

Azure OpenAIをもちいた LLMアプリの企画から 本番構築までの道のり

2023.6.27

株式会社サイダス 取締役CTO

吉田真吾

吉田真吾



■ バックグラウンド

証券システム基盤開発

- 基盤システム開発、Oracleチューニング

エバンジェリスト・SA

- 講演113回/年
- AWS設計・構築・移行

■ 現在のしごと

(株) サイダス 取締役CTO

- タレントマネジメントSaaS 開発・運用

(株) セクションナイン 代表取締役

- AWS、DevOps、サーバーレス開発

■ 実績等

- AWSウルトラクイズ 初代チャンピオン (2012年)
- AWS Samurai 2014 / 2016
- AWS Serverless Hero

好きなMSサービス

→ Azure OpenAI

→ Cognitive Search

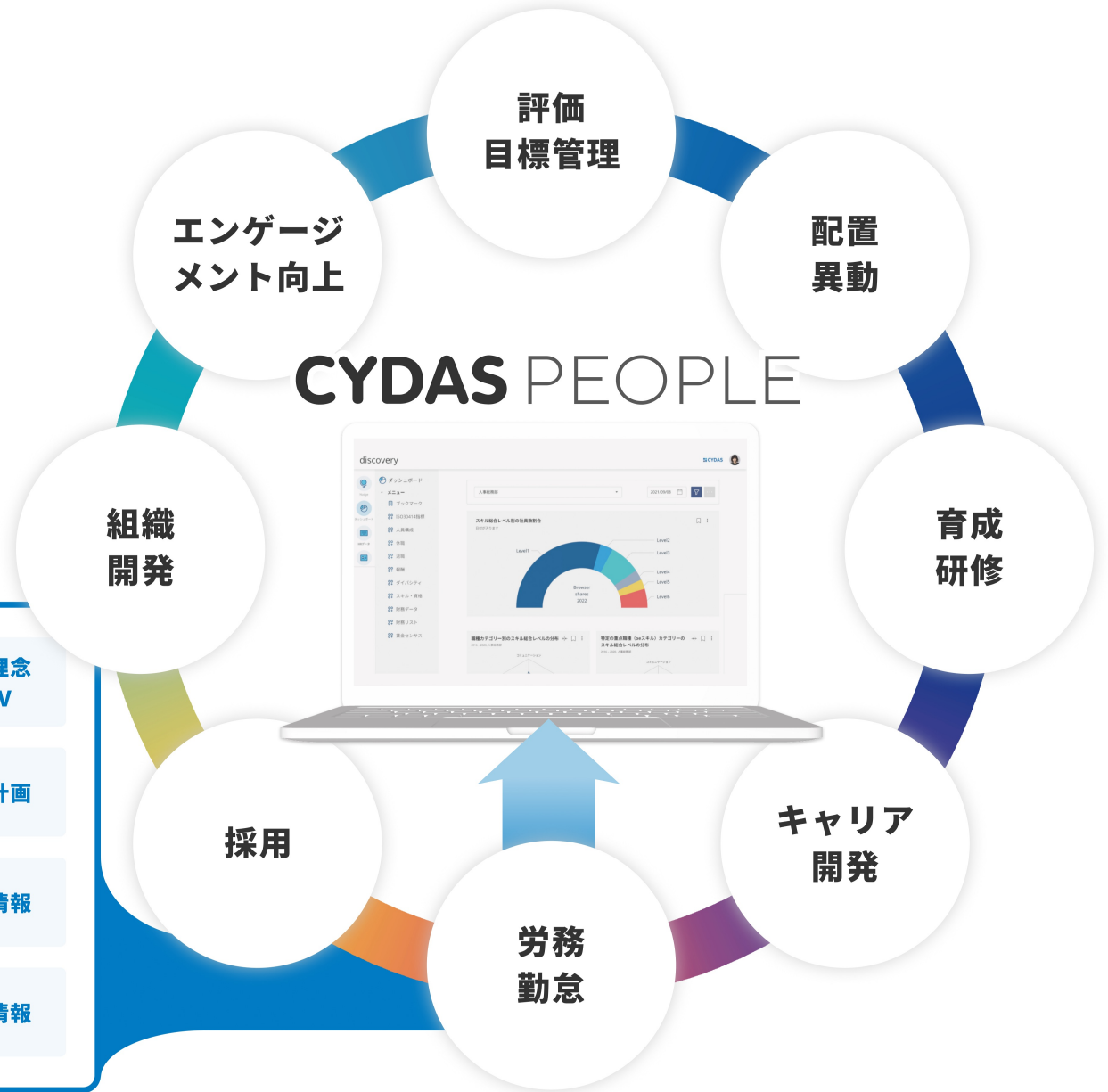
好きなckpt

→ BraV5



タレント マネジメント、 まるっとおまかせ。

公的免許 資格	学歴・職歴	研修履歴	希望職種 働き方	個人の価値観	企業理念 MVV
戸籍・家族	評価	キャリア ビジョン	興味・関心	社員間の 関係性	事業計画
役割 ミッション	1on1・面談	特技・スキル	通勤 希望勤務地	組織図	採用情報
発令情報	言語	勤怠	グレード	給与・賞与	組織情報



RAG(検索拡張生成)アプリ

「CYDAS PEOPLE」に、社員からの問い合わせに自動で答えるChatGPT機能「CYDAS PEOPLE Copilot Chat」

PEOPLE-GPT

CYDAS



https://www.cydac.co.jp/news/press/202304_people-gpt/

HR Expo 春

2週間前

人事FAQ&問い合わせ履歴をベクターデータ化

よくある質問

🔍 キーワードで検索

 勤怠	 給与	 人事考課	 入社と退職
 身上変更	 健康保険、厚生年金保険	 各種証明書の発行	 産休・育休・休職
 カテゴリー9	 カテゴリー10	 カテゴリー11	 カテゴリー12

自社のFAQをベクターデータ化

ニュース

リリースされたニュースがありません

ChatGPTで回答に利用

サイダスお客様サポート
相談チャット

本日はどのようなご用件ですか？



ITパスポートを受験しようと思います、補助は出ますか？

はい、ITパスポートの資格の補助は出ます。ジョブカンで「一般経費支出申請」として申請を行ってください。申請時に以下の項目を必ず入力してください。不明点があれば管理部まで連絡してください。

- ① 申請タイトル
- ② 管理項目
- ③ 費用発生部門
- ④ 支払先名
- ⑤ 発生日。

デモ

🔄 もう一度聞く

補助金の申請方法や条件はどのようなものがあるの...

ITパスポートの受験費用はいくらぐらいかかるの...

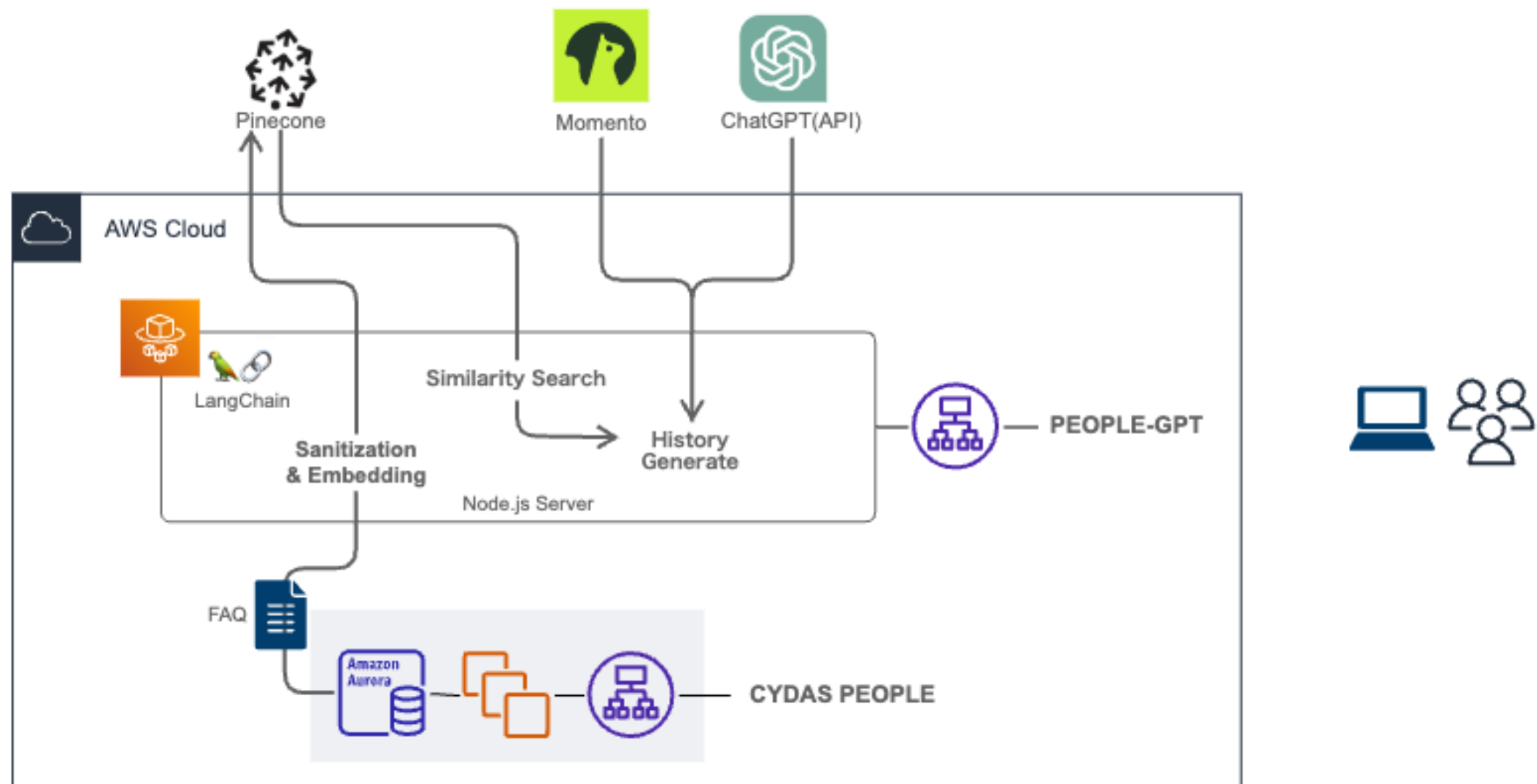
ITパスポートの受験に必要な準備や勉強方法はどのよう...

質問候補を表示しています

何でも聞いてください...

0/2000

[0] DEMOパターン



本番リリリースにむけて

デモの10倍以上やることがある

本番用のLLMアプリケーションをつくる

LLMでクールなものを作るのは簡単だが本番レベルのものを作るのは非常に難しい

1. 課題と解決策

1. 自然言語のあいまいさ(決定論的でない)によるUX不整合 → temperature、プロンプト最適化・バージョニング管理
2. コストとレイテンシー → 推論に使うモデル選定、トークン数の調整、(現状は)レイテンシー解析
3. プロンプティング / ファインチューニング / それ以外
 - プロンプトチューニング：Embeddingsを入れ替える(OSSでのみ可能)
 - ファインチューニング：タスクに最適化させるためのサンプルが性能に影響→が、あまり期待できない
 - 蒸留ファインチューニング：Text-davinci-003(大)でLLaMA(小)のチューニングに使う→Alpaca(小)
 - Embeddings + ベクターストア：Few-shots
4. 後方互換性と前方互換性
 - 新しいモデルは全体的に優れてるかもしれないが、劣っているユースケースもあるはず
 - 前のモデルに対してうまく動いていたプロンプトがうまく動かない場合、どんな意図でプロンプトがチューニングされているか(呪文 = バッドプラクティス)わからなくなる

<https://huyenchip.com/2023/04/11/llm-engineering.html>

本番用のLLMアプリケーションをつくる

2. 複数のタスクで制御されるフロー構成をどうするか、ツールの組み込み
 - プログラミングによる制御フローの実装
 - LLMエージェントによる逐次制御
3. 有望なユースケース
 - AIアシスタント：スケジュール管理、メモ作成、ペアプロ、メール対応、予約、買い物 = タスクをこなす
 - チャットbot：コンパニオン = 文字や音声で相談やアイデアなどの会話
 - プログラミングやゲーム：Webアプリをつくる、セキュリティの脅威を発見する、ゲームを作成する
 - ラーニング：本の要約、講義、理解度テスト、採点やフィードバック、ディベート相手
 - 社内データ抽出：自然言語やQ&A形式で社内のデータを照会する
 - 検索・推薦：キーワードではなく自然言語から推定する検索
 - 営業効率アップ
 - SEO対策
- 結論
 1. 宣伝文句を無視して6ヶ月後に定着してるか確認する
 2. サマリーを読む
 3. 最新のツールに触れるように心がける

<https://huyenchip.com/2023/04/11/llm-engineering.html>





技術選定

データ保護(コンプライアンス)、実行環境、メモリ、検索、ベクターストア

コンプライアンス + AIOAI → Azure

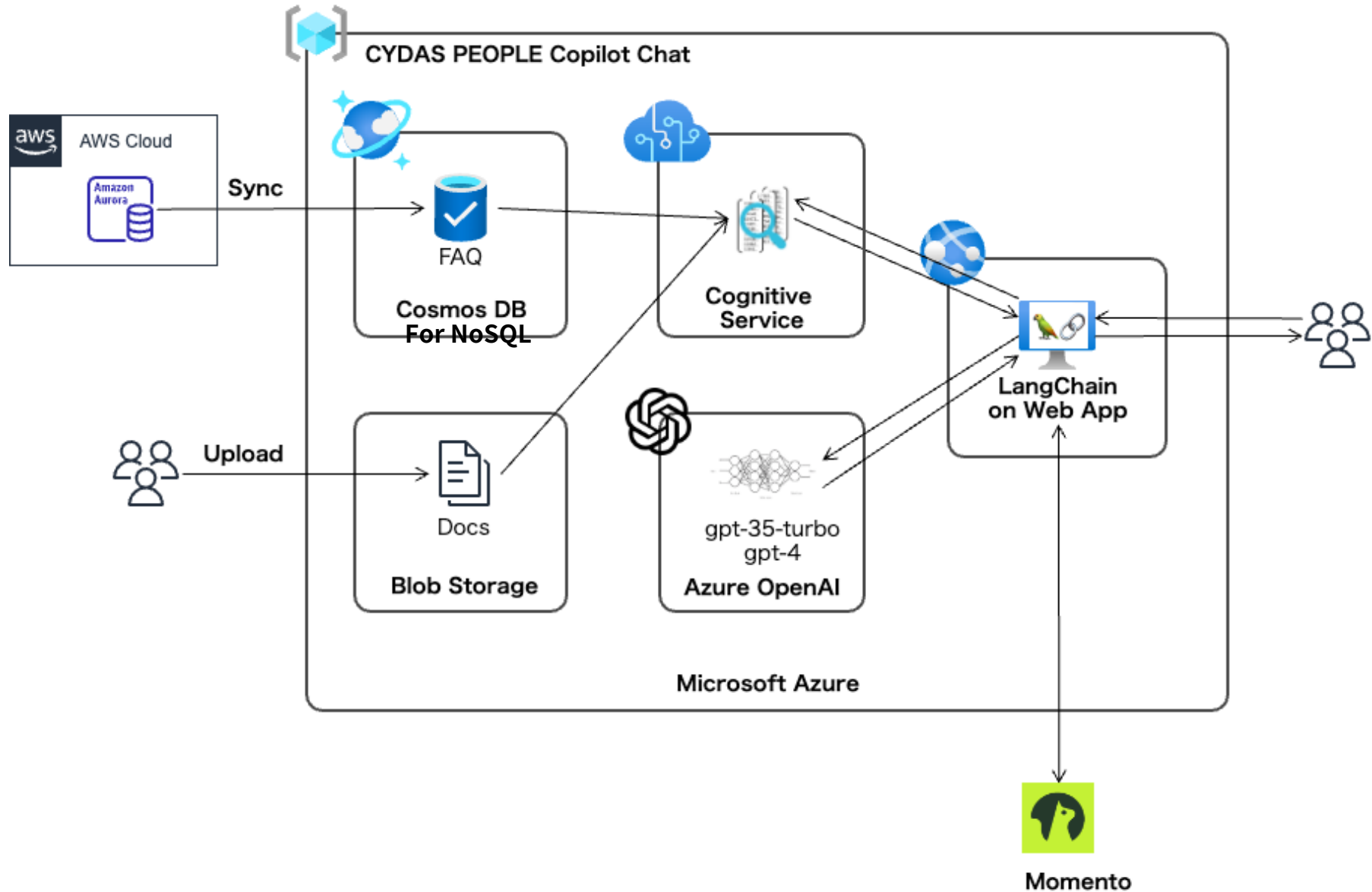
- コンプライアンス
 - ISO 22301 (BCP)
 - ISO 27001 (ISMS) / 27017 / 27018
 - ISO 9001 (品質)
 - SOC 1,2,3 Type 2
 - その他の第三者認証
- コンポーネント
 - Azure OpenAI
 - Cognitive Search

OpenAI(API版)とAzure OpenAIの違い

プロダクト名称	① OpenAI API (OpenAI社側が開発・学習させたモデルをAPIを通して提供)	② Azure OpenAI Service (①をMicrosoft Managedサービスとして提供)	③ Azure Cognitive Service for Language (Azure Cog serviceのうちの1つ)
対象ユーザー	<ul style="list-style-type: none"> • 最新のGenerative AI modelに興味のある企業 (特にスタートアップ企業) • 既存のAzureユーザーではない企業 	<ul style="list-style-type: none"> • ①のユーザーに加えてエンタープライズレベルでGenerative AIを実装したい企業 • Azure AIやMicrosoft工コシシステムの既存顧客 	<ul style="list-style-type: none"> • アプリやサービスにAIを組み込みたいAI開発者 • データサイエンス系ロールの採用が難しい企業
できること	テキスト・コード・画像生成の幅広いユースケース <ul style="list-style-type: none"> • 柔軟なテキスト、コード、画像生成API • 多様なアプリケーションを構築する企業 • 文字生成、要約、チャットボット、コード生成、デザインツールでの画像生成など • 最新モデルによるPoV/MVPなど 	①に記載されている項目に加え、 <ul style="list-style-type: none"> • キャパシティを必要とするプロダクション環境 • グローバルなユーザーベースを持ち地理的な多様性を必要とするサービス • AI倫理のフィルタリングや責任あるAIチェックを必要とするサービス • コンプライアンス、データプライバシー、セキュリティのニーズがあるサービス 	広く知られている自然言語処理のユースケース <ul style="list-style-type: none"> • 文字抽出、分類、チャットボット、抽出要約など、多言語の正式なサポート • 高品質、低コスト、多言語機能 • テキストやあらゆる形式のドキュメントを翻訳 • 迅速なターンアラウンドと最小限のチューニングを必要とするプロジェクト
価格/料金形態	<ul style="list-style-type: none"> • 1,000トークンごとに課金、主に学習と推論 • プロトタイプに最適化 (低利用頻度) 	<ul style="list-style-type: none"> • 1,000トークンごとに課金、主に学習、推論、ホスティング • 実運用に最適化されたファインチューニングモデル 	<ul style="list-style-type: none"> • 1,000テキストレコードごとに課金
プロダクション/エンタープライズへの対応	<ul style="list-style-type: none"> • エンタープライズグレードの機能は限定的 • 請求はクレジットカード支払いのみ 	<ul style="list-style-type: none"> • エンタープライズグレードに特化した機能多数 • グローバルアベイラビリティゾーン • Responsible AI (責任あるAI) への対応 • Billing/リソース管理がほかAzure環境と統合容易 	<ul style="list-style-type: none"> • ②同様にエンタープライズグレードに特化 • グローバルアベイラビリティゾーン • Responsible AI対応 • Billing/リソース管理がほかAzure環境と統合容易
セキュリティコンプライアンス	<ul style="list-style-type: none"> • Single Sign On (SSO) 認証 • 3rd partyサービス認証 • SLA 99.5% 	<ul style="list-style-type: none"> • Azure Active Directory(AAD)対応 • ロールベースのアクセス制御 (RBAC) • コンプライアンス、データセキュリティ、SLA保証 	<ul style="list-style-type: none"> • Azure Active Directory(AAD)対応 • ロールベースのアクセス制御 (RBAC) • コンプライアンス、データセキュリティ、SLA保証
その他の特徴	<ul style="list-style-type: none"> • 最新モデルが実装された場合、まず①から利用可能になる (例: ChatGPT gpt-3.5 turbo, GPT-4 etc) 	<ul style="list-style-type: none"> • 最新モデルは①に入ったあとにごちらも実装される • 一部モデルのファインチューニングも可能 	<ul style="list-style-type: none"> • ①/②を利用せずともユースケースやシナリオ上、この③だけでもカバーできる点もある

[本番] Azure OpenAI / Cognitive Cosmos DB / Momento



Webフロント

Web App (App Service)

- CircleCIとのIAM認証連携
- 実行環境の保守管理の容易さ
 - Requirement.txt に環境指定するだけ = 便利
 - ロバスト性
 - VNet経由でAOAIと接続可能



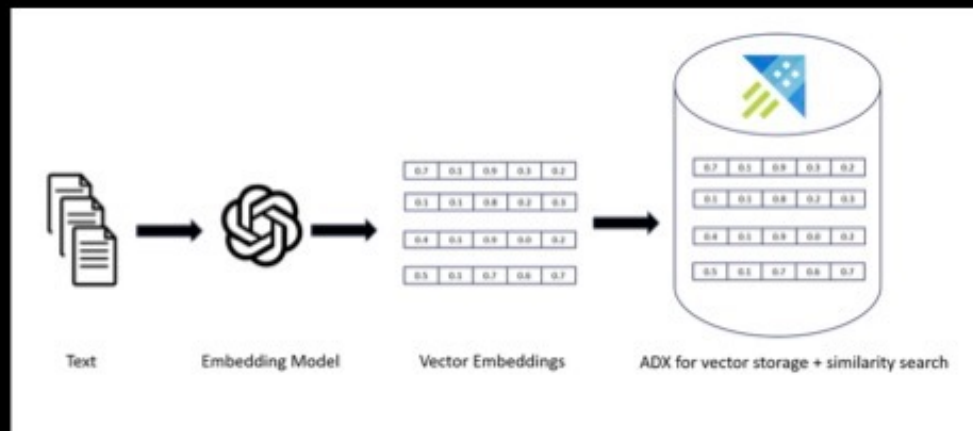
Momento (アプリキャッシュ)

- 完全なサーバーレス
- TTLで揮発するキャッシュ
- 高速かつ安定



(参考) Azure のベクトル検索対応データストア

- [NEW] Azure Cosmos DB (Mongo DB vCore)
- [NEW] Azure Cognitive Search
 - <https://github.com/Azure/cognitive-search-vector-pr>
- Azure Cache for Redis Enterprise
 - <https://techcommunity.microsoft.com/t5/azure-developer-community-blog/vector-similarity-search-with-azure-cache-for-redis-enterprise/ba-p/3822059>
- Azure Data Explorer
 - <https://techcommunity.microsoft.com/t5/azure-data-explorer-blog/azure-data-explorer-for-vector-similarity-search/ba-p/3819626>



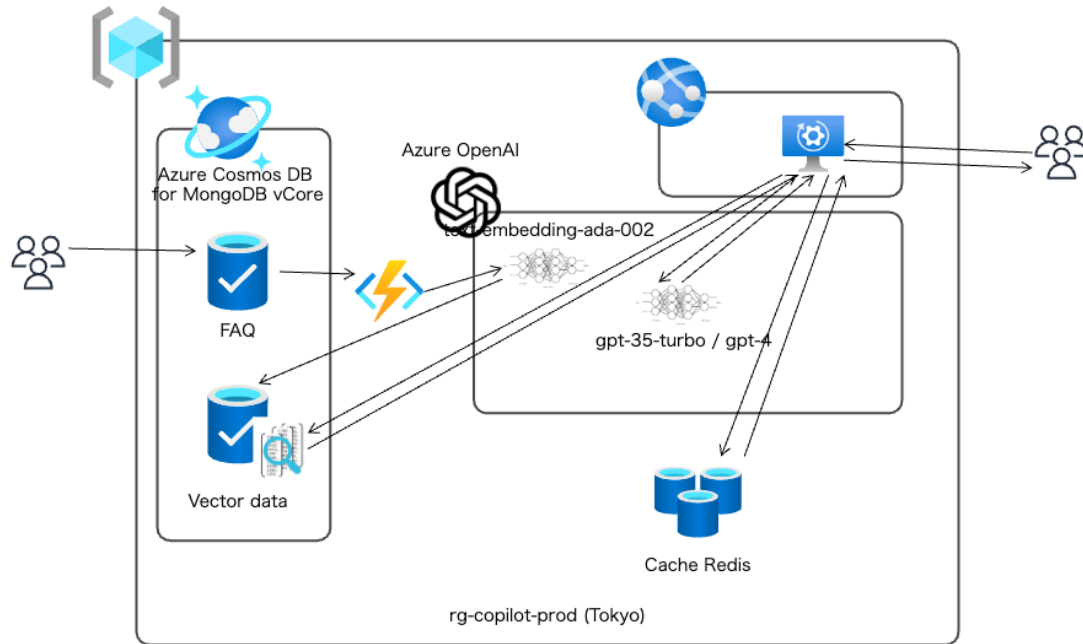
Microsoft Build 2023 で発表された Cosmos DB 注目アップデート

<https://speakerdeck.com/miyake/microsoft-build-2023-cosmos-db-update?slide=9>

Cosmos DB vs Cognitive Search

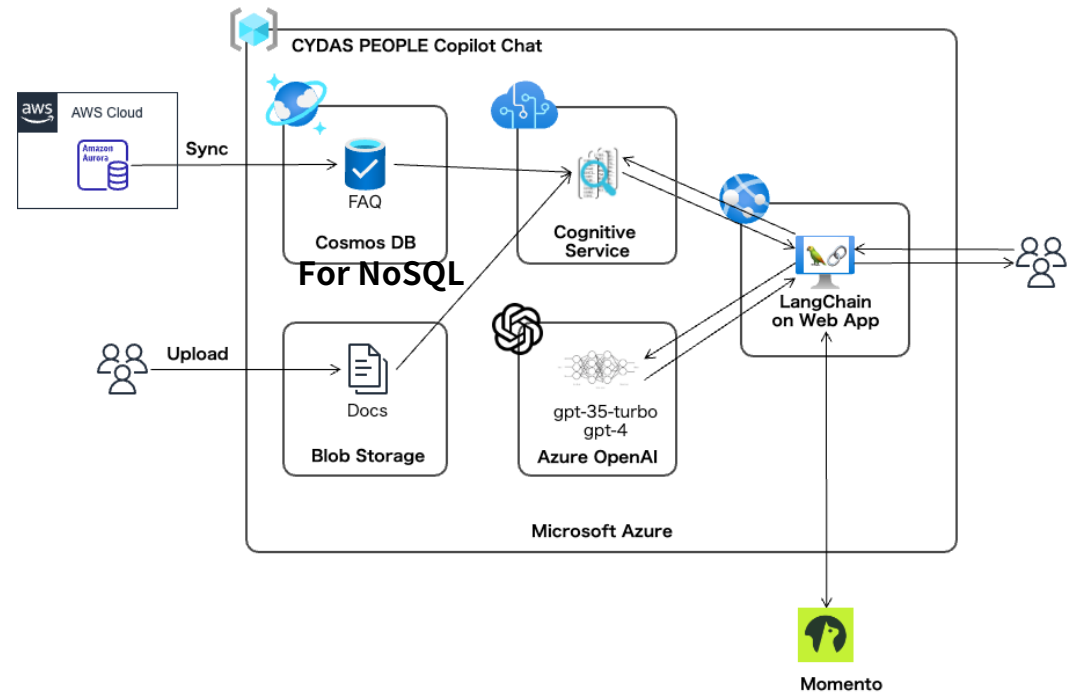
1. Azure Cosmos DB for MongoDB vCore

[1] Cosmos DB with Embeddings パターン



2. Cognitive Search with Cosmos DB & Blob Storage

[本番] Azure OpenAI / Cognitive Cosmos DB / Momento



Cosmos DB vs Cognitive Search

1. Azure Cosmos DB for MongoDB vCore

```
const queryVector = await getEmbeddings(prompt);
let aggCursor = collection.aggregate([
  {
    $search: {
      cosmosSearch: {
        vector: queryVector,
        path: "vectorContent",
        k: 1,
      },
      returnStoredSource: true,
    },
  },
]);
```

2. Cognitive Search with Cosmos DB & Blob Storage

```
const searchResults = await client.search(prompt);
let arr = [];
for await (const result of searchResults.results) {
  arr.push(result);
}
arr.sort((a, b) => b.score - a.score);
let content = "";
if (arr.length) {
  content = arr[0].document.content;
}
```

Cosmos DB vs Cognitive Search

1. Azure Cosmos DB for MongoDB vCore

```
{
  _id: new ObjectId("647ef5bb81ab444849f7b7e8"),
  content: '<xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx>',
  vectorContent: [
    -0.0063294927, 0.0055967104, -0.0017361568, -0.01833973,
    (中略…1536次元全部)… 1436 more items
  ]
}
```

2. Cognitive Search with Cosmos DB & Blob Storage

```
{
  score: 9.943382,
  highlights: undefined,
  document: {
    content: xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx',
    metadata_storage_path: xxxxxxxxxxxx'
  }
}
{
  score: 4.5703325,
  highlights: undefined,
  document: {
    content: 'xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx',
    metadata_storage_path: xxxxxxxxxxxx'
  }
}
:
```

検索

Cognitive Search

- マネージドな検索サービス
 - インデックス作成や検索をまかせられる(Embedding不要)
- 高速・高機能
 - 検索スコア
 - 文書インデックス時のタイトル/パス/ファイル名



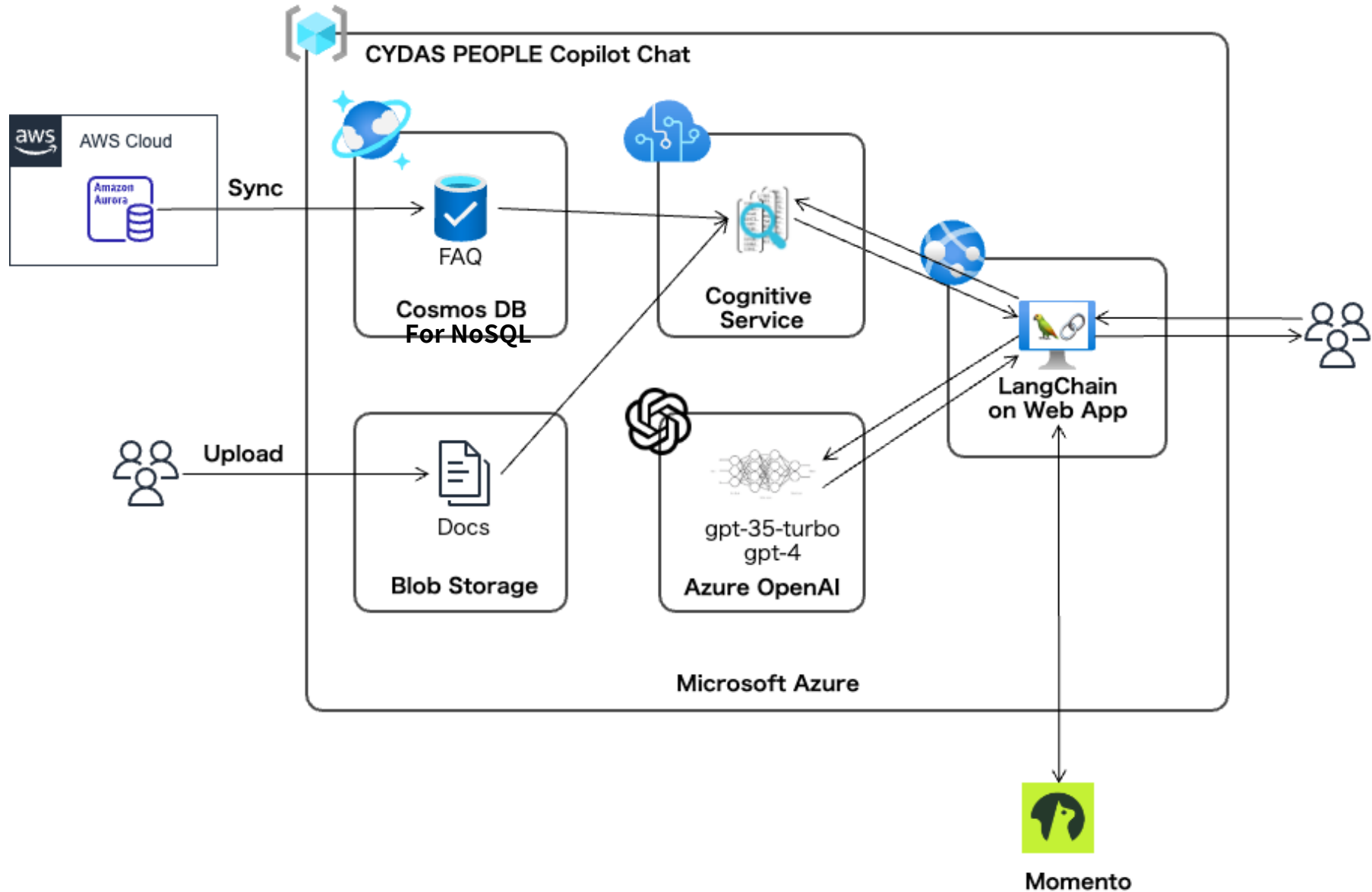
Azure Cognitive Search

Cosmos DB (NoSQL)

- 完全なサーバーレス
- 高速かつ安定



[本番] Azure OpenAI / Cognitive Cosmos DB / Momento



アプリケーションの工夫

- 3つの関連質問：文字入力を最小限に
 - 類似検索結果の2～4位を再利用することで検索最小化
 - 3つない分は汎用的な質問を生成
- 回答を「GPT-4」で、関連質問などその他をgtp-35-turboで
- チャンクサイズを揃える：FAQデータの事前整備、要約
- (今後)Cognitive Search 対象を FAQ + 手動ドキュメントに
 - PDFのチャンクは制御が難しい = 5000文字(ページ)かセンテンス※
ページとほぼ同じ
- (今後)セレクトの実装 = スコアに合わせてデモンストレーションデータの個数を調整する

コンテンツ：安全性・透明性・正確性

- Moderation: コンテンツフィルター
 - 本家OpenAIもLangChainでチェーンできるが冗長
 - 高度なフィルタリングが可能(要申請)
- 敵対的プロンプト抑止
 - プロンプトインジェクションを検知・無視するプロンプト
- バイアスの抑止
 - ふるまいの指定
- ハルシネーション回避：わからないときはわからないと答えさせる
- 参照元情報の表示
 - FAQのリンク
 - PDFのファイル名/リンクの追加
 - Output Parserで応答をJSONで構造化されたデータ形式で受領
- AIによって回答を生成している注意表示
 - 生成AI利用におけるガイドライン(サイダス社)で必須

temperatureはRAGの応答品質にどのくらい影響するのか？

- 温度：生成されるテキストの「創造性」またはランダム性を制御するパラメーターである(GPT-3以降)
 - 温度が高い(ex. 0.7)→より多様で創造的な出力
 - 温度が低い(ex. 0.2)→より決定論的で集中的な出力
 - top_pサンプリング(コアサンプリング)との併用は非推奨

ユースケース	温度	説明
コード生成	0.2	確立されたパターンと慣習に従ったコードを生成する。出力はより決定論的で焦点化される。構文的に正しいコードを生成するのに役立つ。
クリエイティブ・ライティング	0.7	ストーリーテリングのための創造的で多様なテキストを生成する。アウトプットはより探索的で、パターンによる制約が少ない。
チャットボット応答	0.5	一貫性と多様性のバランスのとれた会話応答を生成します。アウトプットはより自然で魅力的。
コード・コメント生成	0.3	より簡潔で適切なコードコメントを生成します。出力はより決定論的で、規約を守る。
データ分析スクリプト	0.2	より正確で効率的なデータ分析スクリプトを生成します。出力はより決定論的で集中的です。
探索的なコーディング	0.6	代替案や創造的なアプローチを模索するコードを生成する。既成のパターンにとらわれないアウトプット

<https://platform.openai.com/docs/api-reference/chat>

<https://community.openai.com/t/cheat-sheet-mastering-temperature-and-top-p-in-chatgpt-api-a-few-tips-and-tricks-on-controlling-the-creativity-deterministic-output-of-prompt-responses/172683>



```
import langchain
from langchain.prompts import PromptTemplate

temperature = 2.0

template = """
あなたは優秀なアシスタントです。以下の知識を参考にしてください。
{retrieval}
---
この知識を参考に、以下の質問に教えてください。
{question}
"""

langchain.verbose = True

prompt = PromptTemplate(
    input_variables=["retrieval", "question"],
    template=template,
    temperature=temperature
)

result = prompt.format(retrieval="りんごは赤い、レモンは黄色い", question="きゅうりは?")
print(result)

response = chat.predict(result)
print(response)
```



```
あなたは優秀なアシスタントです。名前は PEOPLE Copilot です。以下の知識を参考にしてください。
りんごは赤い、レモンは黄色い
---
この知識を参考に、以下の質問に教えてください。
きゅうりは?

色の情報が提供されていないため、きゅうりの色はわかりません。
```

Few-shots

モデル	temp	答え	判定
gpt-3.5-turbo	0.1	赤くも黄色くもなく、きゅうりは緑色です。	
gpt-3.5-turbo	0.1	私の知識にはきゅうりに関する情報が含まれていません。申し訳ありません。	
gpt-3.5-turbo	0.1	色が緑色です。	
gpt-3.5-turbo	0.1	答えることができません。きゅうりに色は決まっています。	
gpt-3.5-turbo	0.1	色は緑です。	
gpt-3.5-turbo	0.1	「きゅうりは緑色です。」と答えることができます。	
gpt-3.5-turbo	0.1	私は人工知能のアシスタントであり、知識の範囲は限られています。お手数をおかけしますが、きゅうりに関する質問には回答できません。他に何かお探しの情報がありましたら、お知らせください。	
gpt-3.5-turbo	0.1	色を指定していないため、きゅうりの色については何も言えません。	
gpt-3.5-turbo	0.1	きゅうりの色に関する情報は提供されていません。	
gpt-3.5-turbo	0.1	きゅうりの色についての情報は提供されていないため、答えることはできません。	
gpt-3.5-turbo	2	きゅうりの色は緑色です。	
gpt-3.5-turbo	2	私は知りません。きゅうりの色は様々な種類があるため、一概に答えることはできません。	
gpt-3.5-turbo	2	色を示す情報が与えられていないため、きゅうりの色については答えることができません。	
gpt-3.5-turbo	2	色の情報が与えられていないため、きゅうりの色に関する答えはありません。	
gpt-3.5-turbo	2	色の指定がないため、きゅうりの色は不明です。	
gpt-3.5-turbo	2	りんごやレモンとは異なり、きゅうりは緑色です。	
gpt-3.5-turbo	2	「きゅうりは緑色です」となります。	
gpt-3.5-turbo	2	色に関する情報が与えられていないため、きゅうりの色についてはわかりません。	
gpt-3.5-turbo	2	色の情報が提供されていないため、きゅうりの色については何も言えません。	
gpt-3.5-turbo	2	色の情報が提供されていないため、きゅうりの色はわかりません。	

「緑色」と答える割合 → 0.1でも2.0でも30%か40%









































もう少し実践的に

質問： ITパスポートを受験しようと思います。補助は出ますか

検索結果：

Document(page_content='ITパスポートの資格の補助はでますか？ ¥n¥nジョブカンで「一般経費支出申請」として申請をすること。その際次の項目は必須なので入力すること。不明点があれば管理部まで連絡すること。¥n①申請タイトル¥n②管理項目¥n③費用発生部門¥n④支払先名¥n⑤発生月。¥n', metadata={'@search.score': 4.32952, 'rid': 'WGxwaUFOWU4xaG9IQUFBQUFBQUFBQT090'}),

RAG同等

gpt-3.5-turbo	0.1	ジョブカンの一般経費支出申請で申請することができます。不明点があれば管理部に直接問い合わせることをお勧めします。	 
gpt-3.5-turbo	0.1	ジョブカンの「一般経費支出申請」として申請することで補助が出ます。申請時に必要な項目を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	0.1	ジョブカンで「一般経費支出申請」として申請をすることで補助を受けることができます。必要な項目を入力し、不明点があれば管理部に直接連絡してく	 
gpt-3.5-turbo	0.1	ジョブカンで「一般経費支出申請」として申請することで補助が出ます。申請時に必要な項目を入力し、不明点があれば管理部に連絡するようにしてくださ	 
gpt-3.5-turbo	0.1	ジョブカンの一般経費支出申請で申請することになっています。不明点があれば管理部まで連絡してください。	 
gpt-3.5-turbo	0.1	ジョブカンの一般経費支出申請で申請することにより、補助を受けることができます。必要な項目を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	0.1	ジョブカンの一般経費支出申請で申請することができます。不明点があれば管理部に直接問い合わせるようにしてください。	 
gpt-3.5-turbo	0.1	ジョブカンで「一般経費支出申請」として申請することで、補助を受けることができます。必要な項目を入力し、不明点があれば管理部に連絡してくださ	 
gpt-3.5-turbo	0.1	ジョブカンの一般経費支出申請で申請することにより、ITパスポートの資格の補助を受けることができます。必要な項目を入力し、不明点があれば管理部に	 
gpt-3.5-turbo	0.1	ジョブカンで「一般経費支出申請」として申請をすることで補助が出ます。必要な項目を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することができます。不明点があれば管理部に直接お問い合わせください。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することによって補助が受けられます。必要な項目を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することができます。不明点があれば管理部に直接お問い合わせください。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することになっています。不明点がある場合は管理部に直接問い合わせてください。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することができます。不明点があれば管理部に直接連絡してください。	 
gpt-3.5-turbo	2	ジョブカンで「一般経費支出申請」として申請することで、ITパスポートの資格の補助を受けることができます。申請時に必要事項を入力し、不明点があれ	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することになっています。不明点があれば管理部に直接問い合わせてください。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請をすることで補助を受けることができます。申請時に必要な項目を入力し、不明点があれば管理部に連絡するようにしてくだ	 
gpt-3.5-turbo	2	ジョブカンの「一般経費支出申請」として申請することで補助が出ます。申請に必要な項目を入力し、不明点があれば管理部に連絡するようにしましょう。	 
gpt-3.5-turbo	2	ジョブカンの一般経費支出申請で申請することにより、ITパスポートの資格の補助を受けることができます。必要事項を記入し、管理部に不明点があれば連	 

回答の質がほぼ同じ = ①補助が出ること②具体的な手順

プロンプトチューニングによる貢献

- ふるまいの指定による変化
 1. ふるまい指定なし
 2. ふるまい指定あり
- プロンプトを日本語で書く場合/英語で書く場合
 3. ふるまい指定あり / 英語









































ふるまいを指定しない

以下のチャット履歴を参考に質問に答えてください。

```
{retrieval}
```

以下の質問に日本語で答えてください。

```
{question}
```

0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで補助を受けることができます。必要な項目を入力し、不明点があれば管理部に連絡するようにして	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することになっています。詳細は管理部にお問い合わせください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することになっています。詳細は管理部まで連絡してください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することになっています。詳細は管理部にお問い合わせください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することができます。ただし、詳細については直接管理部に連絡してください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することになっていますが、具体的な補助の有無については、企業によって異なるため、人事に直接お問い合わせください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ITパスポートの受験料の補助は、福利厚生資格受験料の補助と同様に、ジョブカンの一般経費支出申請で申請することができます。ただし、領収書が必要です。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」で申請することで、ITパスポートの受験費用の補助が出る場合があります。不明点があれば管理部に連絡してください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することによって、ITパスポートの受験料の補助が出ます。ただし、領収書が必要です。不明点があれば管理部に連絡してください。	 
0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請する必要があります。不明点があれば管理部に連絡してください。また、領収書が必要です。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することによって補助が受けられます。ただし、不明点があれば管理部に連絡するようにしてください。申請する際には、申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目が必須になります。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」として申請することで補助が出ます。申請する際には、申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目を入力する必要があります。不明点があれば管理部に連絡してください。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」として申請することで、ITパスポートの受験料の補助が出ます。申請する際は、申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目を必ず入力してください。不明点があれば管理部に連絡してください。また、領収書が必要です。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することで、ITパスポートの資格の補助を受けることができます。申請時には、申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目を必ず入力するようにしてください。不明点がある場合は、管理部に連絡してください。また、申請には領収書が必要です。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することにより、ITパスポート受験料の補助が出る可能性があります。不明な点は管理部に連絡してください。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ITパスポートの資格の補助については、チャット履歴に以下のように記載されています。「ジョブカンで「一般経費支出申請」として申請すること。その際の項目は必須なので入力すること。不明点があれば管理部まで連絡すること。n①申請タイトルn②管理項目n③費用発生部門n④支払先名n⑤発生月。」ですので、ジョブカンの一般経費支出申請を行い、必要な項目を入力することで補助を受けることができます。ただし、詳細については管理部に確認することをおすすめします。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することにより、福利厚生にある資格受験料の補助が受けられます。ただし、領収書が必要です。詳細は管理部へ連絡してください。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請する必要があります。不明点があれば管理部に連絡することをお勧めします。また、領収書が必要です。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することで、ITパスポートの受験料の補助が可能です。不明点があれば管理部に連絡してください。必要な情報として、申請タイトル、管理項目、費用発生部門、支払先名、発生月が必要です。また、領収書が必要です。	 
2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請する必要があります。不明点があれば管理部に連絡してください。また、領収書が必要です。	 

結果はあまり変わらないが、回答の質はむしろ温度が高いほうが良い

ふるまいを指定する

あなたの名前はPEOPLE Copilot Chatです、あなたは以下のようなチャット履歴で従業員の質問に答えるHRエージェントです。

```
----  
{retrieval}
```

上記の知識を日本語で考えながら、以下の質問に答えてください。

ただし、参考にする知識がない場合は、"その件については存じ上げませんので人事に直接お問い合わせください。"と回答してください。

```
{question}
```

gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	はい、ITパスポートの資格の補助は出ます。ジョブカンで「一般経費支出申請」として申請することになります。申請タイトル、管理項目、費用発生部門、支払先名、発生月は必須項目となっています。不明点があれば管理部に連絡することをお勧めします。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ITパスポートの資格の補助は、ジョブカンの一般経費支出申請で申請することができます。申請タイトル、管理項目、費用発生部門、支払先名、発生月を入力する必要があります。不明点があれば管理部に連絡するように指示されています。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	はい、ITパスポートの資格の補助は出ます。ジョブカンで「一般経費支出申請」として申請することになります。申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目は必須になります。不明点があれば管理部まで連絡することになります。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで受験費用の補助が出るそうです。申請する際には必要な項目を入力することが必要です。不明点があれば管理部まで連絡してください。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」として申請することで補助が出ます。申請タイトルや費用発生部門、支払先名、発生月など、必要な項目を入力して申請してください。不明点があれば管理部に連絡するようにしてください。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで補助を受けることができます。申請時に必要な項目がありますので、確認してください。不明点があれば管理部に連絡するようにしてください。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」として申請することで補助が出ます。申請時に必要な項目を確認し、管理部に不明点があれば連絡するようにしましょう。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	はい、ITパスポートの資格の補助は出ます。ジョブカンで「一般経費支出申請」として申請し、必要な項目を入力してください。不明点があれば管理部に直接連絡してください。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ITパスポートの資格の補助は、ジョブカンで「一般経費支出申請」として申請することで補助されます。申請する際に必要な項目がありますので、不明点があれば管理部まで連絡してください。	
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ITパスポートの資格の補助は、ジョブカンの一般経費支出申請で申請することができます。申請タイトル、管理項目、費用発生部門、支払先名、発生月を入力する必要があります。不明点があれば管理部に連絡してください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請する必要があります。具体的には、申請タイトル、管理項目、費用発生部門、支払先名、発生月の入力が必要で、不明点があれば管理部に連絡してください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請をすることで補助が出ます。申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目は必須です。もし不明点があれば、管理部まで連絡してください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」として申請をすることで補助が出ます。次の項目が必須となるので入力することが必要です。申請タイトル、管理項目、費用発生部門、支払先名、発生月です。不明点があれば管理部まで連絡することをお勧めします。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請をすることで、補助を受けることができます。申請の際には、以下の項目が必須となります。申請タイトル、管理項目、費用発生部門、支払先名、発生月となります。不明点があれば管理部に直接お問い合わせください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	補助が出る場合があります。ジョブカンで「一般経費支出申請」として申請をし、必要な項目を入力する必要があります。不明点がある場合は管理部に連絡するようにしてください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで、ITパスポートの資格の補助が受けられます。必要な項目を入力し、不明点があれば管理部に連絡することをおすすめします。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請し、必須項目を入力することで補助を受けることができます。ただし、詳細については、管理部に直接お問い合わせください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	はい、ITパスポートの資格の補助はあります。ジョブカンで「一般経費支出申請」として申請することが必要で、次の項目を入力する必要があります。「①申請タイトル」「②管理項目」「③費用発生部門」「④支払先名」「⑤発生月」。不明点があれば管理部まで連絡することをお勧めします。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで、ITパスポートの資格の補助が出ます。申請の際は、次の項目を必ず入力するようにしてください。申請タイトル、管理項目、費用発生部門、支払先名、発生月です。不明点があれば、管理部まで連絡するようにしてください。	
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することで補助が出ます。申請タイトル、管理項目、費用発生部門、支払先名、発生月が必須項目となります。不明点があれば管理部に連絡してください。	

回答の質が逆転し：温度が低(0.1)くても、具体的な手続き方法を回答できている

プロンプトを英語にする









































Your name is PEOPLE Copilot Chat, you are an HR agent that answer via employees's questions with chat history like below.

{retrieval}

Please answer the question below with thinking about the knowledge above in Japanese.

But if you have no knowledge to refer to, you can answer with "その件については存じ上げませんので人事に直接お問い合わせください。".

{question}

gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することにより補助を受けることができます。不明点があれば管理部に直接連絡してください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することにより補助が出ます。必要な項目を入力し、不明点があれば管理部に連絡することをお勧めします。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することで補助が出ます。必要な項目を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することができます。不明な点があれば管理部に直接お問い合わせください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請にて申請することができます。必要事項を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することができます。不明点があれば管理部まで連絡してください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請し、必要な項目を入力することで補助が出ます。不明点があれば管理部に直接連絡してください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することにより補助が受けられるそうです。必要な項目を入力して申請し、不明点があれば管理部に連絡するようにしてください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請をすることで補助が出ます。申請の際には必要事項を入力する必要があります。不明点があれば管理部に直接連絡してください。	 
gpt-3.5-turbo	0.1	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで補助が出ます。ただし、申請時に必要な項目があるので、不明点があれば管理部に直接連絡してください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請することで補助が出ます。申請タイトル、管理項目、費用発生部門、支払先名、発生月の項目が必須になるので、不明点があれば管理部に直接連絡することをお勧めします。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することになっています。不明点があれば、管理部に直接お問い合わせください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで一般経費支出申請をして、必要な項目を入力することで補助が出ます。不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンで「一般経費支出申請」として申請をすることで補助が出ます。申請の際には必要事項を入力し、不明点があれば管理部に連絡するようにしてください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することができます。ただし、必要書類があるため、管理部に問い合わせることをお勧めします。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することになっていますので、申請していただく必要があります。不明点があれば管理部に直接問い合わせてください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することによって補助が出ます。必要な項目を入力して申請し、不明点があれば管理部に連絡するようにしてください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することで補助が出ます。不明点があれば管理部に直接連絡してください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの「一般経費支出申請」として申請すれば補助が出ます。必要事項を入力し、不明点があれば管理部に連絡してください。	 
gpt-3.5-turbo	2	ITパスポートを受験しようと思います。補助は出ますか	ジョブカンの一般経費支出申請で申請することにより補助が出ます。必要な項目を記入して申請し、不明点があれば管理部に連絡してください。	 

temperatureに関わらず、むしろ回答の質が下がった(手続き方法が案内されなくなった)

応答性能

- 性能影響の大きいLLMの応答性能
- 処理によってモデルを使い分ける
- ストリームモードでUXを向上する



デモ

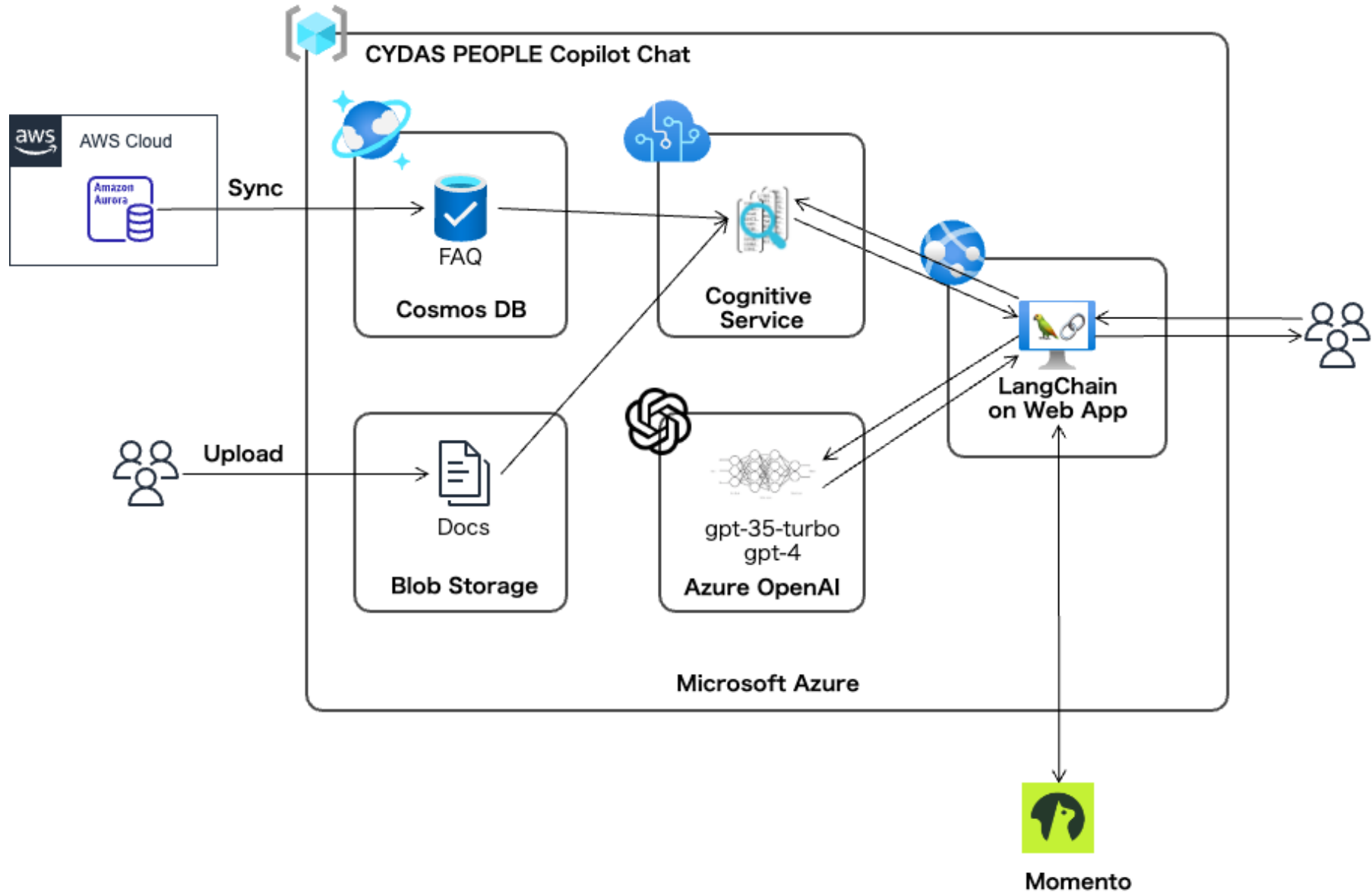
OpenAI (gpt-3.5-turbo)



デモ

AOAI (gpt-35-turbo)

[本番] Azure OpenAI / Cognitive Cosmos DB / Momento



まとめ) OpenAIで本番LLMアプリをつくる

- OpenAIを使うならAzure一択
- Cognitive Search の機能性・非機能性の高さに注目
- アプリの特性に合わせたコンテンツ(入力/出力)の安全性の工夫
- 他のLLMアプリの最適化手法が自分にも万能とはかぎらない
 - temperatureで回答を決定的にする→影響軽微/バランス
 - プロンプトは英語のほうが頭が良い→日本語のほうが質問の意図・回答の丁寧さを再現できる場合あり
- 魔法を科学に着地させるには観測とエビデンス
 - Webアプリと同等の十分なテスト実施
 - パラメータやプロンプトテンプレートのバージョン管理
- 改善要望
 - Azure OpenAI のストリームのチャンクをもっと小さく
 - Cosmos DB for MongoDB vCore の検索スコアパラメータ

4年も待たせてごめんね!!

/serverless/ **DAYS**

TOKYO 2023

2023.09.23(土) カンファレンス

2023.09.24(日) ワークショップ

<https://serverless.connpass.com/>
<https://tokyo.serverlessdays.io/>

DAYS史上最高のラインナップ

- 豪華セッション
 - キーノート：AWS, Momento, TiDB
 - (1)サーバーレス10年の成熟
 - (2)サーバーサイド主体からエッジ/フロントへの開発ライフサイクルの変化
 - (3)生成AIとサーバーレス
- 豪華スポンサー
 - AWS
 - Momento
 - PingCAP (TiDB)
 - Microsoft
 - Cloudflare
 - 【7月から若干募集開始します】