

AI-050

補足資料：Module06

- 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 AI | Azure OpenAI Studio

Azure OpenAI

プレイグラウンド

チャット

入力候補

DALL-E (プレビュー)

アシスタント (プレビュー)

管理

デプロイ

モデル

データ ファイル

クォータ

Azure OpenAI Studio > チャット プレイグラウンド

チャット プレイグラウンド

設定

プロンプト

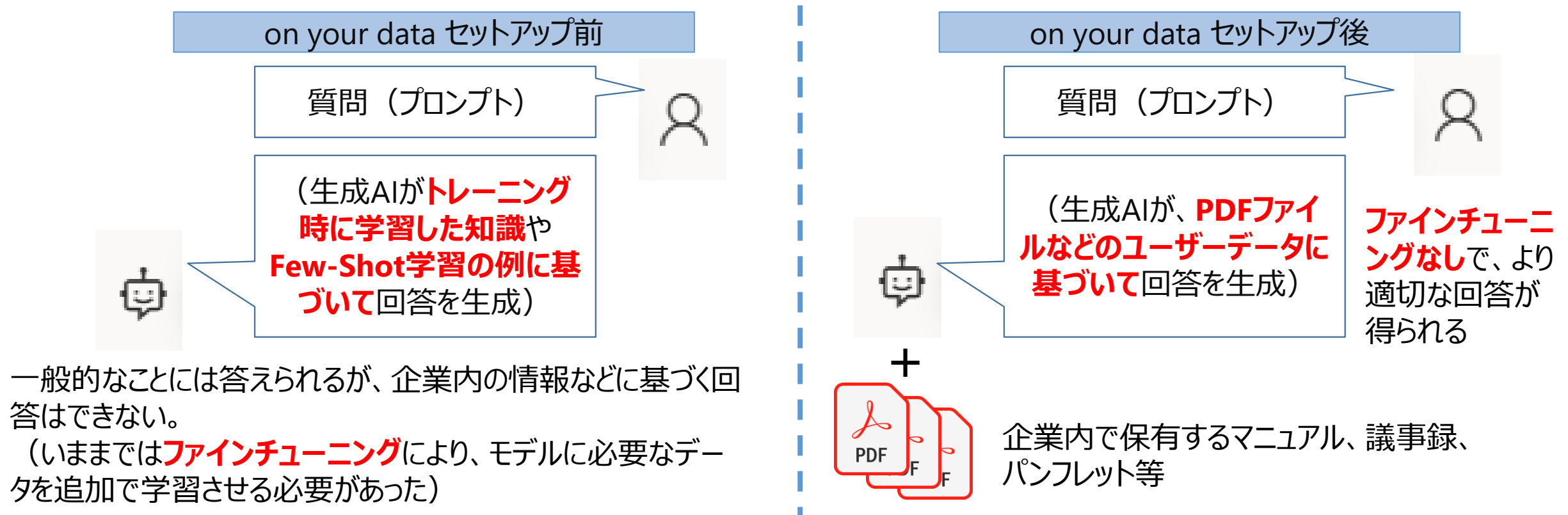
データの追加 (プレビュー)

自分のデータについて質問します。データは、指定したデータソースに保存されたままになります。[データの保護方法について説明します。](#)

+ データソースの追加

チャット
コード

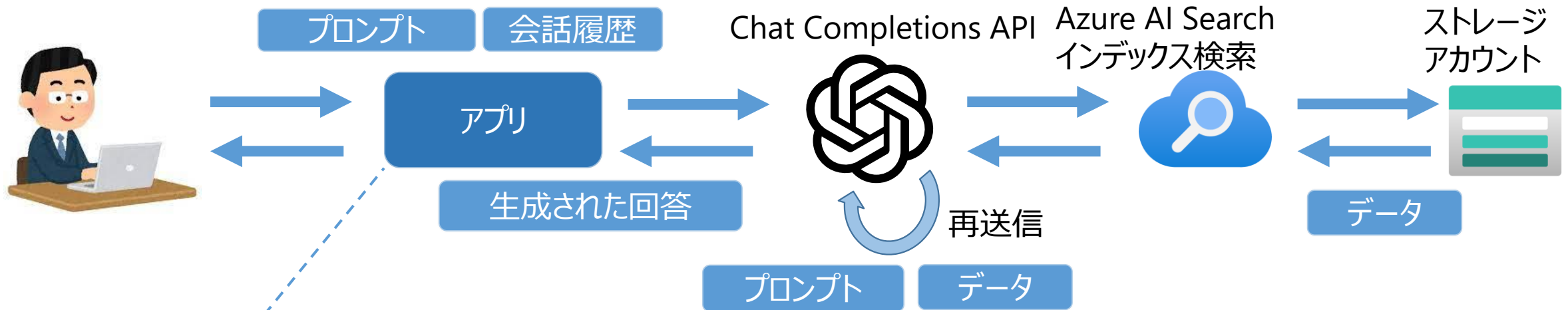
- 生成モデル（gpt-35-turbo / gpt-4等）が、**ユーザーが与えたデータに基づいて回答を生成できるようにする**仕組み



- 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。

- 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）で利用できる
 - 使い方は普通のChat Completions API（Chatプレイグラウンド）とほぼ同じ

- 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

- 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

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。

データの追加



● データソース

○ データ管理

○ レビューして終了

データソースを選択または追加する

データソースは、生成された結果をデータに基にするために使用されます。基になるデータの構築元となるソースとして、既存のデータソースを選択するか、Azure Blob Storage、データベース、検索、URL またはローカルファイルを使用して新しいデータ接続を作成します。

[Azure AI のデータプライバシーとセキュリティの詳細をご覧ください。](#)

データソースを選択する *

URL/web address (preview)

リソース

Azure AI Search リソースを選択する ⓘ *

aisearch20240220

[新しい Azure AI Search リソースを作成する](#)

Azure Blob ストレージ リソースの選択 ⓘ *

labnaodemo

インデックス名

インデックス名を入力してください ⓘ *

f1

データソースの場所

URL/Web アドレスを入力してください ⓘ *

https:// news.yahoo.co.jp/articles/e...

☐ ベクトル検索をこの検索リソースに追加します。 ⓘ

- 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。

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

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。

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。

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。

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。

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