Azure OpenAI Service の新機能

[アーティクル] • 2024/10/22

この記事では、Azure OpenAI の最新リリースと主要なドキュメント更新の概要を示します。

2024 年 10 月

新しいデータ ゾーン標準デプロイの種類

データ ゾーン標準デプロイは、Azure OpenAI のその他すべての種類のデプロイと同じ Azure OpenAI リソースで利用できます。ただし、Azure のグローバル インフラストラクチャを利用して、トラフィックを要求ごとに最適な可用性の Microsoft によって定義されたデータ ゾーン内のデータ センターに動的にルーティングできます。 データ ゾーン標準では、Azure の地理ベースのデプロイの種類よりも高い既定のクォータが提供されます。 データ ゾーン標準デプロイは、`gpt-4o-2024-08-06`、`gpt-4o-2024-05-13`、および `gpt-4o-mini-2024-07-18` のモデルでサポートされます。

詳細については、「展開の種類 ガイド」を参照してください。

グローバル バッチ GA

Azure OpenAI グローバル バッチが一般公開されました。

Azure OpenAI Batch API は、大規模で大量の処理タスクを効率的に処理するように設計されています。 個別のクォータ、24 時間のターゲット ターンアラウンド、グローバル スタンダードと比較した場合の 50% 低いコストで要求の非同期グループを処理します。 バッチ処理では、一度に 1 つの要求を送信するのではなく、1 つのファイル内で多数の要求を送信します。 グローバル バッチ要求には、オンライン ワークロードの中断を回避する個別のエンキュー トークン クォータがあります。

主なユース ケースは次のとおりです。

- **大規模なデータ処理**: 広範なデータセットを並列ですばやく分析します。
- **コンテンツ生成**: 製品の説明や記事など、大量のテキストを作成します。
- **ドキュメントの校閲と要約**: 長いドキュメントの校閲と要約を自動化します。
- **カスタマー サポートの自動化**: 多数の問い合わせを同時に処理して迅速な対応を実現します。

- **データの抽出と分析**: 膨大な量の非構造化データから情報を抽出して分析します。
- **自然言語処理 (NLP) タスク**: 大規模なデータセットに対して感情分析や翻訳などのタスクを実行します。
- **マーケティングとパーソナル化**: パーソナル化されたコンテンツとレコメンデーションを大規模に生成します。

詳細については、「グローバル バッチ デプロイの概要」で確認できます。

o1-preview と o1-mini モデルの制限付きアクセス

`o1-preview` と `o1-mini` モデルを API アクセスとモデル デプロイで使用できるようになりました。 **登録が必要であり、Microsoft の適格性条件に基づいてアクセスが許可されます。**

アクセスの要求: 制限付きアクセス モデルの申請

既に承認され、早期アクセス プレイグラウンドを通してモデルにアクセスできるお客様は、改めて申請する必要はなく、自動的に API アクセスを許可されます。 アクセスが許可されたら、モデルごとにデプロイを作成する必要があります。

API のサポート:

o1 シリーズ モデルのサポートが API バージョン `2024-09-01-preview` に追加されました。

`max_tokens` パラメーターは非推奨となり、新しい `max_completion_tokens` パラメーターに置き換えられました。 **o1 シリーズ** モデルは、`max_completion_tokens` パラメーターでのみ機能します。

利用可能なリージョン:

モデルは、米国東部 2 とスウェーデン中部の承認されたお客様の標準とグローバル標準のデプロイで利用できます。

音声とオーディオ用の新しい GPT-4o Realtime API (パブリック プレビュー)

Azure OpenAI GPT-4o オーディオは、GPT-4o モデル ファミリの一部であり、低遅延の "音声入力、音声出力" の会話操作をサポートします。 GPT-4o オーディオ `realtime` API は、リアルタイムで低遅延の会話操作を処理するように設計されており、カスタマー サポート エージェント、音声アシスタント、リアルタイム翻訳ツールなど、ユーザーとモデルの間のライブ対話を含むユース ケースに最適です。

`gpt-4o-realtime-preview` モデルは、米国東部 2 リージョンとスウェーデン中部リージョンのグローバル デプロイで使用できます。

詳しくは、GPT-4o リアルタイム オーディオのドキュメントを参照してください。

グローバル バッチ サポートの更新

グローバル バッチで GPT-4o (2024-08-06) がサポートされるようになりました。 詳しくは、グローバル バッチの概要ガイドを参照してください。

2024 年 9 月

Azure OpenAI Studio UX の更新

9 月 19 日より、Azure OpenAI Studio にアクセスすると、既定でレガシのスタジオ UI が表示されなくなります。 必要な場合は、今後数週間の間は UI のトップバーにある **[Switch to the old look]\(従来の外観に切り替える\)** トグルを使用して、以前の表示に戻すことができます。 レガシの Studio UI に戻す場合は、その理由をフィードバック フォームにご記入いただけると幸いです。 新しいエクスペリエンスを改善するために、マイクロソフトではこのフィードバックを積極的にモニターしています。

GPT-4o 2024-08-06 プロビジョニング済みデプロイ

GPT-4o 2024-08-06 は、米国東部、米国東部 2、米国中北部、スウェーデン中部のプロビジョニング済みデプロイで利用できるようになりました。 また、グローバルなプロビジョニング済みデプロイでも利用できます。

モデルの提供状況に関する最新情報については、モデルのページを参照してください。

新しいグローバルなプロビジョニング済みデプロイの種類

Global デプロイは、非グローバル デプロイ タイプと同じ Azure OpenAI リソースで利用できます。ただし、Azure のグローバル インフラストラクチャを利用して、トラフィックを要求ごとに最適な可用性のデータ センターに動的にルーティングできます。 グローバルなプロビジョニング済みデプロイでは、Azure グローバル インフラストラクチャを使用して、予測可能な高いスループットを実現するための予約済みのモデル処理容量が提供されます。 グローバルなプロビジョニング済みデプロイは、`gpt-4o-2024-08-06` と `gpt-4o-mini-2024-07-18` のモデルでサポートされます。

詳細については、「展開の種類 ガイド」を参照してください。

制限付きアクセスで利用可能になった新しい o1-preview と o1-mini モデル

Azure OpenAI の `o1-preview` と `o1-mini` モデルは、集中と能力を高めて推論と問題解決のタスクに取り組むために設計されています。 これらのモデルは、ユーザーの要求の処理と理解により多くの時間を費やし、これまでのイテレーションと比較して、科学、コーディング、数学などの分野で非常に強力になっています。

o1 シリーズの主な機能

- 複雑なコード生成: 開発者をサポートするための、アルゴリズム生成と、高度なコーディング タスクの処理の機能。
- 高度な問題解決: 包括的なブレーンストーミング セッションや多面的な課題への対処に最適。
- 複雑なドキュメント比較: 契約、ケース ファイル、法的ドキュメントなどを分析して微妙な違いを特定するのに最適。
- 命令のフォローとワークフロー管理: 短いコンテキストを必要とするワークフローの管理に特に効果的。

モデルのバリエーション

- `o1-preview`: `o1-preview` は、 `o1` シリーズのモデルより高い機能を備えています。
- `o1-mini`: `o1-mini` は、 `o1` シリーズのモデルより高速で安価です。

モデルのバージョン: `2024-09-12`

アクセスの要求: 制限付きアクセス モデルの申請

制限事項

`o1` シリーズのモデルは現在プレビュー段階であり、最新の GPT-4o モデルで利用できる画像理解や構造化出力など、他のモデルで使用できる一部の機能は含まれていません。 多くのタスクでは、一般提供されている GPT-4o モデルの方がまだ適している場合があります。

安全性

OpenAI では、モデルが安全でない要求を拒否するのに役立つ新しい手法など、 `o1` モデルをいっそう安全にする手段が組み込まれています。 これらの進歩により、 `o1` シリーズは最も堅牢なモデルの一部になっています。

可用性

`o1-preview` と `o1-mini` には、米国東部 2 リージョンで、AI Studio の早期アクセス プレイグラウンドを通じて、制限付きでアクセスできます。 `o1` モデルのデータ処理は、それらを利用できる場所とは異なるリージョンで行われる可能性があります。

早期アクセス プレイグラウンドで `o1-preview` と `o1-mini` モデルを試すには、**登録が必要であり、Microsoft の適格性基準に基づいてアクセスが許可されます**。

アクセスの要求: 制限付きアクセス モデルの申請

アクセスが許可されたら、次のようにする必要があります。

1. https://ai.azure.com/resources に移動し、`eastus2` リージョンでリソースを選びます。 このリージョンに Azure OpenAI リソースがない場合は、それを作成する必要があります。
2. `eastus2` の Azure OpenAI リソースを選んだ後、左上のパネルの [**プレイグラウンド**] で [**早期アクセス プレイグラウンド (プレビュー)**] を選びます。

2024 年 8 月

GPT-4o 2024-08-06 の構造化出力

- すべての米国リージョンとスウェーデン中部での標準およびグローバル デプロイで利用できます。
- このモデルでは、構造化出力のサポートが追加されます。

GPT-4o mini プロビジョニング済みデプロイ

GPT-4o mini は、カナダ東部、米国東部、米国東部 2、米国中北部、スウェーデン中部でプロビジョニングされたデプロイで利用できるようになりました。

モデルの提供状況に関する最新情報については、モデルのページを参照してください。

GPT-4o ファインチューニング (パブリック プレビュー)

GPT-4o ファインチューニングは現在、米国中北部とスウェーデン中部においてパブリック プレビューで Azure OpenAI で利用できます。

詳細については、こちらのブログ記事を参照してください。

新しいプレビュー API のリリース

API バージョン `2024-07-01-preview` は、最新のデータプレーン作成および推論 API のリリースです。 API バージョン `2024-05-01-preview` と置き換えられ、次のサポートが追加されます。

- Batch API のサポートが追加されました
- ベクトル ストア チャンク戦略パラメーター
- ファイル検索ツールが出力するべき `max_num_results`。

詳細については、「リファレンス ドキュメント」を参照してください

GPT-4o mini が利用できるリージョン

- GPT-4o mini は、米国東部およびスウェーデン中部リージョン内のスタンダードおよびグローバル スタンダード デプロイで利用できます。
- GPT-4o mini は、米国東部、スウェーデン中部、および米国西部リージョン内のグローバル バッチ デプロイで利用できます。

評価ガイド

- モデル評価の概要 に関する新しいブログ記事。 このガイドをモデルのアップグレードと廃止プロセスの一環として使用することをお勧めします。

早期アクセス プレイグラウンドで利用可能な最新の GPT-4o モデル (プレビュー)

2024 年 8 月 6 日、OpenAI は主力製品である GPT-4o モデルの最新バージョンであるバージョン `2024-08-06` を発表しました 。 GPT-4o `2024-08-06` は以前のバージョンのすべての機能に加えて以下を備えています。

- 複雑で構造化された出力をサポートする強化された機能。
- 最大出力トークン数が 4,096 から 16,384 に増加しました。

Azure のお客様は、新しい AI Studio の早期アクセス プレイグラウンド (プレビュー) 内で GPT-4o `2024-08-06` を今すぐテストできます。

これまでの早期アクセス プレイグラウンドとは異なり、AI Studio の早期アクセス プレイグラウンド (プレビュー) では、特定のリージョン内にリソースを用意する必要はありません。

> **注意**
>
> 早期アクセス プレイグラウンド (プレビュー) を通じて実行されるプロンプトと補完は、任意の Azure OpenAI リージョンで処理される可能性があり、現在は、Azure サブスクリプションごとに 1 分間に 10 個の要求という制限の対象になります。 この制限は将来変更される可能性があります。
>
> Azure OpenAI Service の不正使用監視は、変更が承認されている場合でも、すべての早期アクセス プレイグラウンド ユーザーに対して有効になっており、既定のコンテンツ フィルターも有効で変更することができません。

GPT-4o `2024-08-06` をテストするには、こちらのリンクを使用して Azure AI 早期アクセス プレイグラウンド (プレビュー) にサインインしてください。

グローバル バッチ デプロイが使用可能になりました

Azure OpenAI Batch API は、大規模で大量の処理タスクを効率的に処理するように設計されています。 個別のクォータ、24 時間のターゲット ターンアラウンド、グローバル スタンダードと比較した場合の 50% 低いコストで要求の非同期グループを処理します。 バッチ処理では、一度に 1 つの要求を送信するのではなく、1 つのファイル内で多数の要求を送信します。 グローバル バッチ要求には、オンライン ワークロードの中断を回避する個別のエンキュー トークン クォータがあります。

主なユース ケースは次のとおりです。

- **大規模なデータ処理**: 広範なデータセットを並列ですばやく分析します。
- **コンテンツ生成**: 製品の説明や記事など、大量のテキストを作成します。
- **ドキュメントの校閲と要約**: 長いドキュメントの校閲と要約を自動化します。
- **カスタマー サポートの自動化**: 多数の問い合わせを同時に処理して迅速な対応を実現します。
- **データの抽出と分析**: 膨大な量の非構造化データから情報を抽出して分析します。
- **自然言語処理 (NLP) タスク**: 大規模なデータセットに対して感情分析や翻訳などのタスクを実行します。
- **マーケティングとパーソナル化**: パーソナル化されたコンテンツとレコメンデーションを大規模に生成します。

詳細については、「グローバル バッチ デプロイの概要」で確認できます。

2024 年 7 月

GPT-4o mini でファインチューニングが使用可能になりました

GPT-4o mini ファインチューニングは現在、スウェーデン中部と米国中北部においてパブリック プレビューで利用できます。

Assistants ファイル検索ツールが課金されるようになりました

Assistants のファイル検索ツールの使用に追加料金が発生するようになりました。 詳細については、価格に関するページを参照してください。

GPT-4o mini モデルのデプロイが利用可能に

GPT-4o mini は、2024 年 7 月 18 日 に初めて発表された最新の Azure OpenAI モデルです:

"*GPT-4o mini* は、お客様が驚くべき速度と低コストで素晴らしいアプリケーションを提供することを可能にします。 *GPT-4o mini* は、*Massive Multitask Language Understanding (MMLU)* の測定で *82%* のスコアを付けるなどスコアが *70%* である *GPT-3.5 Turbo* よりもかなりスマートであり、*60%* 以上低コストです。*1* このモデルは、拡張された *128K* コンテキスト ウィンドウを提供し、*GPT-4o* の強化された多言語機能を統合し、世界中の言語に対してより高い品質をもたらします。"

このモデルは現在、米国東部リージョン内の標準デプロイとグローバル標準デプロイの両方で利用できます。

モデル クォータの詳細についてはクォータと制限に関するページを参照し、モデルの可用性に関する最新情報についてはモデルに関するページを参照してください。

新しい責任ある AI の既定のコンテンツ フィルタリング ポリシー

新しい既定のコンテンツ フィルタリング ポリシー `DefaultV2` は、GPT モデル シリーズ (テキスト) に対して最新の安全性とセキュリティに関する軽減策を提供します。これには、次のものが含まれます。

- ユーザー プロンプト (フィルター) に対する脱獄攻撃のプロンプト シールド
- モデル補完時のテキスト (フィルター) について保護されたマテリアルの検出

- モデル補完時のコード (注釈) について保護されたマテリアルの検出

既存のリソースとデプロイに対するコンテンツ フィルターに変更はありません (既定またはカスタムのコンテンツ フィルター構成は変更されないままです) が、新しいリソースと GPT の展開は、新しいコンテンツ フィルター ポリシー `DefaultV2` を自動的に継承します。 お客様は、安全性の既定値を切り替えて、カスタム コンテンツ フィルタリング構成を作成することを選択できます。

詳細については、既定の安全ポリシーに関するドキュメントを参照してください。

新規の一般提供 API リリース

API バージョン `2024-06-01` は、最新の一般提供データ プレーン推論 API リリースです。 API バージョン `2024-02-01` と置き換えられ、次のサポートが追加されます。

- 埋め込み `encoding_format` および `dimensions` パラメーター。
- チャット入力候補 `logprobs` および `top_logprobs` パラメーター。

詳細については、data プレーン推論リファレンスのドキュメントを参照してください。

gpt-4o のグローバル標準デプロイで利用可能なリージョンの拡大

GPT-4o は、次のグローバル標準デプロイで使用できるようになりました :

- australiaeast
- brazilsouth
- canadaeast
- eastus
- eastus2
- francecentral git
- germanywestcentral
- japaneast
- koreacentral
- northcentralus
- norwayeast
- polandcentral
- southafricanorth
- southcentralus
- southindia
- swedencentral

- switzerlandnorth
- uksouth
- westeurope
- westus
- westus3

グローバル標準クォータの詳細については、quota と制限に関するページを参照してください。

2024 年 6 月

提供終了日の更新

- `gpt-35-turbo` 0301 の提供終了日を、2024 年 10 月 1 日以降に更新しました。
- `gpt-35-turbo` と `gpt-35-turbo-16k` 0613 の提供終了日を、2024 年 10 月 1 日に更新しました。
- `gpt-4` と `gpt-4-32k` 0314 の非推奨となる日を 2024 年 10 月 1 日に、提供終了日を 2025 年 6 月 6 日に更新しました。

モデルの非推奨と提供終了に関する最新情報については、「モデル提供終了ガイド」を参照してください。

微調整のためのトークン ベース課金

- Azure OpenAI の微調整課金は、トレーニングの総経過時間ではなく、トレーニング ファイルのトークンの数に基づくようになりました。 これにより、一部のトレーニング実行のコストが大幅に削減され、微調整コストの見積もりがはるかに簡単になります。 詳細については、公式発表 を参照してください。

新しいリージョンで GPT-4o がリリースされました

- GPT-4o は、次のリージョンでも使用できるようになりました。
  - 標準のリージョン デプロイ用のスウェーデン中部。
  - プロビジョニングされたデプロイ用のオーストラリア東部、カナダ東部、東日本、韓国中部、スウェーデン中部、スイス北部、および米国西部 3。

モデルの提供状況に関する最新情報については、モデルのページを参照してください。

Assistants 用のカスタマー マネージド キー (CMK) のサポート

Assistants のスレッドとファイルで、次のリージョンの CMK がサポートされるようになりました。

- 米国西部 3

2024 年 5 月

GPT-4o プロビジョニング済みデプロイ

`gpt-4o` バージョン: `2024-05-13` は、標準デプロイとプロビジョニングされたデプロイの両方で使用できます。 プロビジョニング済みと標準のモデル デプロイでは、テキストと画像/ビジョンの両方の推論要求を受け入れます。 リージョン別のモデルの提供状況については、プロビジョニングされたデプロイのモデル マトリックスを参照してください。

Assistants v2 (プレビュー)

Assistants API の更新が一般公開されました。 次の更新が含まれています。

- ファイル検索ツールとベクトル ストレージ
- トークン使用の管理のために最大完了トークンと最大プロンプト トークンをサポート。
- 指定したツールを使用するようアシスタントに強制する `tool_choice` パラメーター。 アシスタント ロールでメッセージを作成して、スレッドのカスタム会話履歴を作成できるようになりました。
- `temperature`、 `top_p`、 `response_format` のパラメーターのサポート。
- ストリーミングとポーリングのサポート。 Python SDK のヘルパー関数を使用して、実行とストリーム応答を作成できます。 ポーリング不要でオブジェクトの状態の更新を共有できるポーリング SDK ヘルパーも追加されました。
- Azure OpenAI Studio を使用した Logic Apps と関数呼び出しの実験。 Logic Apps に実装されている REST API を関数としてインポートすると、Studio はユーザー プロンプトに基づいて、関数を (Logic Apps ワークフローとして) 自動的に呼び出します。
- AutoGen by Microsoft Research では、幅広いアプリケーションで大規模言語モデル (LLM) ワークフローを手軽に構築できるマルチエージェント会話フレームワークが提供されます。 Azure OpenAI アシスタントは、アシスタントを AutoGen ベースのマルチエージェント ワークフローにシームレスに追加できる、新しい実験的なエージェントである `GPTAssistantAgent` を介して AutoGen に統合されました。 これにより、タスクやドメインに特化した複数の Azure OpenAI アシスタントを協働させて、複雑なタスクに取り組めます。

- 微調整された `gpt-3.5-turbo-0125` モデルが次のリージョンでサポートされます。
  - 米国東部 2
  - スウェーデン中部
- 次のリージョンでリージョン サポートが展開されました。
  - 東日本
  - 英国南部
  - 米国西部
  - 米国西部 3
  - ノルウェー東部

詳細については、アシスタントに関するブログ記事を参照してください。

GPT-4o モデルの一般提供 (GA)

GPT-4o ("o" は "オムニ" の意) は、2024 年 5 月 13 日に発表された OpenAI の最新モデルです。

- GPT-4o はテキストと画像を 1 つのモデルに統合し、複数のデータ型を同時に処理できるようにします。 このマルチモーダル アプローチにより、人間とコンピューターの対話における精度と応答性が向上します。
- GPT-4o は、英語以外の言語とビジョン タスクで優れたパフォーマンスを提供しながら、英語のテキストとコーディング タスクにおいて GPT-4 Turbo に匹敵し、AI 機能の新しいベンチマークを設定します。

リージョン別のモデルの提供状況については、モデルのページを参照してください。

グローバル標準の展開の種類 (プレビュー)

グローバル展開は、非グローバル オファーと同じ Azure OpenAI リソースで利用できます。ただし、Azure のグローバル インフラストラクチャを利用して、トラフィックを要求ごとに最適な可用性のデータ センターに動的にルーティングできます。 グローバル標準では、新しいモデルに対して最大の既定クォータが提供され、複数のリソース間で負荷を分散する必要はありません。

詳細については、「展開の種類 ガイド」を参照してください。

微調整に関する更新

- 現在、GPT-4 の微調整はパブリック プレビューで利用できます。
- `2024-05-01-preview` API リリースの一環として、シード、イベント、完全な検証統計、チェックポイントのサポートが追加されました。

DALL-E および GPT-4 Turbo Vision GA の構成可能なコンテンツ フィルター

DALL-E 2 および 3、GPT-4 Turbo with Vision GA (`turbo-2024-04-09`)、GPT-4o のデプロイ用のカスタム コンテンツ フィルターを作成できます。 コンテンツのフィルター処理

すべての Azure OpenAI カスタマーが利用できる非同期フィルター

ストリーミング シナリオでの待機時間を改善するための、フィルターの非同期的な実行を、すべての Azure OpenAI カスタマーが利用できるようになりました。 コンテンツのフィルター処理

プロンプト シールド

プロンプト シールドは、Azure OpenAI モデルを利用するアプリケーションを、直接攻撃 (ジェイルブレイク) と間接攻撃の 2 種類の攻撃から保護します。 間接攻撃 (間接プロンプト攻撃またはクロスドメイン プロンプトインジェクション攻撃とも呼ばれます) は、生成 AI モデルを搭載したシステムに対する攻撃の一種で、アプリケーション開発者やユーザーが直接作成していない情報をアプリケーションが処理するときに発生する可能性があります。 コンテンツのフィルター処理

2024-05-01-preview API リリース

- 詳細については、「API バージョン ライフサイクル」を参照してください。

GPT-4 Turbo モデルの一般提供 (GA)

GPT-4 Turbo の最新 GA リリースは次のとおりです。

- `gpt-4` **バージョン** `turbo-2024-04-09`

これは、次のプレビュー モデルに代わるものです。

- `gpt-4` **バージョン** `1106-Preview`
- `gpt-4` **バージョン** `0125-Preview`
- `gpt-4` **バージョン** `vision-preview`

OpenAI と Azure OpenAI GPT-4 Turbo GA モデルの違い

- OpenAI の最新の `0409` ターボ モデル バージョンでは、すべての推論要求に対して JSON モードと関数呼び出しがサポートされています。
- Azure OpenAI の最新の `turbo-2024-04-09` バージョンでは、現在、画像 (ビジョン) 入力による推論要求を行う場合、JSON モードと関数呼び出しの使用はサポートされていません。 テキスト ベース入力の要求 (`image_url` とインライン イメージがない要求) では、JSON モードと関数呼び出しがサポートされています。

gpt-4 vision-preview との違い

- Azure AI 固有の Vision 拡張機能と GPT-4 Turbo with Vision の統合は、`gpt-4` **バージョン**: `turbo-2024-04-09` ではサポートされません。 これには、光学式文字認識 (OCR)、オブジェクト グラウンディング、ビデオ プロンプト、画像を含むデータの処理の改善が含まれます。

> **ⓘ 重要**
>
> 光学式文字認識 (OCR)、オブジェクト グラウンディング、ビデオ プロンプトなどのビジョン拡張機能のプレビュー機能は廃止され、`gpt-4` バージョン: `vision-preview` が `turbo-2024-04-09` にアップグレードされると使用できなくなります。 現在これらのプレビュー機能のいずれかに依存している場合、このモデルの自動アップグレードは破壊的変更になります。

GPT-4 Turbo のプロビジョニングされたマネージド可用性

- `gpt-4` **バージョン** `turbo-2024-04-09` は、標準デプロイとプロビジョニングされたデプロイの両方で使用できます。 現在、このモデルのプロビジョニングされたバージョンでは、イメージ/ビジョン推論要求は**サポートされていません**。 このモデルのプロビジョニングされたデプロイでは、テキスト入力のみ受け入れます。 標準のモデル デプロイでは、テキストと画像/ビジョンの両方の推論要求を受け入れます。

GPT-4 Turbo with Vision GA のデプロイ

Studio UI から GA モデルをデプロイするには、`GPT-4` を選択し、ドロップダウン メニューから `turbo-2024-04-09` バージョンを選択します。 `gpt-4-turbo-2024-04-09` モデルの既定のクォータは、GPT-4-Turbo の現在のクォータと同じになります。 リージョンのクォータ制限を参照してください。

2024 年 4 月

米国東部 2 とスイス西部という 2 つの新しいリージョンで微調整がサポートされるようになりました

次の機能をサポートする微調整が利用できるようになりました。

米国東部 2

- `gpt-35-turbo` (0613)
- `gpt-35-turbo` (1106)
- `gpt-35-turbo` (0125)

スイス西部

- `babbage-002`
- `davinci-002`
- `gpt-35-turbo` (0613)
- `gpt-35-turbo` (1106)
- `gpt-35-turbo` (0125)

各リージョンでのモデルの提供状況とファインチューニングのサポートに関する最新情報は、モデルのページをご確認ください。

マルチターン チャット トレーニング例

微調整でマルチターン チャット トレーニング例がサポートされるようになりました。

GPT-4 (0125) は Azure OpenAI On Your Data で使用できます

Azure OpenAI on Your Data を使用して対応リージョンで GPT-4 (0125) モデルを使用できるようになりました。

2024 年 3 月

Azure OpenAI Studio でのリスクと安全性の監視

Azure OpenAI Studio では、コンテンツ フィルター構成を使用するデプロイごとにリスクと安全性のダッシュボードが提供されるようになりました。 これを使用して、フィルター処理アクティビティの結果を確認します。 その後、フィルター構成を調整して、より適切にビジネス ニーズに対応し、責任ある AI 原則を満たすことができます。

リスクと安全性の監視の使用

Azure OpenAI On Your Data の更新

- Azure OpenAI On Your Data で使用する Elasticsearch ベクトル データベースに接続できるようになりました。
- データ インジェスト中にチャンク サイズ パラメーターを使用して、インデックス内の特定のデータ チャンクのトークンの最大数を設定できます。

2024-02-01 一般提供 (GA) API がリリースされました

これは最新の GA API リリースであり、以前の `2023-05-15` GA API リリースに代わるものです。 このリリースでは、Whisper、DALLE-3、微調整、On Your Data など、最新の Azure OpenAI GA 機能のサポートが追加されています。

アシスタント、テキスト読み上げ (TTS)、一部の "On Your Data" データソースなどのプレビュー段階の機能には、プレビュー API バージョンが必要です。 詳しくは、API バージョンのライフサイクル ガイドをご覧ください。

Whisper 一般提供 (GA)

Whisper 音声テキスト変換モデルは、REST と Python の両方で GA になりました。 クライアント ライブラリ SDK は現在、パブリック プレビュー段階にあります。

クイックスタートに従って、Whisper をお試しください。

DALL-E 3 一般提供 (GA)

DALL-E 3 画像生成モデルは、REST と Python の両方で GA になりました。 クライアント ライブラリ SDK は現在、パブリック プレビュー段階にあります。

クイックスタートに従って、DALL-E 3 をお試しください。

DALL-E 3 の新しいリージョン サポート

`SwedenCentral` に加え、 `East US` と `AustraliaEast` の Azure リージョンでも、Azure OpenAI リソースで DALL-E 3 にアクセスできるようになりました。

モデルの非推奨と提供終了

Azure OpenAI Service でのモデルの非推奨と提供終了 を追跡するページが追加されました。 このページでは、現在使用可能、非推奨、提供終了のモデルに関する情報を提供します。

2024-03-01-preview API がリリースされました

`2024-03-01-preview` は、 `2024-02-15-preview` と同じ機能をすべて備えており、埋め込み用の以下の 2 つの新しいパラメータが追加されています。

- `encoding_format` では、埋め込みを生成する形式を `float` または `base64` で指定できます。 既定値は、 `float` です。
- `dimensions` では、出力される埋め込みの数を設定できます。 このパラメータは、新しい第 3 世代埋め込みモデル (`text-embedding-3-large` および `text-embedding-3-small`) でのみサポートされています。 通常、埋め込みが大きくなると、コンピューティング、メモリ、ストレージの観点からコストが高くなります。 ディメンション数を調整できるので、全体的なコストとパフォーマンスをより詳細に制御できます。 `dimensions` パラメーターは、OpenAI 1.x Python ライブラリのすべてのバージョンではサポートされていません。このパラメーターを利用するには、最新バージョンの `pip install openai --upgrade` にアップグレードすることをお勧めします。

現在、プレビュー API バージョンを使って最新の機能を利用している場合は、API バージョンのライフサイクルに関する記事を参照して、現在お使いの API バージョンのサポート期間を確認することをお勧めします。

GPT-4-1106-Preview アップグレード プランの更新

2024 年 3 月 8 日に予定されていた `gpt-4` 1106-Preview から `gpt-4` 0125-Preview へのデプロイ アップグレードは行われなくなりました。 "自動更新を既定にする" と "期限切れになったときにアップグレードする" に設定された `gpt-4` バージョン 1106-Preview と 0125-Preview のデプロイは、安定バージョンのモデルがリリースされた後にアップグレードが開始されます。

アップグレード プロセスの詳細については、 モデルに関するページを参照してください。

2024 年 2 月

GPT-3.5-turbo-0125 モデルが利用可能になりました

このモデルではさまざまな機能強化が組み込まれました。たとえば、要求された形式での応答精度の向上、英語以外の言語の関数呼び出しに対してテキスト エンコードの問題が発生していたバグの修正などです。

リージョン別のモデルの提供状況とアップグレードについては、モデルのページを参照してください。

利用可能な第 3 世代埋め込みモデル

- `text-embedding-3-large`
- `text-embedding-3-small`

OpenAI の報告によると、テストでは、大規模と小規模の第 3 世代埋め込みモデルのいずれも、MIRACL ベンチマークで多言語検索の平均パフォーマンスが向上しており、さらに MTEB ベンチマークで、第 2 世代の text-embedding-ada-002 モデルよりも優れた英語タスクのパフォーマンスを維持しています。

リージョン別のモデルの提供状況とアップグレードについては、モデルのページを参照してください。

GPT-3.5 Turbo のクォータ統合

GPT-3.5-Turbo モデル (16k を含む) の異なるバージョン間の移行を簡単にするため、すべての GPT-3.5-Turbo クォータを 1 つのクォータ値に統合する作業が行われます。

- クォータの引き上げが承認されたお客様は、以前の引き上げを反映した統合された合計クォータを保有します。
- モデル バージョン全体の現在の合計使用量が既定値より少ないお客様は、統合された新しい合計クォータを既定で取得します。

GPT-4-0125-preview モデルが利用可能

`gpt-4` モデルのバージョン `0125-preview` が、米国東部、米国中北部、米国中南部の各リージョンの Azure OpenAI Service で利用できるようになりました。 `gpt-4` バージョン `1106-preview` のデプロイを使用しているお客様は、今後数週間以内に `0125-preview` に自動的にアップグレードされます。

リージョン別のモデルの提供状況とアップグレードについては、モデルのページを参照してください。

Assistants API パブリック プレビュー

Azure OpenAI では、OpenAI の GPT を利用できる API がサポートされるようになりました。 Azure OpenAI Assistants (プレビュー) を使用すると、カスタム命令やコード インタープリターなどの高度なツール、およびカスタム関数を使用して、自分のニーズに合わせて調整された AI アシスタントを作成できます。 詳細については、以下をご覧ください。

- クイックスタート
- 概念
- 詳細な Python の使用方法
- コード インタープリター
- 関数呼び出し
- Assistants モデルと利用可能なリージョン
- Assistants Python および REST リファレンス
- Assistants サンプル

OpenAI テキスト読み上げ音声パブリック プレビュー

Azure OpenAI Service では、OpenAI の音声を使用したテキスト読み上げ API がサポートされるようになりました。 指定したテキストから AI で生成された音声を取得します。 詳細については、「概要ガイド」を参照し、「クイックスタート」を試してください。

> ⓘ **注意**
>
> Azure AI 音声でも、OpenAI のテキスト読み上げ音声がサポートされています。 詳細については、「**Azure OpenAI Service または Azure AI 音声を介した OpenAI テキスト読み上げ音声**」ガイドを参照してください。

新しいファインチューニング機能とモデルのサポート

- 継続的なファインチューニング
- ファインチューニングと関数呼び出し
- gpt-35-turbo 1106 のサポート

独自のデータに基づく Azure OpenAI の新しいリージョン サポート

これで、次の Azure リージョンで Azure OpenAI On Your Data を使用できるようになりました。

- 南アフリカ北部

Azure OpenAI On Your Data の一般提供

- Azure OpenAI On Your Data が一般公開されました。

2023 年 12 月

Azure OpenAI On Your Data

- ストレージ アカウント、Azure OpenAI リソース、Azure AI 検索サービス リソースのセキュリティ サポートなど、Azure OpenAI On Your Data に対する VPN とプライベート エンドポイントの完全なサポート。
- 仮想ネットワークとプライベート エンドポイントを使用してデータを保護することで、Azure OpenAI On Your Data を安全に使用するための新しい記事。

GPT-4 Turbo with Vision が利用可能

Azure OpenAI サービスの GPT-4 Turbo with Vision はパブリック プレビュー中です。 GPT-4 Turbo with Vision は、OpenAI によって開発された大規模なマルチモーダル モデル (LMM) であり、画像を分析し、それらに関する質問に対してテキストでの応答を提供できます。 自然言語処理とビジュアル解釈の両方が組み込まれています。 拡張モードでは、Azure AI Vision 機能を使用して、画像から追加の分析情報を生成できます。

- Azure OpenAI Playground を使用して、ノー コード エクスペリエンスで GPT-4 Turbo with Vision の機能を探索します。 詳細については、クイック スタート ガイドを参照してください。
- GPT-4 Turbo with Vision を使った Vision の機能強化は、Azure OpenAI Playground で利用できるようになりました。また、光学式文字認識、オブジェクト グラウンディング、"データの追加" の画像サポート、ビデオ プロンプトのサポートが含まれています。
- REST API を使用してチャット API を直接呼び出します。
- 現在、利用可能なリージョンは`SwitzerlandNorth`、`SwedenCentral`、`WestUS`、および`AustraliaEast` に制限されています

- GPT-4 Turbo with Vision の既知の制限事項と、その他のよく寄せられる質問の詳細をご覧ください。

2023 年 11 月

Azure OpenAI On Your Data での新しいデータ ソース サポート

- Azure Cosmos DB for MongoDB 仮想コアと URL/Web アドレスをデータ ソースとして使って、サポートされている Azure OpenAI モデルでのデータの取り込みとチャットを行うことができるようになりました。

GPT-4 Turbo プレビューと GPT-3.5-Turbo-1106 のリリース

両モデルとも OpenAI の最新リリースであり、指示実行、JSON モード、再現可能な出力、並列関数呼び出しが改善されています。

- **GPT-4 Turbo プレビュー**には、最大 128,000 トークンのコンテキスト ウィンドウがあり、4,096 の出力トークンを生成できます。 その最新のトレーニング データには 2023 年 4 月までの情報が含まれています。 このモデルはプレビューであり、運用環境では使わないことをお勧めします。 このプレビュー モデルのデプロイはすべて、安定版リリースが利用可能になると自動的にインプレースで更新されます。

- **GPT-3.5 Turbo 1106** には、最大 16,385 トークンのコンテキスト ウィンドウがあり、4,096 の出力トークンを生成できます。

リージョン別のモデルの提供状況については、モデルのページを参照してください。

モデルには、リージョンごとに独自の一意のクォータ割り当てがあります。

DALL-E 3 パブリック プレビュー

DALL-E 3 は、OpenAI の最新の画像生成モデルです。 画像の品質が向上し、より複雑なシーンが表示され、画像内のテキストをレンダリングするときのパフォーマンスが向上しています。 また、縦横比のオプションが増えています。 DALL-E 3 は、OpenAI Studio と REST API を通じて使用できます。 OpenAI リソースは、`SwedenCentral` Azure リージョンに存在する必要があります。

DALL-E 3 には、画像を強化し、バイアスを減らし、自然な変動を増やすためのプロンプト書き換えが組み込まれています。

クイックスタートに従って、DALL-E 3 をお試しください。

責任ある AI

- **拡張された顧客の構成可能性**: Azure OpenAI のすべてのお客様は、ヘイト、暴力、性的、自傷のカテゴリに対して、すべての重大度レベル (低、中、高) を構成できるようになりました。これには重大度の高いコンテンツのみをフィルター処理することが含まれます。 コンテンツ フィルターを構成する

- **すべての DALL-E モデルのコンテンツ資格情報**: すべての DALL-E モデルから生成された AI 画像には、AI が生成したコンテンツであることを示すデジタル資格情報が含まれるようになりました。 画像アセットを表示するアプリケーションでは、AI で生成された画像に資格情報を表示するために、オープンソースの Content Authenticity Initiative SDK が利用されています。 Azure OpenAI のコンテンツ資格情報

- **新しい RAI モデル**
  - **脱獄リスク検出**: 脱獄攻撃は、システム メッセージに設定されたルールを回避または中断するようにトレーニングされた動作を生成 AI モデルに示させる目的で設計されたユーザー プロンプトです。 脱獄リスク検出モデルは省略可能 (既定ではオフ) で、注釈とフィルター モデルで使用できます。 これはユーザー プロンプトで実行されます。
  - **保護済み素材テキスト**: 保護済み素材テキストは、大規模言語モデルによって出力される可能性のある既知のテキスト コンテンツ (曲の歌詞、記事、レシピ、一部の Web コンテンツなど) を記述するものです。 保護済み素材テキスト モデルは省略可能 (既定ではオフ) で、注釈とフィルター モデルで使用できます。 これは LLM の完了時に実行されます。
  - **保護済み素材コード**: 保護済み素材コードは、ソース リポジトリを適切に引用することなく大規模言語モデルによって出力される可能性のある、パブリック リポジトリからの一連のソース コードと一致するソース コードを記述するものです。 保護済み素材コード モデルは省略可能 (既定ではオフ) で、注釈とフィルター モデルで使用できます。 これは LLM の完了時に実行されます。

  コンテンツ フィルターを構成する

- **ブロックリスト**: お客様は、自分のフィルターにカスタム ブロックリストを作成して、プロンプトと入力候補のコンテンツ フィルターの動作をすばやくカスタマイズできるようになりました。 カスタム ブロックリストを使用すると、そのフィルターでパターンのカスタマイズされたリスト (特定の用語や正規表現パターンな

ど) に対してアクションを実行できます。 カスタム ブロックリストに加えて、Microsoft の不適切な表現のブロックリスト (英語) も提供しています。 ブロックリストを使用する

2023 年 10 月

新しい微調整モデル (プレビュー)

- `gpt-35-turbo-0613` が、微調整に使用できるようになりました。
- `babbage-002` と `davinci-002` が、微調整に使用できるようになりました。 これらのモデルは、以前微調整に使用できたレガシ ada、babbage、curie、davinci ベース モデルに代わるものです。
- 微調整は、特定のリージョンでのみ利用できます。 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。
- 微調整されたモデルには、通常のモデルとは異なるクォータ制限があります。
- チュートリアル: GPT-3.5-Turbo の微調整

Azure OpenAI On Your Data

- 取得したドキュメントの数と厳密度を決定する、新しいカスタム パラメーター。
  - 厳密度の設定では、クエリに関連するドキュメントの分類に使用するしきい値を設定します。
  - 取得したドキュメントの設定では、応答の生成に使用されるデータ インデックスの上位スコアのドキュメントの数を指定します。
- Azure OpenAI Studio でデータ インジェスト/アップロードの状態を確認できます。
- BLOB コンテナーでのプライベート エンドポイントと VPN のサポート。

2023 年 9 月

GPT-4

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。 お客様は、GPT-4 と GPT-4-32k を使用するための待機リストに申し込む必要がなくなりました (制限付きアクセス登録要件はすべての Azure OpenAI モデルに

引き続き適用されます)。 提供状況はリージョンによって異なる場合があります。 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。

GPT-3.5 Turbo Instruct

Azure OpenAI Service で GPT-3.5 Turbo Instruct モデルがサポートされるようになりました。 このモデルのパフォーマンスは `text-davinci-003` と同等であり、Completions API で使用できます。 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。

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。

新しいリージョン

- Azure OpenAI は、スウェーデン中部およびスイス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。

リージョンのクォータ制限の引き上げ

- 特定のモデルとリージョンについて、既定のクォータ制限の最大値に引き上げられます。 これらのモデルとリージョンにワークロードを移行すると、より大きい 1 分あたりのトークン数 (TPM) を利用できます。

2023 年 8 月

独自のデータに基づく Azure OpenAI (プレビュー) の更新

- Azure OpenAI On Your Data を Power Virtual Agents にデプロイできるようになりました。
- Azure OpenAI On Your Data 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、米国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)

- Azure OpenAI On Your Data がプレビューでお使いいただけ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 と 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。

現在、`2023-03-15-preview` API をお使いの場合は、GA の `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` をお使いの場合、この API は GA のままですが、最新のチャット入力候補機能は含まれません。

> **ⓘ 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。

- **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) とも呼ばれます) 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、Azure 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。

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 | Microsoft Q&A でヘルプを表示する