

2024.1.31

# AzureでRAGをガンガン試行錯誤してみても 得たナレッジを紹介します！

KDDIアジャイル開発センター株式会社(KAG)

三宅 潤也

# 自己紹介

- **三宅 潤也(@jnymyk)**

- KDDIアジャイル開発センター株式会社(**KAG**)/KDDI株式会社

- 業務：

- 生成AIチャットボットの業務活用プロジェクトのプロダクトオーナー
- 開発チームのスクラムマスター・開発者

- 趣味：

- 深夜ラジオ（オードリーのANN in 東京ドーム参戦します！）
- ニチアサ（変身ベルト、武器のおもちゃが増えてきました・・・）

- 好きな**Azureサービス**：AI Search、Cosmos DB



# コンテンツ

- **はじめに**
  - RAGとは？
  - AzureにおけるRAGの実装
- **[結論]やってみてよかったこと**
- **事例紹介**
  - デバサポAIチャット概要
  - 独自情報について
  - 試行錯誤あれこれ
  - デモ
- **まとめ：やってみてよかったこと**

# はじめに

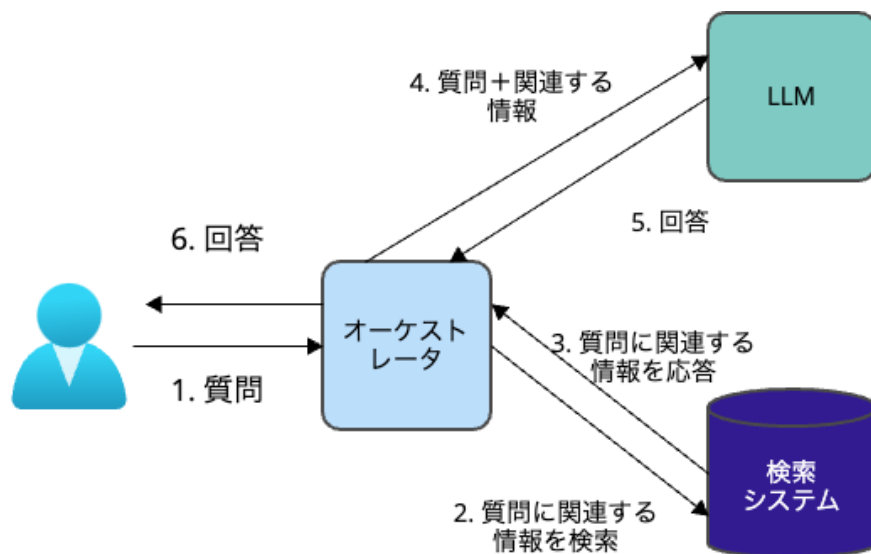
- KAG/KDDIで実施したPoCプロジェクト、**デバサポAIチャットの企画・開発を通して得たナレッジ**を共有します！！！！
- 企画から運用開始まで**3ヶ月間の短期集中プロジェクト**のためとにかくやってみよう！精神で突き進みました🔥(私は**PO/SM/Devの3足のわらじ**状態でした🔥)
- **改善アプローチのフィードバック**大歓迎です👏👏
  - チームで発案したもの、Azure SAさんからのアイデアなど、とにかくやってみよう精神に基づいたアプローチのため、論文などでアカデミックに検証しているソースがあればぜひ知りたいです！  
(手法、効果、さらなる改善アプローチなど)

# [結論]やってみてよかったこと

- データ取り込みについて（独自情報ファイルのロード～インデックス化）
  - データの読み込みについて
    - AI回答の精度を高めるため、トークン上限超過を回避するため
      - データの内容に**冗長なものやノイズになるものが含まれている場合は要約してから取り込む**
  - データ分割について
    - AIのトークン上限回避のため**チャンク分割**する
      - チャンク分割時のパラメータはchunk\_sizeの値の20%程度の値を目安にchunk\_overlapに指定すると良さそう
  - データの保存について
    - 業務活用時に想定される**キーとなるデータを埋め込んでおく**
      - 例：機種名をキーに質問することが想定される場合、メタデータに機種名を埋め込んでおく
- データ検索について
  - 業務活用時に想定される**キーとなる情報でフィルタして関連データを検索**する
  - AI Searchから**検索するドキュメント数を増やす**

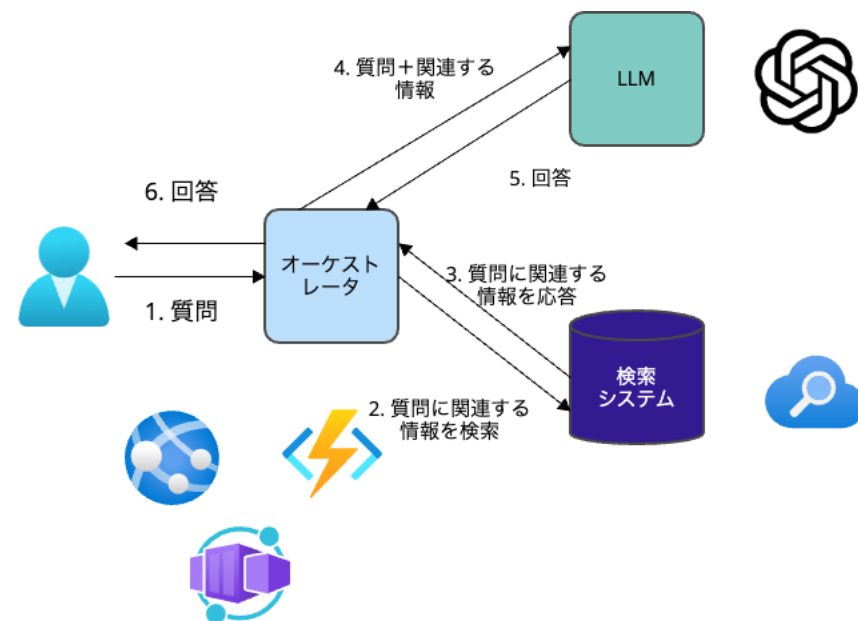
# Azureで実現するRAGとは？

- RAGとは？
  - 社内文章などの独自情報を元に大規模言語モデル(LLM)から回答させるための手法
    - 独自情報をベクトル化してDBに保存し、入力に関連した情報をDBから検索してLLMに渡すプロンプトに独自情報を付加することで実現



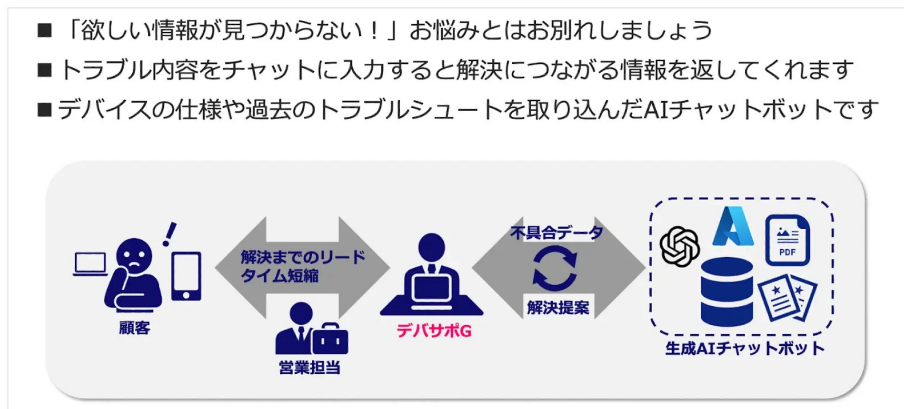
# Azureで実現するRAGとは？

- AzureにおけるRAGの実装
  - 必要なAzureリソースと構成要素
    - LLMサービス：**Azure OpenAI Service**
    - 検索サービス（独自情報の保存と検索・ベクトルストア）：**Azure AI Search**
    - オークストレーター（バックエンド）：**自前でアプリケーションを実装** or **Azure OpenAI Add your data**など
    - フロントエンド：自前でチャットUIを実装 or OSSを利用
    - オプション
      - 独自情報の格納：**Blob Storage**
      - 会話履歴の保存：**Cosmos DB**



# 事例紹介(デバサポAIチャット)

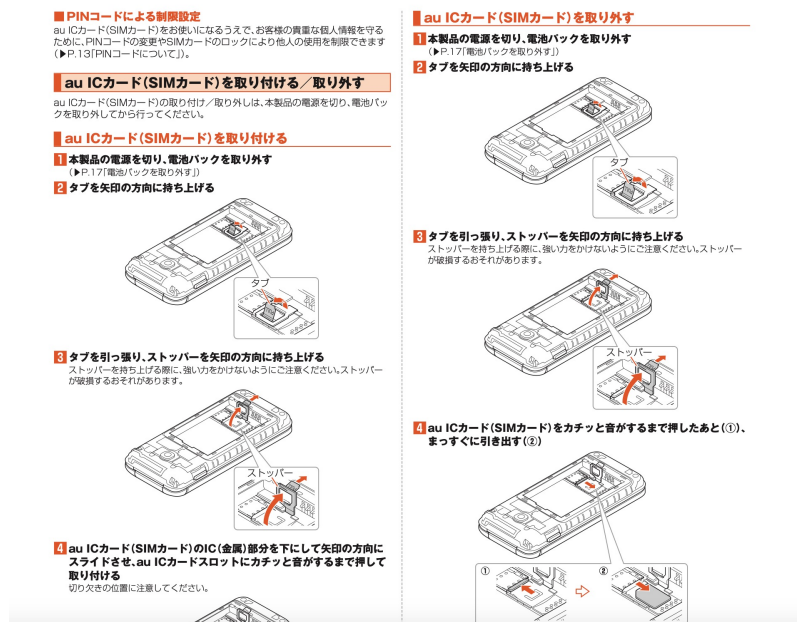
- **セキュアに社内文章データを活用**できるAIチャット
- スマートフォンやモバイルWiFiルータのデバイスの仕様や、過去のトラブルシュートなどの独自情報を取り込むことで、**過去のナレッジデータから法人のお客様からの不具合申告の解決**につながる情報をAIが回答
- **プレスリリース**： KDDIアジャイル開発センター、Azure OpenAI ServiceとAmazon Bedrockを用いて生成AIを活用したプロダクトを複数開発
  - <https://prtmes.jp/main/html/rd/p/000000009.000115171.html>
- KAGエンジニアブログ記事： Azure&RAGアーキテクチャによる生成AIチャットボット「デバサポAIチャット」を作ってみてわかったこと(法人デバイスサポート業務活用PoC)
  - <https://developers.kddi.com/blog/29moPnzUaqSPJsu0v2pKg9>





# 事例紹介(独自情報について)

- AIチャットボットに取り込んだ独自情報：
  - 法人向けスマートフォンおよびモバイルWiFiルーターの**取扱説明書**：数百ページほどのボリュームの**PDF形式のファイル**。9機種分を取り込み。
  - 法人のお客様からの**不具合申告の過去対応一覧・対応履歴・チャット履歴**：**CSV形式のファイル**。過去に対応した**不具合申告情報を約1000件**ほど取り込み。



取扱説明書PDFファイルのイメージ  
GRATINA KYF42 取扱説明書[詳細版]より引用

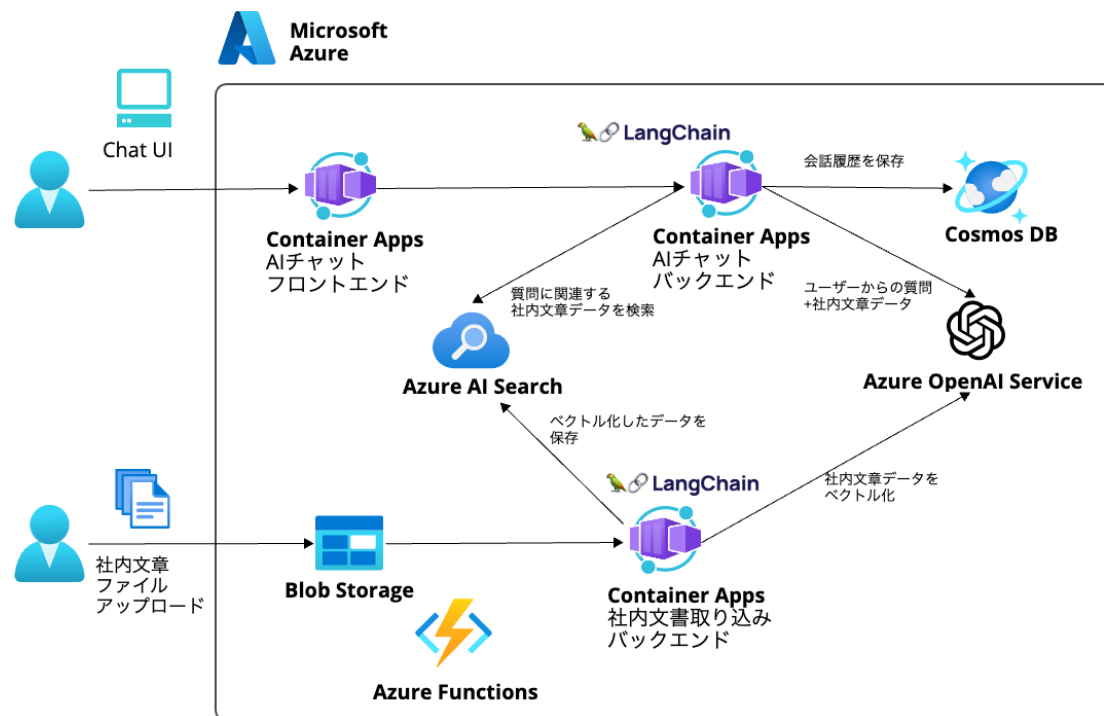
管理番号	送信欄	タイトル	ご相談内容	発生届未状況	発生場所・時	発生場所 (住)	再現性の有無	再現頻度 (確)	メールフォル	不達遅延の状	不達遅延の状
MF2307006	2023/8/25 6:28,キャリア 七部,仮営業1部 営業 八部さま, 営業 太郎さま おつかれさまです。 仮チャネル営業部 仮サポートG キャリア 七部です。 ご連絡いただき、誠にありがとうございます。 内容承知しました。 ※毎月お盆期間中は受付が滞ります。 2023/8/23 4:18,キャリア 八部,仮営業 次郎さん	端末 マンゴ その他 端末未?		?	?	?	?	?	?	?	?
MF2302014	先日はTeamsチャット、通話でのご連絡にてご対応いただきありがとうございます。 上記の「連絡履歴」タブのMD(履歴) / 発言者内容添付ファイル名 2023/7/28 8:52,キャリア 次部,営業 三郎さん 承知しました。では一旦今時点でクローズしますね。 何かあればまたご連絡ください。	端末 バイブ その他 端末未?		?	?	?	?	?	?	?	?
MF2306002	★以下、任意で構いませんので、ご確認いただけたらと思います。 ----- AARON-3 ご利用ありがとうございます。 移動端末不具合向けに専用チャットボットをご用意しております。	端末 キウイ その他 端末未?		?	?	?	?	?	?	?	?

過去対応のCSVファイルのイメージ  
(内容はダミーのサンプルファイル)

# 事例紹介(アーキテクチャ)

- オークストレーターをContainer Apps上に実装
- PoCということで社内文章ファイルのアップロードUIは省略しAzureポータルからBlob Storageに直接アップロード

## デバサポAIチャット アーキテクチャ



# 事例紹介(デモ)

+ 新規チャット

Search...

チャット履歴

New Conversation

録音したものが再生できな...

録音したものが再生できな...

[管理番号]MF23020011 問...

[管理番号]MF23020014 問...

通話中に突然再起動する事...

通話中に突然再起動する事...

[機種名]kyf42 SIMカードの...

New Conversation

フォルダを作成する

会話をクリア

設定

## Device Support G Chatbot UI

デバイスの仕様や、過去の対応事例をお答えします。

入力ボックスに以下のフォーマットで質問を記載してください。

**対象機種を絞って質問をする場合**

1行目 : [機種名]xxx 例) [機種名]kyf42

2行目以降 : 質問を記載 例) 着信時に勝手にスピーカーになる現象はありますか?ある場合は管理番号・概要・対処法を教えてください。

**対象管理番号を絞って質問をする場合**

1行目 : [管理番号]xxx 例) [管理番号]MF23020014

2行目以降 : 質問を記載 例) 問い合わせの概要と対処法を教えてください。

**対象機種や対象管理番号を絞らずに質問をする場合**

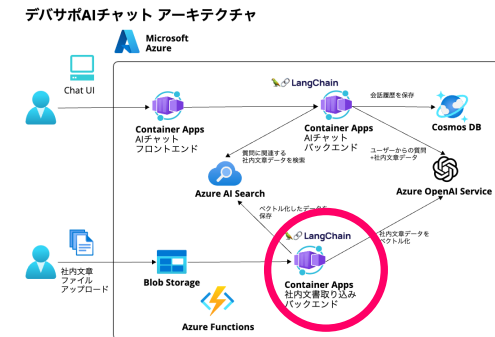
1行目 : 質問を記載 例) 着信時に電話に出られない事象はありますか?ある場合は機種名と概要と対処法を教えてください。

質問を入力してください。Shift+Enterで改行できます。

AIが生成した文章のため不正確な場合があります。生成結果についてフィードバックにご協力をお願いします。

# 事例紹介（試行錯誤のあれこれ①）

- データの取り込みについて
  - AIからの回答精度を高める、トークン上限超過を回避することを目的に**過去対応履歴CSVファイル**のうち、**チャットに関する情報の要約**を実施。
  - 独自情報ファイルを読み込み際に以下観点で要約をしてからチャンク分割。
    - チャットのやりとりで**議論された問題や課題**は何か？
    - それぞれの**問題や課題**に対して、**提案された解決策や調査**などあるか？
    - 解決案や調査に対する**結果および行動**などを記述
    - 「お疲れ様です、XXです」ような**冗長なあいさつ表現を削除**



MF23020011 年月日時分.発着番号.内容.添付ファイル名  
2023/3/2 0:25.キャリア 三郎.営業 九郎さんへ

キャリア 三郎です。  
プロ品 太郎さんからの回答で、落ち着いたと思いますので、クローズします。

桜の花株式会社の回答  
■問題発生原因MDM設定端末のアップデート時に、自動録音機能のユーザーデータ読取りミスが発生し、通話時の録音で音が入らない問題が発生。(ユーザーデータを読み取ってからの録音が開始される弊社の再現試験では発生しないため、work profile(MDM)による影響が考えられます。  
■対応ユーザーデータの修正は、次期MR(ソフトウェアのアップデート)で適用される予定です。※次期MR日程は未定となっております。召喚が完了次第、回答いたします。  
■回避策 端末の再起動実施・端末再起動時に、ユーザーデータ読取りが行われ、設定されるため、問題が解決されます。※念のため、再起動後、自動録音を確認をお願い致します。

今回のサポート品質はいかがでしたでしょうか。  
品質向上のため、簡単なアンケートにご協力いただけますと幸いです。  
(最大5問：所要時間2分程度)  
<https://karikari.nankai.co.jp/mf23020011/>  
2023/2/10 7:00.市内 静雄.営業 九郎さん

お疲れ様です。仮サポートGのキャリア 七郎です。

先ほどは端末の受け入れ対応を頂きありがとうございました。  
引き続きよろしくお願いいたします。

2023/2/10 4:30.営業 九郎.キャリア 七郎さん  
お疲れ様です。仮営業3部 営業 九郎です。

先日キャリア 三郎さんからお話ありましたが、桜の花株式会社からお客様の端末戻ってきてますでしょうか？

2023/2/3 1:02.営業 九郎.キャリア 三郎さん  
お疲れ様です。仮営業3部 営業 九郎です。

チャットでお伝えした件、確認漏れとして記載します。  
事後再現端末追加  
回線番号：0856789012A  
ユーザー、ユーザー 三郎 きぶろう (担当者)；  
2023/2/2 7:26.キャリア 三郎.キャリア 三郎です。  
スタッフホールディングス株式会社からの申告をまとめています。  
録音されるはずの音声データが、m4aのファイルに入っていないとの事です。  
確認中次第で発表させていただきます。せっかく、実施する端末があるので桜の花株式会社にお問い合わせし、\*KARikayu 01録音できないpptx

要約前

MF23020011 問題や課題：  
- MDM設定端末のアップデート時に、自動録音機能のユーザーデータ読取りミスが発生し、通話時の録音で音が入らない問題が発生している。  
- 桜の花株式会社の再現試験では問題が発生しないため、work profile(MDM)による影響が考えられる。  
- 次期MR(ソフトウェアのアップデート)でユーザーデータ読取りの修正が適用される予定だが、次期MRの日程は未定。  
- 回避策として、端末の再起動を行うことで問題が解決する可能性がある。

提案された解決策や調査：  
- 次期MRでのユーザーデータ読取りの修正を待つ。  
- 端末の再起動を行い、問題が解決するか確認する。

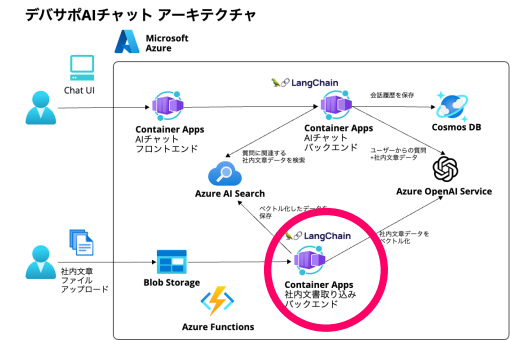
結果と行動：  
- 次期MRの日程は未定のため、具体的な解決案は提案されていない。  
- 端末の再起動を行い、問題が解決するか確認するように指示されている。  
- サポート品質向上のためのアンケートへの参加を依頼されている。

要約後

# 事例紹介（試行錯誤のあれこれ②）

- データの分割について

- AOAIのトークン上限回避のため**チャンク分割**を実施。
- LangChainのRecursiveCharacterTextSplitterを利用してchunk\_size=1000を固定として、chunk\_overlap=0, 100, 200(デフォルト), 500で実施。
- リファレンスとなる5パターンの質問への回答を比較した結果以下の結果となった。
  - chunk\_overlapの値によって劇的に精度が良くなることはなかった。**
  - 質問のパターンによって、200, 500それぞれで良い回答を返していた。
  - chunk\_overlapの値が大きくなるほどチャンク数が増えコストがかかるため、**chunk\_overlap=200を採用。**



## パターン3

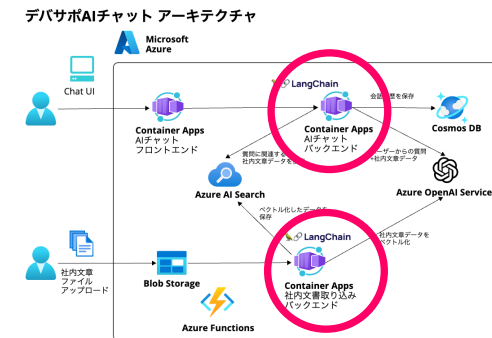
着信時に電話に出られない事象はありますか?ある場合は機種名と概要と対処法を教えてください。

- chunk\_overlap=0
  - 概要は正しい。管理番号・対応経緯・対処法はいまいち。
- chunk\_overlap=100
  - 対応経緯の精度が上がり、問題の原因を答えられた。管理番号は間違っている。
- chunk\_overlap=200
  - 概要をより詳細に答えてくれている。ソフトウェアアップデートで修正されることを教えてくれた。管理番号は間違っている。
- chunk\_overlap=500
  - 概要・対応経緯・対処法を答えられた。問題の原因とソフトウェアアップデートで修正されることも回答できた。管理番号は間違っていた。

ある質問パターンと各chunk\_overlapごとの回答

# 事例紹介（試行錯誤のあれこれ③）

- 複数機種の取り扱い説明書や過去対応履歴をAI Searchに取り込んだ際に、入力と異なる機種についての情報を検索して回答してしまう事象が発生
- 特定の情報からのみ検索するような仕組みを実装
  - 検索システムへのデータの登録について
    - 業務活用時に想定されるキーとなるデータをAI Searchにインデックス化するときに埋め込んでおく
      - 例：スマホの機種名をキーに質問することが想定される場合、AI Searchのメタデータに機種名を埋め込んでおく
  - 検索システムからのデータの検索について
    - 業務活用時に想定されるキーとなる情報でフィルタして関連データを検索する
      - 例：スマホの機種名をキーとして検索対象ドキュメントをフィルタして対象を絞って上で検索する



# 事例紹介（試行錯誤のあれこれ④）

- データの検索について

- AI Searchから検索するドキュメント数を増やしてみた

- k=3, 6, 10(GPT-4-32利用。トークン長上限に近い設定)に変更した際にリファレンスとなる質問に対する回答を比較
- k=6では3よりも詳細に回答できていた。k=10の場合、回答できない場合があったり6の場合と回答に大きな差分がなかったため、**k=6**を採用

Q:

[機種名]KARI01

着信時に勝手にスピーカーになる現象はありますか?ある場合は管理番号・概要・対処法を教えてください。

## k=3の回答

はい、着信時に勝手にスピーカーになる現象が報告されています。管理番号はMF23020014で、概要は「端末 パイナップルコンサルティング株式会社（KARI01／着信時のスピーカー挙動不良）」です。対処法については、現在のところ具体的な情報は提供されていません。

参照データ：[サンプル\_AARON-3]F\_area\_20230901152437\_20230906\_通信欄追加.csv

管理番号：MF23050061, MF23020014

## k=6の回答

はい、着信時に勝手にスピーカーになる現象があります。管理番号はMF23020014で、概要は「端末 パイナップルコンサルティング株式会社（KARI01／着信時のスピーカー挙動不良）」です。対処法については、現在のところ解決策や具体的な結果は記載されていません。ただし、端末の再起動を行うことで一時的に問題が解消する可能性があります。詳細な対処法や解決策については、お客様の要望として他の機種への変更を希望していることが記載されていますので、それに基づいて対応が行われる可能性があります。

参照データ：[サンプル\_AARON-3]F\_area\_20230901152437\_20230906\_通信欄追加.csv

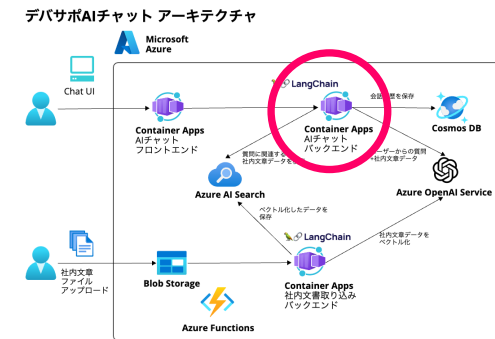
管理番号：MF23070062, MF23020011, MF23020014, MF23050061

## k=10の回答

申し訳ありませんが、提供された情報には着信時に勝手にスピーカーになる現象に関する具体的な情報が含まれていません。そのため、該当する管理番号や概要、対処法についてはわかりません。お客様の要望としては、他の機種への変更希望があるようですので、メーカーやキャリアに問い合わせ、他の機種への変更の可否や手続きについて確認することをおすすめします。

参照データ：[サンプル\_AARON-3]F\_area\_20230901152437\_20230906\_通信欄追加.csv

管理番号：MF23070062, MF23020011, MF23060002, MF23050061, MF23020014



# [結論]やってみてよかったこと

- データ取り込みについて（独自情報ファイルのロード～インデックス化）
  - データの読み込みについて
    - AI回答の精度を高めるため、トークン上限超過を回避するため
      - データの内容に**冗長なものやノイズになるものが含まれている場合は要約してから取り込む**
  - データ分割について
    - AIのトークン上限回避のため**チャンク分割**する
      - チャンク分割時のパラメータはchunk\_sizeの値の20%程度の値を目安にchunk\_overlapに指定すると良さそう
  - データの保存について
    - 業務活用時に想定される**キーとなるデータを埋め込んでおく**
      - 例：機種名をキーに質問することが想定される場合、メタデータに機種名を埋め込んでおく
- データ検索について
  - 業務活用時に想定される**キーとなる情報でフィルタして関連データを検索**する
  - AI Searchから**検索するドキュメント数を増やす**



# 技術以外のふりかえり

- For 生成AIを活用して誰かの業務課題を解決したい方（主に**エンジニア、デザイナー**など）
  - 要求される**回答の精度が高くない用途**から適用していきましょう！
    - 最初から完全自動化、省力化による社外ユーザーの利用を想定しているプロジェクトはがんじがらめになりやすく、まず作って効果を確認してみよう！の流れになりにくい
  - **不確実性の高いプロダクト**なので、**アジャイル開発**との相性が非常に良い👍
    - 事前に仕様を固めて、全体の詳細な設計を詰めて・・・はマッチしにくい
  - プロダクトバックログアイテムの**見積もりはどんぶり勘定でもOK!** とにかくやってみよう！
    - あれこれ机上で検討・議論するよりも**動くものが正義**

# 技術以外のふりかえり

- For 生成AIで解決したい業務課題をお持ちの方（**営業、企画などエンジニア以外の方**）
  - **ChatGPT**やパブリッククラウド各社が提供している**マネージドサービス**などで**ライトに試して**感触を掴んでみましょう！
    - **生成AIやRAGで何ができるか**、ざっくりでもイメージが掴めているとエンジニアとの意思疎通が非常にやりやすいです。
  - 柔軟で動きが速く、小回りがきく社内・社外パートナーと組みましょう！
    - **試行錯誤とフィードバックのサイクル**をたくさん回しましょう



Be a Change Leader.

アジャイルに力を与え

共に成長し続ける社会を創る