Azure OpenAI Service とは

[アーティクル] • 2023/05/01

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 および ChatGPT (gpt-35-turbo) モデル シリーズがプレビューで利用可能になりました。 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、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 (プレビュー)** GPT-3 ベース シリーズ **新しい ChatGPT (gpt-35-turbo) (プレビュー)** 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* * 現在は利用できません。 ** 米国東部と西ヨーロッパでは、現在新規のお客様は微調整を利用できません。 米国ベースのトレーニングには、米国中南部をご利用ください
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
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**
リージョン別の提供状況	米国東部 米国中南部 西ヨーロッパ

機能	Azure OpenAI
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。

コンテキスト内学習

Azure OpenAI で使用されるモデルでは、生成の呼び出しで提供される自然言語での指示と例を使用して、要求されているタスクと必要なスキルを識別しています。 この方法を使用する場合、プロンプトの最初の部分には、自然言語の指示や、必要な特定のタスクの例が含まれます。 次に、最も可能性の高い次のテキスト部分を予測することで、モデルによってタスクが完了します。 この手法は、"コンテキスト内" 学習と呼ばれます。 これらのモデルは、この手順では再トレーニングされませんが、代わりにプロンプトに含めるコンテキストに基づいて予測を行います。

コンテキスト内学習には、主に 3 つのアプローチがあります。少数ショット、ワンショット、ゼロショットです。 これらの方法は、モデルに与えられるタスク固有のデータの量によって異なります。

少数ショット: この場合、ユーザーは、期待される応答形式と内容を示すいくつかの例を通話プロンプトに含めます。 次の例は、複数の例を提供する数個のプロンプトを示しています (最後の答えはモデルによって生成されます)。

```
    Convert the questions to a command:
    Q: Ask Constance if we need some bread.
    A: send-msg `find constance` Do we need some bread?
    Q: Send a message to Greg to figure out if things are ready for
Wednesday.
    A: send-msg `find greg` Is everything ready for Wednesday?
    Q: Ask Ilya if we're still having our meeting this evening.
    A: send-msg `find ilya` Are we still having a meeting this evening?
    Q: Contact the ski store and figure out if I can get my skis fixed
before I leave on Thursday.
    A: send-msg `find ski store` Would it be possible to get my skis fixed
before I leave on Thursday?
    Q: Thank Nicolas for lunch.
    A: send-msg `find nicolas` Thank you for lunch!
    Q: Tell Constance that I won't be home before 19:30 tonight – unmovable
meeting.
    A: send-msg `find constance` I won't be home before 19:30 tonight. I
have a meeting I can't move.
    Q: Tell John that I need to book an appointment at 10:30.
    A:
```

通常、1 つのプロンプトの最大入力長に収まる数に応じて、例の数は 0 から 100 の範囲です。 最大入力長は、使用する特定のモデルによって異なる場合があります。 少数ショット学習を使用すると、正確な予測に必要なタスク固有のデータの量を大幅に削減できます。 このアプローチは、通常、微調整されたモデルよりもパフォーマンスは正確ではありません。

ワンショット: この場合は、1 つの例のみが提供されることを除き、少数ショットの方法と同じです。

ゼロショット: この場合、モデルに例は提供されず、タスク要求のみが提供されます。

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。 これらのモデルは現在プレビュー段階です。 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。

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。

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。