AI-050

補足資料：Module06

Azure OpenAI Service で独自のデータを使用する

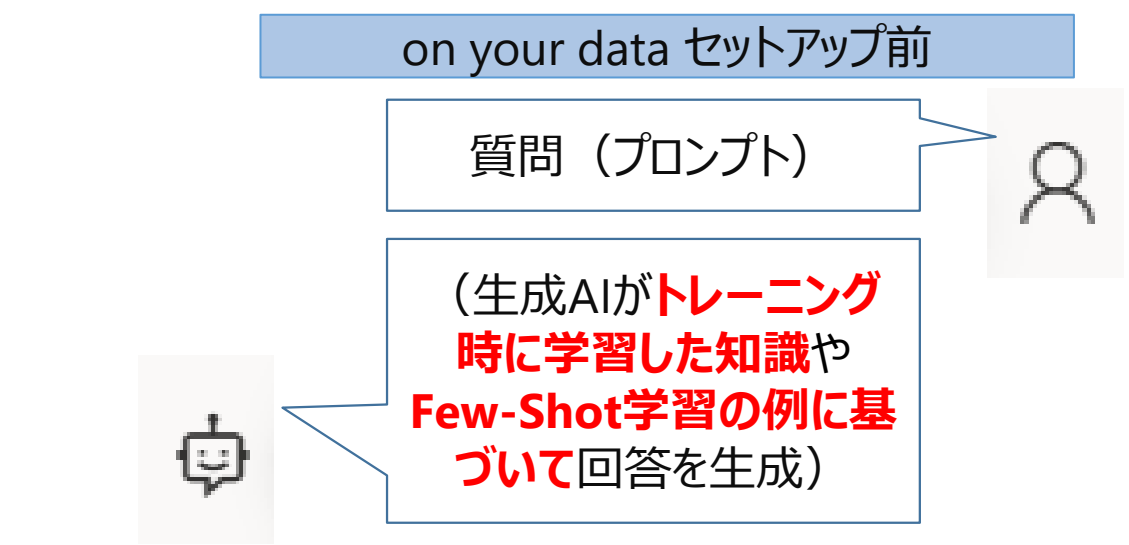
- 独自のデータの使用方法を理解する
- 独自のデータ ソースを追加する
- 独自のデータを使用してモデルとチャットする

Azure OpenAI Service で独自のデータを使用する - Training | Microsoft Learn

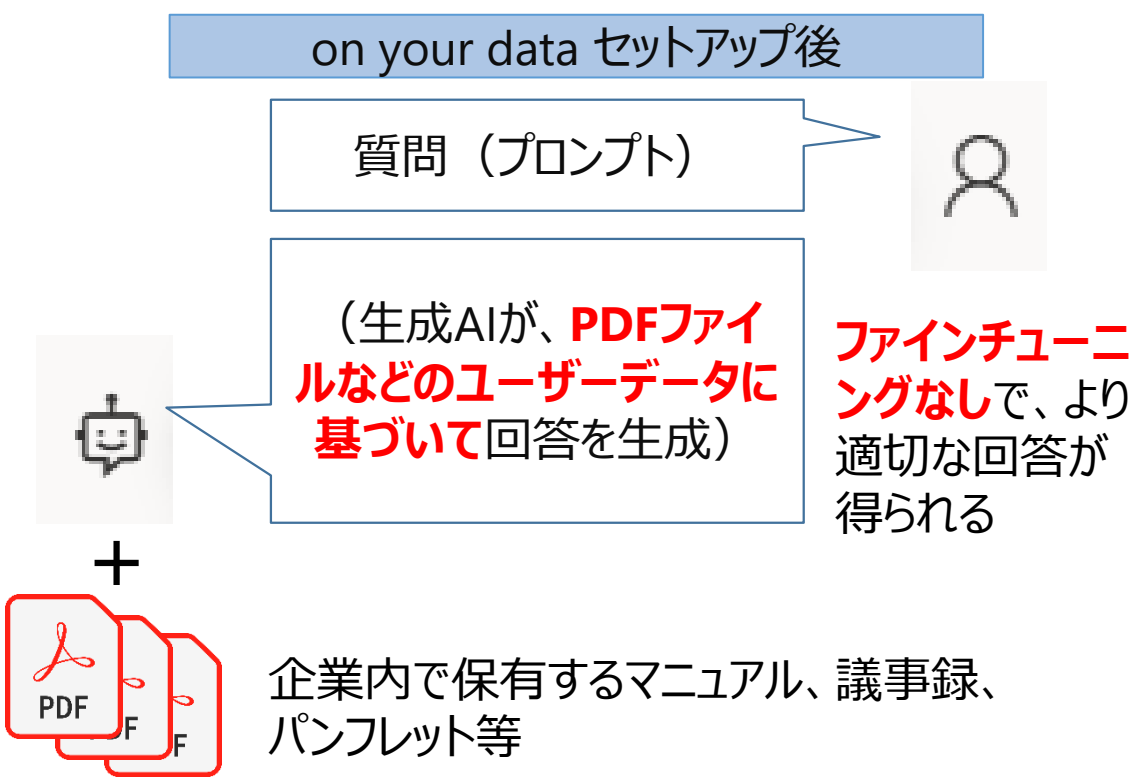
独自データ（On Your Data）とは？

独自データ（On Your Data）とは？

- 生成モデル（gpt-35-turbo / gpt-4等）が、**ユーザーが与えたデータに基づいて回答を生成できるようにする**仕組み



一般的なことには答えられるが、企業内の情報などに基づく回答はできない。
（いままでは**ファインチューニング**により、モデルに必要なデータを追加で学習させる必要があった）



[Public Preview: Azure OpenAI Service On Your Data | Azure の更新情報 | Microsoft Azure]

独自データ（On Your Data）とは？

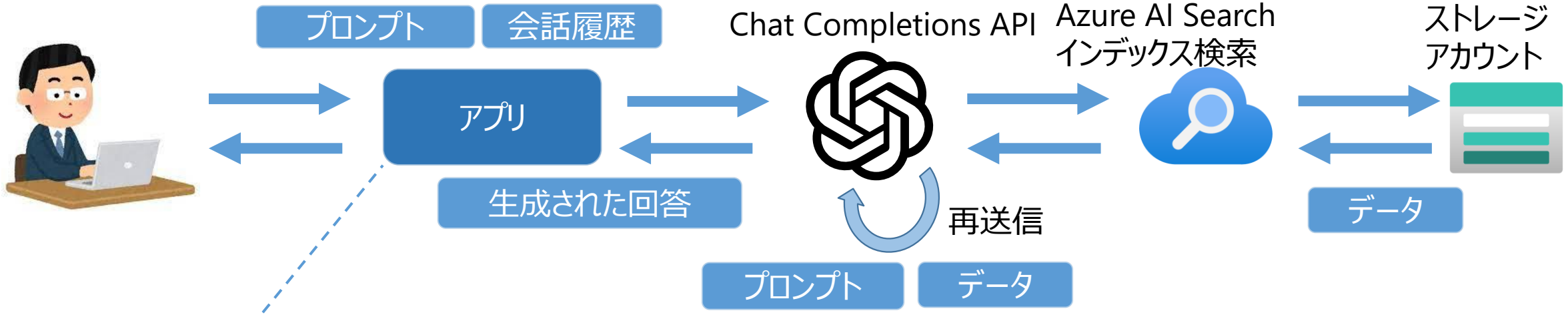
- データ形式として.txt / .md / .html / Word / PowerPoint / PDFに対応
- データは**ストレージアカウント**に記録される※
  - ※データソースで「ファイルのアップロード」を指定した場合。
  - 既存のAzure AI Search、Azure Blob Storage、Azure Cosmos DB for MongoDB vCoreなどをデータソースとして利用することも可能
- データの追加・変更・削除も可能（インデックスの手動 / 自動更新が可能）
- 内部的には**Azure AI Search**を使用して高速な検索を実現している
- **データの追加はAzure OpenAI Studioから行う**ことが推奨されている
  - 投入したデータが適切な大きさのかたまり（チャンク）に分割されて、インデックスが作成されるため。

独自データ（On Your Data）の特徴

- モデルの**ファインチューニング**をせずに、**ユーザーが提供する独自のデータ**に基づくコンテンツ生成が可能となる
  - ファインチューニングの手間とコストが削減できる
    - ファインチューニングを行うには、トレーニング用のデータの準備とトレーニングが必要。時間と手間とトレーニングコストがかかる
    - ファインチューニングによって作られたモデル（カスタムモデル）の利用には追加の運用コストがかかる
- Azure OpenAI Studioから簡単に設定できる
- gpt-35-turbo / gpt-4 （Chat Completions API）で使用できる
  - 使い方は普通のChat Completions API（Chatプレイグラウンド）とほぼ同じ

独自データ（On Your Data）の動作

- データはストレージアカウントに格納され、Azure AI Searchを使用して、データからインデックス（索引）が作られる
- インデックスを使用して、プロンプトに関係のあるデータが高速に検索され、取得される
- データの検索・取得は Chat Completions API 側で自動的に実行される
- ユーザーの**プロンプト**と**会話履歴**から、適切なデータソースが選択され、データが取得される。



Azure OpenAI StudioのChatプレイグラウンド
または独自のアプリ / システム

- 取得した**データ**が追加された**プロンプト**が生成モデルに**再送信**される

独自データ（On Your Data）のコスト

- on your data自体の利用にはコストはかからない
  - 使用されたトークン、一緒に使ったAzureリソースについては料金が発生

- 内部的に、**ストレージアカウント**と**Azure AI Search**を使用するので、その部分のコストは追加で発生する。

- 内部的に、ユーザーのプロンプトに、取得したデータが追加され、モデルに送信されるため、それによる入力・出力トークン数の変動があり、そのためのコストの増減がありうる。

独自データ（On Your Data）の考慮事項

- 独自データを使用したRAG利用時のトークン
  - モデルの呼び出しには、システムメッセージ、ユーザープロンプト、会話履歴、取得した検索ドキュメント、内部プロンプト、モデルの応答トークンが含まれる
  - モデルによって、システムメッセージ、モデルからの応答のトークン数は異なる
- プロンプトの長さや会話履歴の両方を制限することを考慮する
  - **Chain of thought （CoT）**が効果的（モジュール3で紹介）
- 独自データの形式や内容が、モデルからの応答の品質に影響する
  - 文字の一覧が理想（それ以外はノイズになりえる）
  - ファイルにテーブルや列、箇条書きなどの特別な書式設定がある場合は、データ準備スクリプトを使用してデータを準備する

Azure OpenAI Service で独自のデータを使用する - Azure OpenAI | Microsoft Learn

独自データ（On Your Data）の使用例

F1ドライバーで2023年のチャンピオンは誰ですか？また2022年のポイント獲得数と2023年の差が大きいドライバーは？

...

申し訳ありませんが、2023年のF1ドライバーズチャンピオンについての情報は、私の最終情報更新日時点で未確定です。また、2022年と2023年のポイント獲得数に関しても、現時点での詳細な比較データを提供することはできません。

F1の成績やドライバーの情報は、F1公式サイトや主要なスポーツニュースサイトで定期的に更新されています。シーズン進行中には、各レース後にポイントスタンディングが更新されるので、そちらをチェックすることをお勧めします。

独自データ（On Your Data）の使用例

参考：on your data

- Azureの独自機能（**OpenAI社側では「on your data」という機能は提供していない**）。ユーザーのデータに基づく回答が可能になる。RAGアーキテクチャをすばやく実装できる。

- Assistants API: もともとOpenAI社が2023年11月に提供を開始したAPI。開発者が独自のアプリケーション内でエージェントのようなエクスペリエンスを構築するためのAPI。

- Azure OpenAIが追加した機能「on your data」、OpenAIが「Assistants API」をRAGとして提供。

Fine-tuning vs RAG

比較項目	ファインチューニング（Fine-tuning）	Retrieval Augmented Generation（RAG）
目的	◎ 特定のタスクに最適化されたカスタムモデルの作成	◎ 外部データソースを使用してモデルの精度と適応性を向上
コスト	△ トレーニングとカスタムモデルのホスティングに高コストがかかる	◎通常、初期コストが低く、適応可能なオプション
データの必要性	△ 大量の高品質なデータが必要	◎ 必要なデータ量は少なく、外部データソースを利用
適用範囲	◎ 特定のスタイルやトーンで出力をカスタマイズする場合に最適	◎最新情報や長文の情報を処理する場合に有効
技術的要件	△ 高度な専門知識が必要	◎ 比較的シンプルで迅速に実装可能
モデルの柔軟性	△ 固定されたカスタムモデルになるため柔軟性が低い	◎外部データを随時更新可能で柔軟性が高い
実装の容易さ	△ 専門的な知識と時間が必要	◎ 比較的簡単にセットアップ可能
パフォーマンス評価	◎ 明確なベースラインと定量的な評価が可能	△定量的な評価が難しい場合がある
まとめ	特定のタスクやフォーマットにモデルを適応させるために既存のモデルを再トレーニングします。 高度な技術的専門知識が必要で、大量の高品質なデータが不可欠です。 コストが高く、柔軟性に欠けることがあります。	外部データソースをリアルタイムで取得し、プロンプトに組み込むことでモデルの性能を向上させます。 比較的低コストで実装が容易で、適応性が高いです。 最新情報や長文情報の処理に特に有効ですが、定量的な評価が難しい場合もあります。

知識チェック 第1問

Azure OpenAI on your data を使用すると、開発者は何を行うことができますか?

a. 独自の AI チャット モデルを作成する

b. 承認されたサブスクリプションなしで Azure OpenAI にアクセスする

c. ユーザーが与えた任意のデータに基づいて回答を生成できる

知識チェック 第1問

Azure OpenAI on your data を使用すると、開発者は何を行うことができますか?

a. 独自の AI チャット モデルを作成する

不正解。独自のモデル作成には「ファインチューニング」などを利用する。

b. 承認されたサブスクリプションなしで Azure OpenAI にアクセスする

不正解。Azure OpenAI Serviceの利用前に申請を行い、承認を得る必要がある。

c. ユーザーが与えた任意のデータに基づいて回答を生成できる

正解。PDF、テキストファイル、Word、PowerPoint等のデータに基づいて回答を生成できる。

知識チェック 第2問

Azure OpenAI on your data を使用する場合、データを追加するための推奨される方法は何ですか?

a. Azure OpenAI on your data で利用できる任意のデータ ソース オプションを使用する。

b. Azure OpenAI Studio を使用して検索リソースとインデックスを作成する。

c. Azure OpenAI Studio を使用せずにストレージ アカウント内のファイルに接続する。

知識チェック 第2問

Azure OpenAI on your data を使用する場合、データを追加するための推奨される方法は何ですか?

a. Azure OpenAI on your data で利用できる任意のデータ ソース オプションを使用する。

不正解。入力データが分解されない。

b. Azure OpenAI Studio を使用して検索リソースとインデックスを作成する。

正解。Azure OpenAI Studioを使用すると、入力データが適切な単位（最大1024トークンの「チャンク」）に分解されてから、各チャンクごとにインデックスが作成される。

c. Azure OpenAI Studio を使用せずにストレージ アカウント内のファイルに接続する。

不正解。入力データが分解されない。

知識チェック 第3問

Azure OpenAI on your data を使用する場合に推奨されるプロンプト エンジニアリング手法は何ですか?

a. 短いプロンプトと、必要最小限の会話履歴を含める。

b. できるだけ多くの会話履歴を含める。

c. 1 つの長いプロンプトを使用して、必要なすべての情報を提供する。

知識チェック 第3問

Azure OpenAI on your data を使用する場合に推奨されるプロンプト エンジニアリング手法は何ですか?

a. 短いプロンプトと、必要最小限の会話履歴を含める。

正解。プロンプトと会話履歴に基づいてデータが取得される

b. できるだけ多くの会話履歴を含める。

効率的ではない

c. 1 つの長いプロンプトを使用して、必要なすべての情報を提供する。

効率的ではない