

顧客向け AIチャットボットの基礎

RAGと構築・運用のポイント



AIチャットボットとは？

従来のチャットボットとの違いとRAGの概要

従来型 vs AIチャットボット

従来のチャットボット

シナリオ型や辞書型が主流。

- ✕ 決められた質問にしか回答できない
- ✕ 柔軟な対話や文脈の理解が苦手
- ✕ FAQや定型的な案内に利用

AIチャットボット (RAG)

LLMとRAG技術を活用。

- ✓ 文脈を理解し、自然な対話が可能
- ✓ 自社ナレッジに基づき回答を動的生成
- ✓ 複雑な問い合わせや情報検索に対応

中核技術: RAGとは？

-  RAG = 検索拡張生成 (Retrieval-Augmented Generation)。
-  従来のLLMの弱点(ハルシネーションや知識の古さ)を克服する技術です。
-  自社のドキュメントなど、外部ナレッジベースから関連情報を「**検索**」します。
-  検索結果をプロンプトに含め、それに基づいて正確な回答を「**生成**」します。

RAGの仕組み(簡易フロー)



1. 検索 (Retrieval)

ユーザーの質問に関連する情報を、ベクトルデータベースから検索します。



2. 拡張 (Augmentation)

見つかった情報を元の質問(プロンプト)と組み合わせ、LLMに渡します。



3. 生成 (Generation)

LLMが提供された情報に基づいて、文脈に沿った正確な回答を生成します。

AIチャットボットの構築

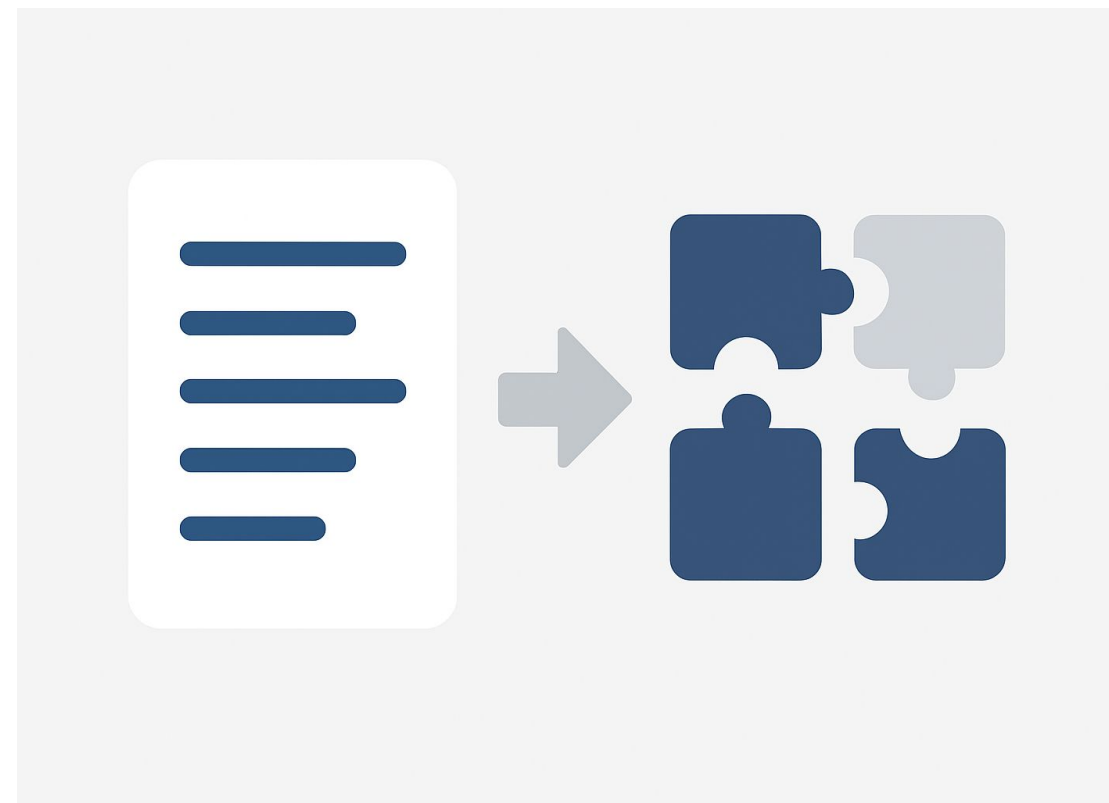
ナレッジベースを作成する3つのステップ

ステップ1: チャンキング (Chunking)

ドキュメントの分割

AIが情報を検索・処理しやすくするために、元のドキュメント（PDF、HTML、テキストファイルなど）を小さな「チャンク」（塊）に分割します。

適切なサイズに分割することが、後の検索精度に大きく影響します。



チャンキングのポイント

なぜ分割サイズが重要か？

AIが回答を生成するためには、適切な文脈(情報)が必要です。

- ・**小さすぎる場合** : 文脈が失われ、質問に答えるための情報が不足します。
- ・**大きすぎる場合** : ノイズ(不要な情報)が多くなり、検索精度が低下します。

主な分割アプローチ

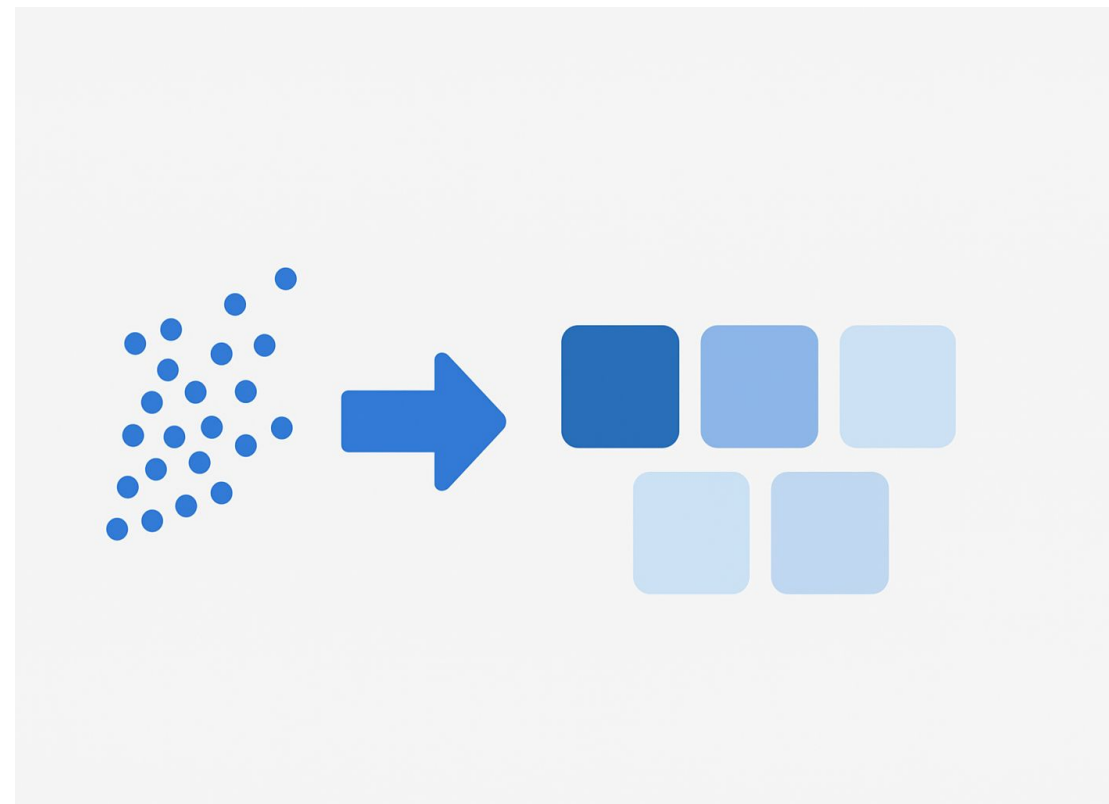
- ・**固定サイズ** : シンプルに文字数やトークン数で区切ります。実装が容易です。
- ・**意味論的分割** : 文書構造(段落、見出し)や文の意味に基づいて分割し、情報のまとまりを維持します。

ステップ2: Embedding (埋め込み)

テキストのベクトル化

分割した各チャンクを、「Embeddingモデル」を使って「ベクトル」(数値の配列)に変換します。

このベクトルは、テキストの「意味」や「文脈」を数値的に表現したもので、AIが単語の表面的な一致ではなく、意味の近さで検索できるようになります。



Embeddingのポイント

単語の一致 vs 意味の検索

従来の検索は「単語の一致」に基づきます（例：「AI 導入」と「AI 活用」は別物）。

Embeddingによるベクトル検索は、「**意味の近さ**」に基づきます。これにより、「AIを導入したい」という質問に対し「AIの活用事例」という情報を見つけることができます。

意味の「距離」で判断

Embeddingモデルは、単語や文を「ベクトル空間」上の点として配置します。

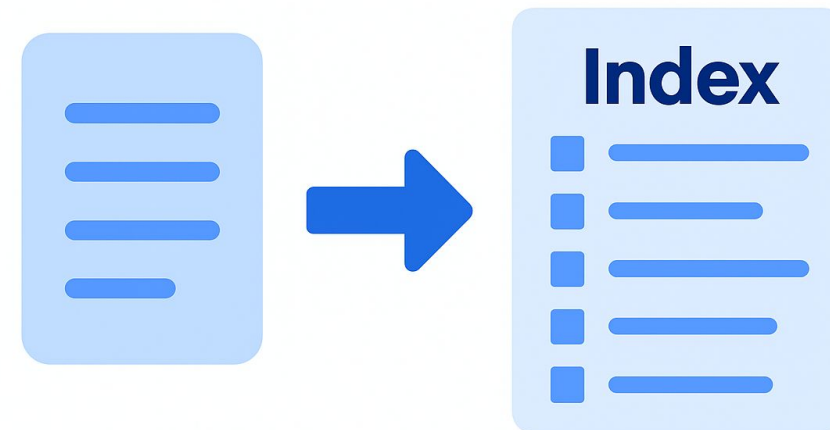
意味が近い単語（例：「犬」「猫」）は空間上で**近くに**、意味が遠い単語（例：「犬」「車」）は**遠くに**配置され、この「距離」に基づいて関連性を判断します。

ステップ3: 索引作成 (Indexing)

ベクトルデータベースへの保存

生成されたベクトルデータを「ベクトルデータベース」に保存し、インデックスを作成します。

これにより、ユーザーの質問(もベクトル化される)と意味的に最も近い情報を高速に検索することが可能になります。これがAIの「知識の源」となります。



日々の運用

チャットボットの鮮度と精度を保つために

運用タスク (1): 監視と更新



1. 日々の運用作業

ユーザーログを監視し、AIが回答できなかった質問や、回答が不適切だったケースを特定します。



2. ドキュメント更新時

新製品情報、仕様変更、料金改定などが発生した際、速やかにナレッジベースを更新（追加・修正・削除）します。



3. 表現のバリエーション

回答が単調にならないよう、プロンプトを調整します。同じ質問でも異なる「言い回し」に対応できるかテストします。

運用タスク (2): 品質と調整



4. 毎週の品質確認

主要な質問や、ログで特定された「回答が怪しい質問」をリスト化し、週次で回答の精度をチェックします。



5. 見せ方の調整

回答が長すぎ/短すぎないか、箇条書きを使うべきかなど、プロンプトを調整し、回答の「見せ方」を最適化します。



6. 例外と本人確認

クレームや個人情報を含む質問、本人確認が必要な手続きなど、AIが対応すべきでない質問を定義し、有人対応へ誘導します。

運用タスク (3): 棚卸しと継続

7. 月次の棚卸し

データベース内の情報が全体として最新かを確認します。古いキャンペーン情報や、廃止された製品情報など、不要なナレッジが「ノイズ」として残っていないかを月次でレビューし、削除します。

品質維持のサイクル

チャットボットの価値は「作って終わり」ではありません。「**ログ確認** → **DB更新** → **品質確認** → **プロンプト調整**」というPDCAサイクルを継続的に回し続けることが、精度維持の鍵となります。

- ✓ 顧客向けAIチャットボットの成功には、**RAG (検索拡張生成)** の仕組みが鍵となります。
- ✓ 高品質な知識DBの構築には、「**チャンキング**」「**Embedding**」「**索引作成**」の3ステップが重要です。
- ✓ AIの回答精度を維持するため、ログ監視から棚卸しまで、**継続的な運用サイクル** が不可欠です。

会社概要

社名	株式会社ヴィクセス
所在地	神奈川県川崎市中原区木月 1-32-3 2F
代表者名	林 高行
設立	2011年10月11日
資本金	12,452,000円
業務内容	DX顧問、DX顧問バンク(現: エクセレント・パートナーズ)



林 高行 (株式会社ヴィクセス 代表)

経歴

東京工業大学大学院修了後、Citibank, N.A.を経てみずほ証券リスク統括部にてデリバティブのクオンツ業務に従事後、2011年にヴィクセスを立ち上げました。

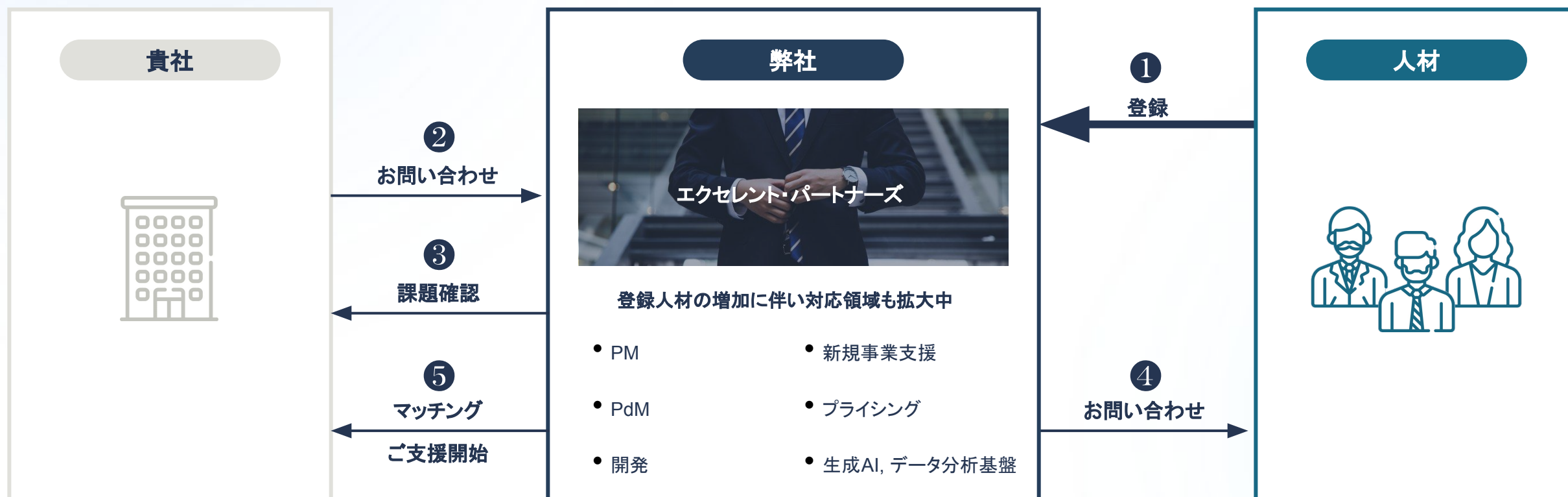
強み

- 金融工学
- SaaS、特にクラウド BI開発・運用の経験
- BtoB SaaS 支援の経験

エクセレント・パートナーズとは

エクセレント・パートナーズとは、上位0.1%人材を、お客様のお悩みに合わせアサイン・課題解決するサービスです。
課題解決まで人材が伴走し、貴社の事業成長を御支援致します。

エクセレント・パートナーズとはあなたの会社の支援する人材を紹介するプラットフォームです



エクセレント・パートナーズの強み

高い技術力はもちろんのこと、高いコミュニケーション能力、他者への強いサポート姿勢、能動的な提案、自走能力、常に安定した精神、チームワークなど、組織の成長を阻害せず、むしろ高める人材が、貴社の組織にて御支援を致します。

上位0.1%の人材

国立大学、慶應義塾大学、早稲田大学などの一流大学出身者、もしくは独自のコネクションによって集められた人材のみで構成された精鋭チーム。卓越した問題解決能力と論理的思考力を備えています。英語でのコミュニケーションも可能な人材がおります。

卓越した技術力

エンジニアリング、AI開発、プロジェクトマネジメント、プロダクトマネジメント、新規事業開発など、幅広い専門分野をカバー。あらゆるニーズに対応可能です。

円熟したメンタリティ

技術力だけでなく、チームワーク力、能動的な提案力、自走する力を兼ね備えた人材のみを厳選。プロジェクトの成功に必要な総合力を発揮します。



お問い合わせはこちらから

<https://excellent-partners.io/>

株式会社 ヴイクセス ・ 所在地：神奈川県川崎市中原区木月 1 - 32 - 3 2F