



Azure OpenAI Service リファレンスアーキテクチャ

# Azure OpenAI プレイグラウンド パッケージ (企業内向けの”生成AI チャット利用機能”と”社内ナレッジ検索機能”)

## 株式会社ベーシック

☎ 03-3204-2241(代表)

☎ 03-3204-8490

✉ [contact-us@basic.co.jp](mailto:contact-us@basic.co.jp)

🌐 <https://www.basic.co.jp>



A large, faint watermark of the Basic logo is visible in the bottom right corner of the slide.



## 目次

- シナリオ概要
- アプリUI
- アーキテクチャ
- 考慮事項
- デプロイ方法



# シナリオ概要

Basic

# シナリオ概要

## Azure OpenAIプレイグラウンド パッケージ (企業内向けの”生成AI チャット利用機能”と”社内ナレッジ検索機能”)

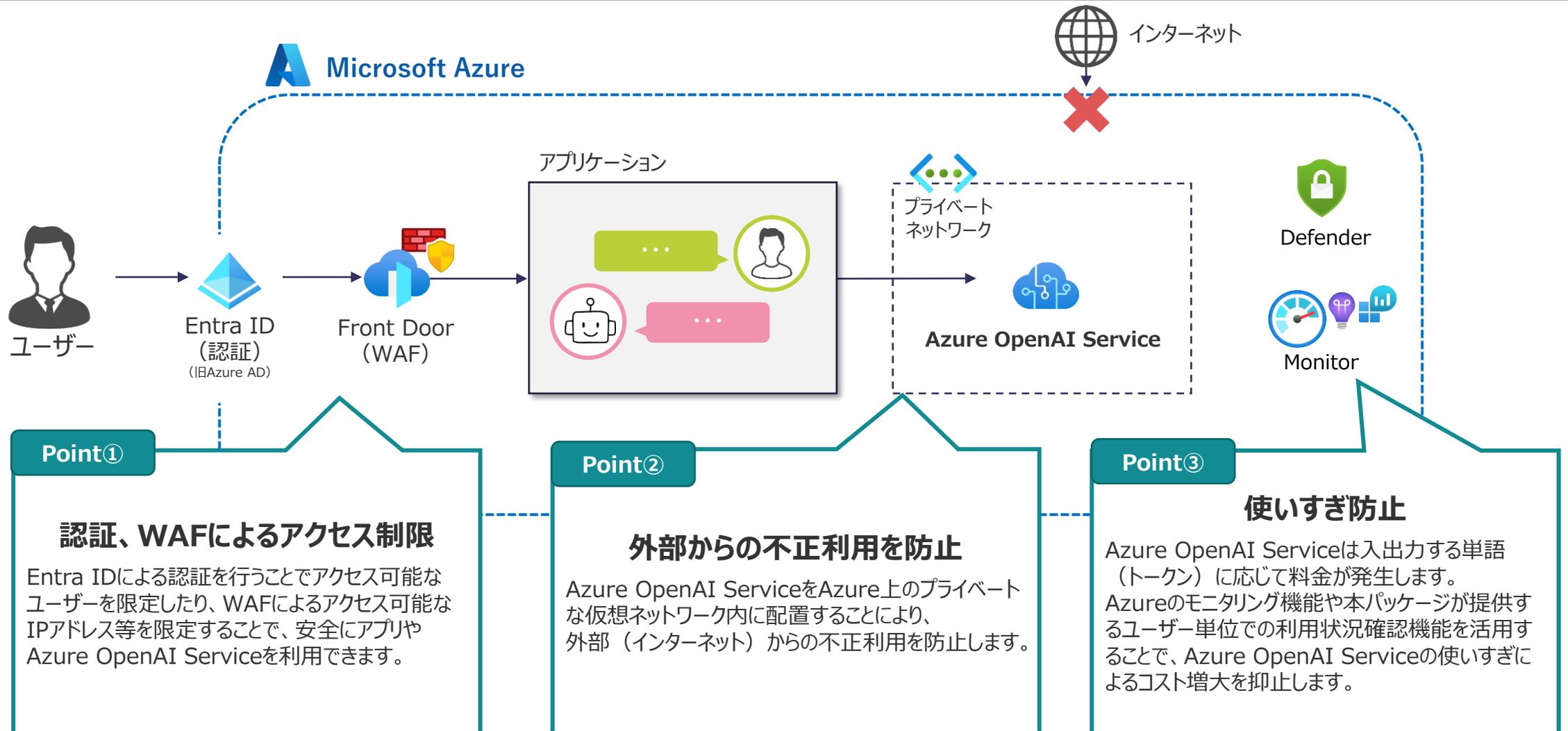
本パッケージは、Azure OpenAI Service及び各種PaaSサービスをAzure上のプライベートな仮想ネットワーク内に配置し利用することで、お客様の企業内に**ChatGPTのような生成AIを安心・安全に利用できる環境**の提供が可能になります。

パブリックな環境で提供されているChatGPTとは異なり、チャットのやり取りが学習データとして利用されることもありません。機能面においても、プロンプト・エンジニアリングによるプロンプトの調整や、Azureのモニタリング機能を用いたユーザー単位での利用状況確認機能の提供が可能になります。

また、Azure Cognitive Searchを組み合わせることで、特定の社内ドキュメント群を検索し、生成AIで返答を行う機能が提供でき、企業内の情報検索を会話ベースで行うことが容易になります。

# シナリオ概要 | Azure OpenAIプレイグラウンド パッケージ

お客様の企業内にAzure OpenAI Serviceを安心・安全に利用できる環境を提供します





# アプリUI



Basic

# アプリUI | 生成AI チャット利用機能① (QAチャット)

Entra ID(旧Azure AD)認証やIP制限で、限られたユーザーのみアクセス可能 (社内ユーザー利用を想定)

The screenshot shows the Azure OpenAI Playground interface. On the left, there is a dark sidebar with the text "Azure OpenAI Playground" and a "Back" button. Below "Back" is a "Get Help" link. The main area is a chat window. At the top left, it says "AI" and "こんにちは！何かお手伝いできることはありますか？". Below this, it says "Token:0 (入力:0 出力:0)". In the center, there is a blue bubble representing the user's input: "You AzureOpenAIセミナーのタイトルと目次を作成してください。". Below that, the AI's response is shown in a white bubble: "AI もちろんです！以下にAzureOpenAIセミナーのタイトルと目次の例を示します。 タイトル: AzureOpenAIセミナー - AIの力を引き出すためのAzureの活用方法 目次: 1. イントロダクション - AzureOpenAIセミナーの概要 - AIとAzureの関係性の紹介 2. Azureの基礎知識 - Azureの概要と特徴 - Azureの主要なサービスの紹介 3. AzureでのAIの活用方法 - Azure Cognitive Servicesの紹介 - Azure Machine Learningの概要 - Azureでのデータの前処理とモデルのトレーニング 4. AzureとOpenAIの連携 - OpenAIの概要と特徴 - AzureとOpenAIの統合方法 - AzureでのOpenAIモデルのデプロイと使用方法 5. Azure OpenAIの活用事例の紹介 - 実際のビジネスケースにおけるAzure OpenAIの活用事例の紹介". At the bottom of the chat window, there is a text input field with the placeholder "メッセージを入力してください" and a blue send button.

# アプリUI | 生成AI チャット利用機能② (アップロードしたドキュメントに基づいたQAチャット)

電子文書をアップロードして、対象の文書に対する質問に回答するQAチャット機能

The screenshot displays the Azure OpenAI Playground interface. On the left, a dark sidebar contains the text "Azure OpenAI Playground", a "Back" button, and a "Get Help" link. The main content area features a document upload section with a dashed border and an upload icon. Below this, a list of uploaded files shows "GOTOEATキャンペーン.pdf" with a search icon. The document preview displays a Japanese poster with the headline "新しい生活様式に基づいた外食の楽しみ方" and a message: "食事の前には、手洗い・消毒をしましょう。" Below the preview, a text input field contains the question "GoToEatキャンペーン事業の目的". A "質問する" button is positioned below the input. The AI response text reads: "GoToEatキャンペーン事業の目的は、感染予防対策に取り組みながら頑張っている飲食店を応援し、食材を供給する農林漁業者を応援することです。キャンペーンは、新しい生活様式に基づいた外食の楽しみ方を提案し、感染防止対策が実施されている店舗を利用することを推奨しています。" At the bottom, a list of reference files shows "GOTOEATキャンペーン.pdf" with a search icon.

# アプリUI | 社内ナレッジ検索機能①（社内ドキュメント群へのQAチャット）

事前に連携した社内文書に対し、様々な質問に回答するQAチャット機能

The screenshot displays the Azure OpenAI Playground interface for a Q&A chat. On the left, a dark sidebar contains the text "Azure OpenAI Playground", "Q&A Chat", and "Keyword Search". At the bottom of the sidebar are "Back" and "Get Help" links. The main chat area shows a conversation:

- AI:** こんにちは！何かお手伝いできることはありますか？  
Token:0 (入力:0 出力:0)
- You:** 家族手当について
- AI:** ファイルの中にその質問に対する答えが見つかりました。  
公式の抜粋は第9条で、家族手当は次の対象家族を扶養している従業員に対して支給されます。  
(1)配偶者: 月額10,000円  
(2)18歳未満の子1人につき月額5,000円（高校在学中は卒業まで）  
ただし、他の対象学校はこれに準じます。  
Token:1811 (入力:162 出力:1649)

At the bottom of the chat area, there is a text input field with the placeholder text "メッセージを入力してください" and a blue send button.

# アプリUI | 社内ナレッジ検索機能② (社内ドキュメントのフリーワード検索)

事前に連携した社内文書に対しては、キーワードに対する文書検索も可能

The screenshot displays the Azure OpenAI Playground interface. On the left, there is a dark sidebar with the text 'Azure OpenAI Playground', 'Q&A Chat', and 'Keyword Search'. The main area shows a search input field with the text '手当' (allowance) and a blue '検索' (Search) button. Below the search bar, the results are displayed under the heading '検索結果 (全6件)'. Three document entries are visible:

- 各種手当規程\_p1.doc**  
(役職手当) 第2条 役職手当は、係長以上の地位にある役職者に対し、その職責に応じ支給する。  
(職務手当) 第3条 職務手当は、職務に応じ、特別に資格および技術を有する者に対し、支給する。  
2 自転車通勤手当が支給される者には、第10条に定める通勤手当は支給されない。  
(在宅勤務手当) 第6条 在宅勤務手当は、在宅勤務制度を希望する社員が所定の書面で申請し、会社が認めた場合に支給する。  
(特別手当) 第7条 特別手当は、会社備別に支糸合する必要があると認めた場合に支給する。
- 各種手当規程\_p2.doc**  
(住宅手当) 第8条住宅手当は、世帯主の従業員に対し次の区分により支給する。  
2 毎月の勤手当の上限額は30,000円とする。  
届出が遅れた場合、または虚偽の申告をした場合、通勤手当を支払わない。  
また、支給済みの通勤手当に関しては返還を命ずることがある。  
6 自転車通勤者の通勤手当は、第5条の定めによるものとし、本条に定める通勤手当は支給されない。
- 通勤手当支給規程\_p2.doc**  
(申請手続) 第4条 通勤手当の支給を受けようとする者は、別に定める「通勤手当支給申請書」を、所属長を経由して総務部長に提出するものとする。  
2 次に該当する場合には、遅滞なく、別に定める「通勤手当変更申請書」を提出しなければならない。  
(1) 住居が異動し、通勤経路に変更が生じた場合 (2) 利用交通機関の乗車料金が改定された場合 (3) 交通事情の変化により、通勤経路又は利用交通機関の変更が妥当な場合 (4) その他会社の認める事由により通勤手当を変更する必要があるが生じた場合 (支給方法) 第5条 通勤手当は、1か月全休 (年次有給休暇、慶弔休暇、特別休暇を含む) の場合は支給しない。  
2 新调入社員は初任給及び自己都合による退職の場合の当月の通勤手当は、戻りによる割増を算定する。

At the bottom left of the sidebar, there are 'Back' and 'Get Help' options.

# アプリUI | Tokenの利用状況確認機能①

The screenshot shows the Azure AI Playground interface. On the left is a dark sidebar with the title 'Azure AI Playground' and navigation options: 'Q&A Chat' and 'Keyword Search'. At the bottom of the sidebar are 'Logout' and 'Get Help' links. The main chat area has a light gray background. It shows a conversation starting with an AI message: 'こんにちは！何かお手伝いできることはありますか？' with a token count of 'Token:0 (入力:0 出力:0)'. This is followed by a user message in a blue bubble: '妻と5歳の子供がいます。支給される家族手当はいくらですか？'. The AI then responds with a list of allowance rates: '家族手当は次のようになっています： - 配偶者：月額10,000円 - 18歳未満の子供1人につき：月額5,000円 (ただし高校在学中は卒業まで)'. It concludes with 'したがって、あなたの場合、家族手当は月額15,000円になります。' and a token count of 'Token:1830 (入力:111 出力:1719)'. A red callout box with a pointer to the token count contains the text 'チャット利用時に適宜Tokenの利用料が確認可能'. At the bottom of the chat area is a text input field with the placeholder 'メッセージを入力してください' and a blue send button.

チャット利用時に適宜Tokenの利用料が確認可能

# アプリUI | Tokenの利用状況確認機能② (モニタリング)

ホーム > 1tadvopenaistorage002jpe

1tadvopenaistorage002jpe | ストレージ ブラウザー

検索

1tadvopenaistorage002jpe

お気に入り

最近表示した項目

BLOB コンテナ

ファイル共有

キュー

テーブル

openaitokens

すべて表示

テーブル > openaitokens

認証方法: アクセス キー (Azure AD のユーザー アカウントに切り替える)

フィルターを追加する

最初の 100 項目の表示

PartitionKey	RowKey	Timestamp	UserId	TotalTokens	PromptTokens	CompletionTokens
1fddc71b-77ef-4bf9-a2...	openai_063f9f57-bfd6-4...	2023-05-17T01:23:38.91...	1fddc71b-77ef-4bf9-a2...	490	299	191
1fddc71b-77ef-4bf9-a2...	openai_071b8761-c207...	2023-05-16T08:15:13.81...	1fddc71b-77ef-4bf9-a2...	156	46	110
1fddc71b-77ef-4bf9-a2...	openai_1e202499-5a11...	2023-05-16T08:05:25.52...	1fddc71b-77ef-4bf9-a2...	566	523	43
1fddc71b-77ef-4bf9-a2...	openai_21d12478-c978...	2023-05-16T08:14:33.01...	1fddc71b-77ef-4bf9-a2...	1446	1285	161
1fddc71b-77ef-4bf9-a2...	openai_2e25ff3c-a487-4...	2023-05-16T08:10:12.79...	1fddc71b-77ef-4bf9-a2...	1330	1181	149
1fddc71b-77ef-4bf9-a2...	openai_3e9531d6-abed...	2023-05-16T08:16:26.06...	1fddc71b-77ef-4bf9-a2...	675	229	446
1fddc71b-77ef-4bf9-a2...	openai_4420f0ef-8325-4...	2023-05-16T05:03:23.57...	1fddc71b-77ef-4bf9-a2...	364	327	37
1fddc71b-77ef-4bf9-a2...	openai_5baf680b-6fe3-...	2023-05-16T08:06:05.59...	1fddc71b-77ef-4bf9-a2...	945	590	355
1fddc71b-77ef-4bf9-a2...	openai_6f2d5401-061f...	2023-05-16T05:02:09.16...	1fddc71b-77ef-4bf9-a2...	241	77	164
1fddc71b-77ef-4bf9-a2...	openai_70a364fc-6981-...	2023-05-16T08:20:55.04...	1fddc71b-77ef-4bf9-a2...	1607	1122	485
1fddc71b-77ef-4bf9-a2...	openai_73ae780e-6caf-...	2023-05-16T08:18:13.02...	1fddc71b-77ef-4bf9-a2...	1092	704	388
1fddc71b-77ef-4bf9-a2...	openai_783bfb00-90a0...	2023-05-16T08:09:22.85...	1fddc71b-77ef-4bf9-a2...	1627	1494	133
1fddc71b-77ef-4bf9-a2...	openai_7d120802-61c3...	2023-05-16T08:04:08.85...	1fddc71b-77ef-4bf9-a2...	492	46	446
1fddc71b-77ef-4bf9-a2...	openai_7da13970-42f4...	2023-05-16T08:21:32.90...	1fddc71b-77ef-4bf9-a2...	236	47	189
1fddc71b-77ef-4bf9-a2...	openai_83cd2bd5-7aca...	2023-05-16T08:22:43.06...	1fddc71b-77ef-4bf9-a2...	793	263	530
1fddc71b-77ef-4bf9-a2...	openai_8fa09e82-9e71...	2023-05-16T05:04:15.08...	1fddc71b-77ef-4bf9-a2...	513	200	313
1fddc71b-77ef-4bf9-a2...	openai_9acc907f-d706...	2023-05-16T05:01:56.19...	1fddc71b-77ef-4bf9-a2...	60	36	24
1fddc71b-77ef-4bf9-a2...	openai_ab91b807-b2d2...	2023-05-16T05:02:28.12...	1fddc71b-77ef-4bf9-a2...	280	258	22

管理者はAzure OpenAI Serviceのレスポンスに含まれるToken情報を確認できる (誰が、いつ、どれぐらいTokenを利用したかモニタリング可能)

詳細フィル

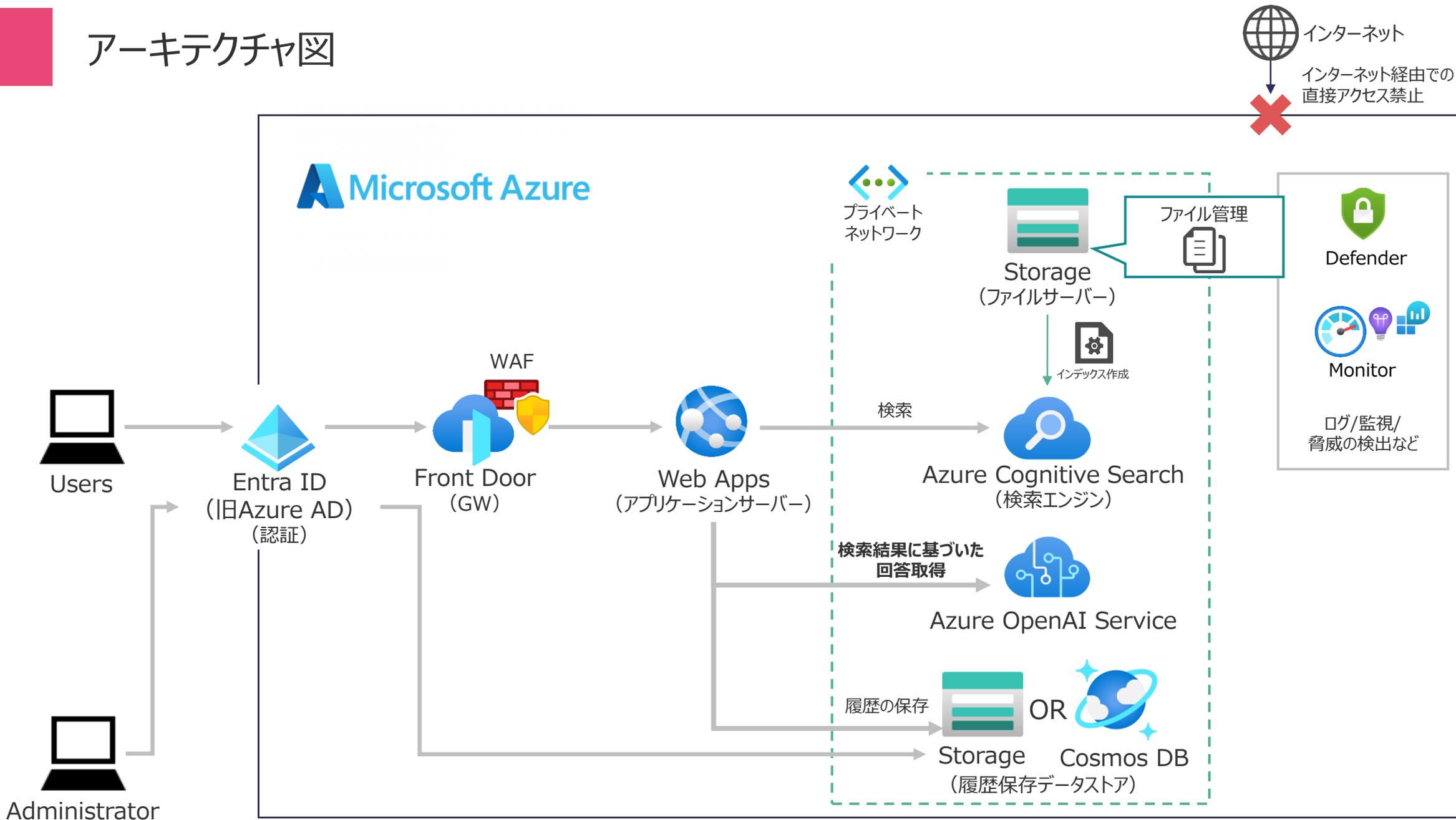


# アーキテクチャ

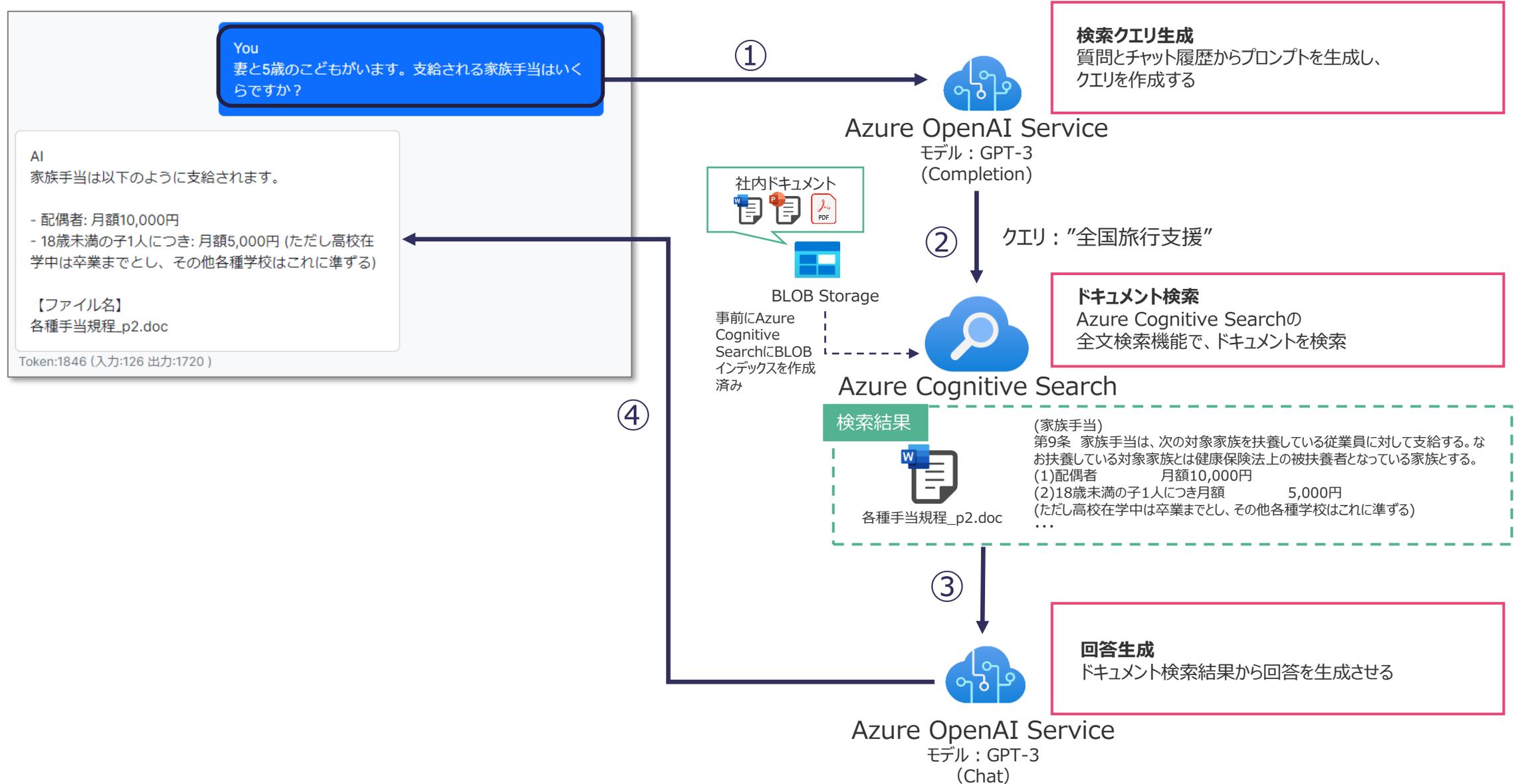


Basic

# アーキテクチャ図



# 【社内ドキュメント群へのQAチャット】のフロー





# 考慮事項

Basic

# 考慮事項 (1)

## 信頼性(可用性)

- Azureのサービスでは「可用性ゾーン」や「リージョン」といった単位で可用性を設計しており、これらを適切に組み合わせることで、ビジネスクリティカルなワークロードの信頼性を実現するように設計することが可能です。詳細は「[Azure リージョンと可用性ゾーンとは](#)」をご参照ください。
- 本シナリオで用いられている Azure App Service、Azure OpenAI、Azure Cognitive Search、Azure Cosmos DB 等のコンポーネントは”ゾーン冗長”、”リージョン冗長”、”geoレプリケーション”など高可用性のオプションや構成を利用可能です。必要となる可用性に応じて導入を検討してください。  
また今回採用しているAzure Front Doorを用いることで、複数リージョン間/複数ゾーン間でのAct-Act構成が可能です。

## 信頼性(回復性)

- アプリケーションの正常性を監視するために、Application Insights を使用すると、カスタマーエクスペリエンスや可用性に影響を及ぼすパフォーマンスの問題についてアラートを生成し、対応することができます。詳細については、「[Application Insights とは何か?](#)」を参照してください。
- 回復性に関するその他の記事については、「[信頼性の高い Azure アプリケーションを設計する](#)」を参照してください。

## セキュリティ

- セキュリティは、重要なデータやシステムの意図的な攻撃や悪用に対する保証を提供します。詳細については、「[セキュリティの重要な要素の概要](#)」を参照してください。
- Webアプリケーションの脆弱性攻撃や、DDoS攻撃に対しては、本シナリオではAzure Front DoorのWAF機能やDDoS Protection機能を用いて対策しております。
- Azure Virtual Networkを用いて、Azure上にプライベートな仮想ネットワークを構築し、Azure OpenAI Serviceや各種PaaSサービス群を上記仮想ネットワーク上に配置することで、外部インターネットからの不正利用をネットワークレベルでの通信防御で防止します。
- 本シナリオは企業内ユーザを前提としているので、Entra ID (旧Azure AD) を用いて、認証認可機能を実現しています。

## 考慮事項 (2)

### コスト最適化

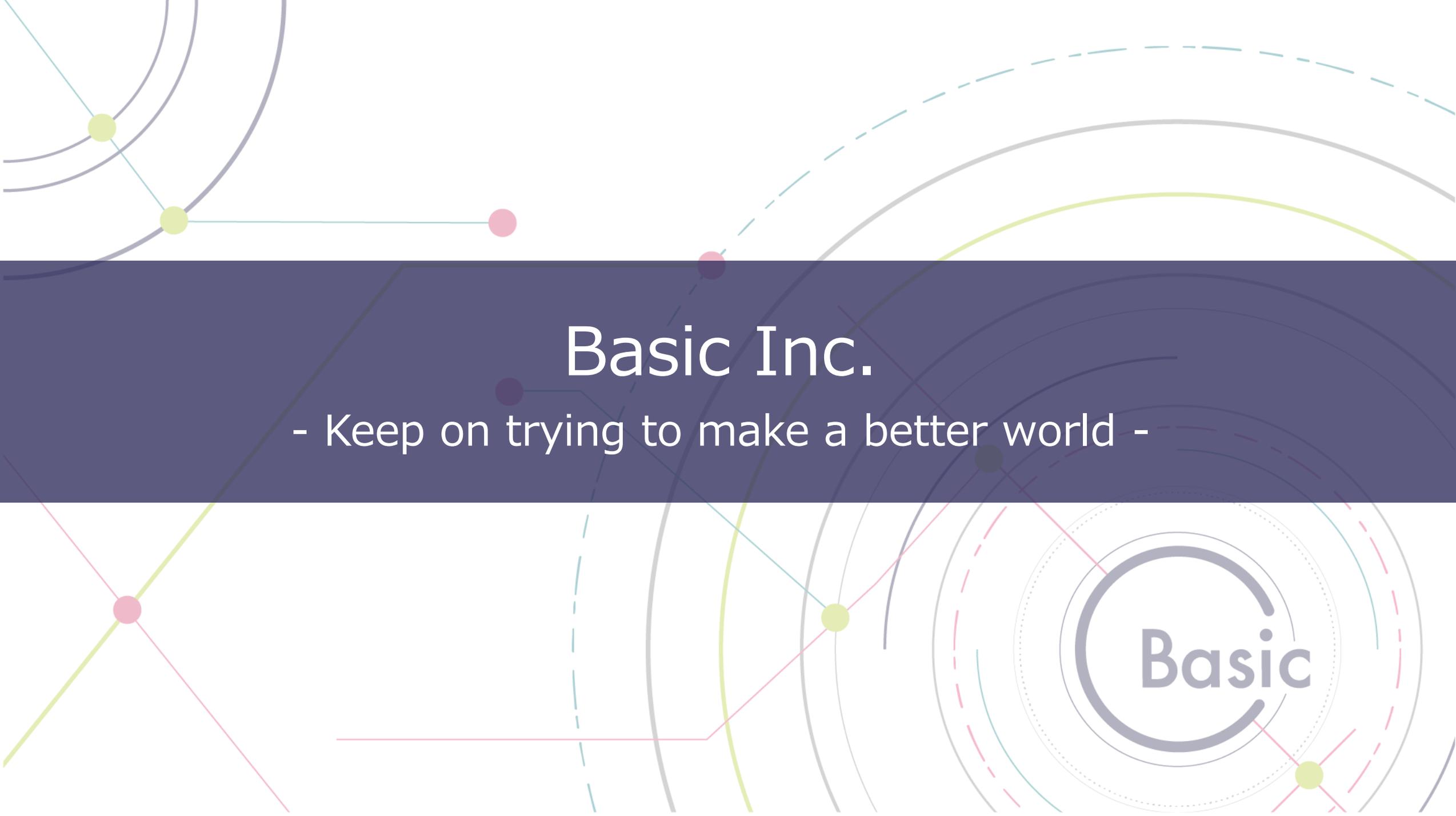
- 不要な費用を削減し、運用効率を向上させる方法を検討することです。詳しくは、「[コスト最適化の柱の概要](#)」に関する記事をご覧ください。
- チャット履歴の保存先には、利用状況に従って、Azure Cosmos DB/Azure Storage (Azure Table) を準備しており、利用状況を確認しながら、適しているサービスを選択ください。
- システムの利用料が増えた場合にも、柔軟性の高いマネージドなPaaSサービス (Azure App Service、Azure Cosmos DB) を用いることで、利用した分だけのコスト費用で最大限のパフォーマンスを発揮することができます。
- 本シナリオでは、Tokenの利用状況確認機能を用意しております。利用状況を確認しながら、必要に応じてリソース増強などの最適化を検討してください。

### オペレーショナルエクセレンス

- システムの健全性の担保、トラブルの解決、利用動向の監視を行うためには適切な監視とログ収集が必要となります。詳細は「[ワークロードの監視](#)」をご参照ください。
- Power BIやAzure App Service上にWebアプリを構築することで、Tokenの利用状況確認などの管理機能の拡充が可能です。
- ソフトウェアのアップデートや脆弱性への対応など、ソフトウェア/インフラ設計の改修を円滑に進められるよう、DevOpsプロセスを確立してください。詳細は「[リリース エンジニアリングの継続的インテグレーション](#)」をご参照ください。

### パフォーマンス効率

- アプリケーションの負荷が高まることを見越し、スケーラビリティの確保をあらかじめ検討することは重要です。詳細は「[スケーリング用のアプリケーションを設計する](#)」をご参照ください。
- Azure App Serviceは負荷に応じて水平にスケールさせることが可能です。詳細については「[自動スケーリングを有効にする方法](#)」をご参照ください。
- 頻出のクエリについてはFront DoorのCDN機能を用いる等のキャッシュ戦略もご検討ください。詳細は「[キャッシュを使用する](#)」をご参照ください。
- また、特定のユーザーにAzure OpenAIの利用が集中することを避けたい場合には、API Managementによるスロットリング導入などをご検討ください。詳細は「[Azure API Management を使用した高度な要求スロットル](#)」をご参照ください。

The background features a dark blue horizontal band across the center. Above and below this band are white areas with abstract geometric patterns. These include various colored lines (green, pink, blue, grey) that intersect and curve. There are also several colored dots (green, pink, purple, black) placed at various points. On the right side, there are concentric circles in different colors (grey, green, blue, red) and styles (solid, dashed, dotted).

# Basic Inc.

- Keep on trying to make a better world -

Basic