ) によって最終的なコストが決まります。 各タイルのコストは 170 トークンです。
    4. 合計コストには、さらに 85 トークンが追加されます。

- **例: 2048 x 4096 の画像 (高詳細度)**

    1. 2048 の正方形に収まるように、最初は 1024 x 2048 にサイズ変更されました。
    2. さらに 768 x 1536 にサイズ変更されました。
    3. カバーするには 6 つの 512px タイルが必要です。
    4. 合計コストは トークンです。 `170 × 6 + 85 = 1105`

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。

DALL-E モデル (一部プレビュー、モデルを参照) 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。

Whisper モデルは、音声からテキストへの文字起こしと翻訳を行うために使用できます。

現在プレビュー段階にある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。

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。

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。

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