

Azure OpenAI Service の新機能

[アーティクル] • 2023/09/15

2023 年 9 月

Whisper パブリックプレビュー

Azure OpenAI Service は、OpenAI の Whisper モデルによる音声テキスト変換 API をサポートするようになりました。指定した音声に基づいて AI で生成されたテキストを取得します。詳細については、[クイックスタート](#)を参照してください。

ⓘ 注意

Azure AI 音声は、バッチ文字起こし API を介した OpenAI の Whisper モデルもサポートしています。詳細については、「[バッチ文字起こしを作成する](#)」ガイドを参照してください。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。

2023 年 8 月

独自のデータに基づく Azure OpenAI (プレビュー) の更新

- データに Azure OpenAI を [Power Virtual Agents](#) にデプロイできるようになりました。
- [データ上の Azure OpenAI](#) でプライベート エンドポイントがサポートされるようになりました。
- [機密ドキュメントへのアクセス権をフィルター処理](#)する機能。
- [スケジュールに従ってインデックスを自動的に更新](#)。
- [ベクトル検索とセマンティック検索のオプション](#)。
- [デプロイされた Web アプリでチャット履歴を表示](#)

2023 年 7 月

関数呼び出しのサポート

- [Azure OpenAI](#) で関数呼び出しがサポートされるようになり、チャット入力候補 API で関数を操作できるようになりました。

入力配列の増加の埋め込み

- Azure OpenAI では、text-embedding-ada-002 バージョン 2 を使用した API 要求あたり [最大 16 の入力を含む配列がサポートされる](#) ようになりました。

新しいリージョン

- Azure OpenAI は、カナダ東部、米国東部 2、東日本、米國中北部リージョンでも使用でき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 6 月

独自のデータに基づく Azure OpenAI を使用する (プレビュー)

- [独自のデータに基づく Azure OpenAI](#) がプレビューで利用できるようになりました。これにより、GPT-35-Turbo や GPT-4 などの OpenAI モデルとチャットし、データに基づいて応答を受信できます。

gpt-35-turbo および gpt-4 モデルの新しいバージョン

- gpt-35-turbo (バージョン 0613)
- gpt-35-turbo-16k (バージョン 0613)
- gpt-4 (バージョン 0613)
- gpt-4-32k (バージョン 0613)

英国南部

- Azure OpenAI が米国南部リージョンで使用でき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用でき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、GPT-35-Turbo 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。
- **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 AI サービス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 AI サービス 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