Knowledge Index 公式ユーザーマニュアル

Knowledge Index サポートチーム

2026-04-04

目次

はじめに

本書は、RAG AI チャットボットプラットフォーム「Knowledge Index」の公式ユーザーマニュアルです。システムへの初期設定から、効果的なコンテンツの登録・管理、回答精度の向上策、日々の運用や利用状況の分析まで、Knowledge Index を最大限に活用するための手順を網羅的に解説しています。

本書の対象読者

- **テナント管理者**: 組織内の Knowledge Index 環境を設定し、コンテンツを管理・運用する方（主に第 2 章〜第 9 章、第 15 章）
- **エンドユーザー**: Web サイト等に組み込まれたチャットボット（Widget）を利用して質問をする方（主に第 3 章）
- **システム管理者・運用担当者**: プラットフォーム全体の保守運用、アクセス権限管理、インフラ構成の把握を行う方（主に第 10 章〜第 13 章）

バージョン情報

- 対象バージョン: Knowledge Index v1.0〜
- 最終更新: 2025 年

第 1 章：Knowledge Index とは

1.1 サービス概要

Knowledge Index は、企業や組織が持つ独自のドキュメント資産（社内規程、マニュアル、製品仕様書、FAQ など）を最大限に活用するための **エンタープライズ向け RAG AI チャットボットプラットフォーム** です。

RAG（Retrieval-Augmented Generation = 検索拡張生成）技術を用いることで、一般的な AI が持つ「一般的な知識」だけでなく、「あなたの組織だけの固有の知識」に基づいた正確な回答を生成します。

1.2 主な特徴

1. **簡単なコンテンツ登録**: Word、Excel、PDF、Markdown、HTML、CSV などをアップロードするか、URL を指定するだけで、自動的にテキストを抽出し、AI が検索可能なベクトルデータ（埋め込み）に変換します。
2. **マルチ LLM 対応**: OpenAI（GPT-4o 等）、Anthropic（Claude 3.5 Sonnet 等）、Google（Gemini 1.5 Pro 等）の最新 AI モデルを切り替えて利用できます。用途や予算に応じた最適なモデル選択が可能です。
3. **柔軟な Widget 連携**: 数行のスクリプトを Web サイトに埋め込むだけで、チャットボットを導入できます。Ciel（公式キャラクター）など複数のアイコンや、用途別デザインテンプレートを備えています。
4. **テナント分離による高い安全性**: 組織（テナント）ごとにデータは空間的・論理的に分離されており、他の企業のデータと混ざることはありません。

1.3 料金プランの概要

Knowledge Index は利用規模に応じた 3 つのプランを提供しています。すべてのプランは「基本料金」＋「AI のトークン利用に応じた従量課金」で構成されます。

プラン	基本料金 (月額)	推奨用途	主な利用制限・特徴
ベーシック	2,980 円	小規模チーム、お試し導入	・管理/運用者数: 20 人・ナレッジ・ストレージ: 200MByte・最大質問数: 2000 回/月・Widget: AI アシスタント
スタンダード	4,800 円	一般企業、部門導入	・管理/運用者数: 50 人・ナレッジ・ストレージ: 500MByte・最大質問数: 5000 回/月・Widget: AI アシスタント＋ FAQ 等

プラン	基本料金 (月額)	推奨用途	主な利用制限・特徴
プロ	9,800 円	全社導入、大規模運用	・管理/運用者数: 100 人・ナレッジ・ストレージ: 1000MByte・最大質問数: 10000 回/月・Widget: スタンダード＋社内ヘルプデスク等・音声対応

※ ご登録後最初の 14 日間は「Knowledge Index 月額利用料」が無料となるトライアル期間です。ただし、実際に生成 AI を利用する際にかかる「API 利用料（従量課金）」につきましては、無料期間中であっても別途発生いたします。

1.4 システム構成の特長

本システムは、高いセキュリティと可用性が求められるエンタープライズ環境向けに設計され、クラウドプラットフォーム **Microsoft Azure の国内リージョン（東日本）**で稼働しています。データが海外に持ち出されることはありません。

第 2 章：はじめに —導入・初期設定

Knowledge Index を利用開始するための第一歩として、テナント（組織）の登録から初期設定を行い、チャットボットが正しく応答できるようになるまでの手順を解説します。

2.1 テナントの利用登録とログイン

Knowledge Index では、組織ごとに「テナント」という単位でデータが完全に分離・保護されます。

テナント登録の手順

1. ブラウザで Knowledge Index のログイン画面にアクセスします。
2.「利用登録」タブをクリックします。
3. 以下の情報を入力します：
   - **テナント名**：会社名や部署名（例：株式会社サンプル）
   - **テナント識別子**：システム内で一意となる ID（英数字・ハイフンのみ。例：company-hr）
   - **管理者メールアドレス**：あなた（初回管理者）のメールアドレス
   - **管理者ユーザー名**：あなたの表示名
   - **パスワード**：8 文字以上（大文字・小文字・数字をすべて含むこと）
4.「利用登録」をクリックします。
5. 入力したメールアドレス宛に確認メールが送信されます。メール内のリンクをクリックするとアカウントが有効化されます。

注意: テナント識別子は登録後に変更できませんので、わかりやすい文字列を設定してください。

ログイン

1. アカウント有効化後、ログイン画面からメールアドレスとパスワードを入力します。
2.「ログイン」をクリックすると、管理画面のダッシュボードが表示されます。

2.2 ダッシュボードの見方

ログイン後に最初に表示されるのがダッシュボードです。

- **ユーザー情報カード**: ご自身の権限や所属テナントが表示されます。
- **システム設定状況カード**: AI を利用するための準備が整っているか（準備完了 / 未完了）が表示されます。**初回ログイン時は必ず未完了**になっています。
- **統計カード**: 登録済みのコンテンツ数や、今までの質問回数などが一目でわかります。

2.3 テナントの初期設定

チャットボットを動かすには、「どの AI を使うか」と「誰の API キーを使うか」を設定する必要があります。

1. ダッシュボードの「テナント設定を開く」ボタン、または左メニューの「初期設定」をクリックします。
2. 以下の 2 つを設定します。

ステップ 1：LLM モデルの選択

AI の「頭脳」となるモデルを選択します。用途やコストに合わせて選べます。

- **チャット用モデル**: ユーザーの質問に回答する AI です。（例：OpenAI の `gpt-4o` や Anthropic の `claude-3-5-sonnet` など）
- **ベクトル埋め込みモデル**: ドキュメントを検索可能なデータに変換する AI です。（例：OpenAI の `text-embedding-3-small` など）

選択して「保存」をクリックします。

ステップ 2：API キーの登録（任意）

Knowledge Index ではシステム側でデフォルトの API キー（OpenAI / Anthropic / Google 等）を用意しているため、特別な設定を行わなくてもすぐに AI をご利用いただけます。その場合は、システムが集計した「API 利用料（従量課金単価）」が毎月のサブスクリプション更新時に加算されます。

もし **お客様ご自身で契約・取得した生成 AI の API キー**をお持ちの場合は、ここで登録することで、システム側からの「API 利用料」の請求を免除することができます。

1.（自前キーを持参する場合）画面下部の「API キー」セクションで「新規登録」をクリックします。
2. **プロバイダー**（OpenAI など）、**モデル**、およびお持ちの **API キー** を入力して保存します。

3. リストに追加された API キーの「有効化」スイッチがオンになっていることを確認します。

   ポイント: システム共通のモデルをそのまま利用される場合は、このステップはスキップして構いません。ご自身の API を利用して課金を免除させたい場合は、設定したプロバイダーとその「有効化」を忘れずに行ってください。

2.4 最初のコンテンツ登録と動作確認（チュートリアル）

設定が完了したら、実際にコンテンツ（社内文書など）を読み込ませて AI に質問してみましょう。

1. ドキュメントを用意する

テスト用に、自社の会社概要や簡単な FAQ が書かれた PDF またはテキストファイル（`test.txt` など）を用意します。

2. コンテンツをアップロードする

1. 左メニューの「コンテンツ管理」をクリックします。
2. 右上の「新規コンテンツ」をクリックします。
3. 準備したファイルを画面上にドラッグ＆ドロップします。
4. タイトルに「テスト用会社概要」と入力し、「保存」をクリックします。
5. 一覧画面に戻ります。ステータスが `PROCESSING`（処理中）から `INDEXED`（インデックス済み）に変わるまで数分待ちます。

3. チャットで質問してみる

1. 左メニューの「チャット」をクリックします。
2. 画面下部の入力欄から、アップロードしたファイルに書かれている内容について質問してみましょう。
   - 例：「会社設立日はいつですか？」
3. 数秒後に AI が回答を生成し、参照したソース（ファイル名）が回答の下に表示されれば成功です！

   ヒント: もし AI が「情報が見つかりません」と答えた場合は、ファイルの処理（`INDEXED`）が完了しているか、または質問の言葉がファイル内の表現と異なっていないか確認してください。

これで初期設定は完了です。続いて、本格的なコンテンツの追加（第 4 章）や、エンドユーザーへのチャットボット公開（第 6 章）に進みましょう。

第 3 章：エンドユーザーガイド

この章は、Knowledge Index のチャットボット（Widget）を利用して質問するエンドユーザーの方を対象としています。管理画面の操作ではなく、**Web サイト上に表示されるチャットウィジェットの使い方** を解説します。

3.1 チャットを始める

Widget の見つけ方

Knowledge Index のチャットボットは、Web サイトの画面右下（または設置場所によって異なる位置）に表示される**丸いアイコンボタン** です。

このボタンをクリック（タップ）すると、チャット画面が開きます。

チャット画面の構成

チャット画面は、以下のパーツで構成されています。

パーツ	位置	説明
ヘッダー	上部	タイトルと操作ボタン（戻る、履歴クリア、表示切替）
メッセージエリア	中央	AI との会話が表示される領域
入力エリア	下部	テキスト入力欄、送信ボタン、音声入力ボタン
ファシリテータ切替	入力欄上部	AI の役割を切り替えるボタン（複数設定されている場合）

3.2 テキストで質問する

基本的な流れ

1. 入力欄に質問を入力します
2. **Enter キー**を押すか、**送信ボタン**（紙飛行機アイコン）をクリックします
3. AI が回答を生成している間、ローディング表示が出ます
4. 回答が表示されます

良い質問の仕方

AI から的確な回答を得るために、以下のポイントを意識してみてください。

ポイント	良い例	悪い例
具体的に聞く	「有給休暇の申請方法を教えてください」	「休みについて」
一つずつ聞く	「経費精算の締め切りはいつですか？」	「経費精算の締め切りと申請方法と承認フローを教えて」
用語を合わせる	社内で使われている正式名称を使う	略語や俗称だけで聞く
前提を伝える	「新入社員ですが、健康診断の受け方は？」	「健康診断って何？」

ヒント: 回答が的外れだと感じたら、質問の表現を変えて再度お試しください。同じ意味でも言い方を変えると、別の情報がヒットすることがあります。

3.3 音声で質問する

キーボード入力の代わりに、**音声で質問する** ことができます。

使い方

1. 入力エリアの **マイクアイコン** をクリックします
2. 初回利用時は、ブラウザからマイクへのアクセス許可を求められます →「**許可**」を選択してください
3. マイクアイコンが **赤色に変わり、点滅** します → 音声認識中です
4. 質問を話してください
5. 発話が完了すると、自動的にテキストに変換され、AI に送信されます
6. AI の回答が表示されます

音声入力を停止する

- 音声認識中にマイクアイコンをもう一度クリックすると **停止**します

音声入力がうまくいかない場合

症状	対処法
マイクアイコンをクリックしても始まらない	ブラウザの設定でマイクを許可しているか確認してください
音声が正しく認識されない	静かな環境で、はっきりとお話しください
エラーが表示される	テキスト入力に切り替えてご質問ください

注意: 音声入力は Web Speech API を使用しており、Google Chrome での利用を推奨します。一部のブラウザでは利用できない場合があります。

3.4 回答の読み方

回答の構成

AI の回答には、テキスト本文に加えて以下の情報が含まれることがあります。

■ソース情報（参照元） 回答の下に「 ソースを表示」というボタンが表示される場合があります。これをクリックすると、AI が回答の根拠とした **元のドキュメントの情報** を確認できます。

- **ファイル名**: 参照されたドキュメントの名前
- **抜粋テキスト**: 参照された箇所のテキスト（チャンクの一部）

ソース情報を確認することで、回答の信頼性を判断したり、元のドキュメントを直接確認することができます。

■フィードバックボタン　回答の下に　（良い）と　（悪い）のボタンが表示されます。

- **　を押す場合**: 回答が役に立った、的確だったとき
- **　を押す場合**: 回答が的外れだった、不十分だったとき

フィードバックは AI の改善に役立てられます。ぜひ気軽にご利用ください。

回答についての注意点

- AI の回答は登録されたドキュメントに基づいて生成されていますが、完全な正確性を保証するものではありません。
- 重要な判断を行う場合は、回答のソース情報を確認し、必要に応じて担当部署にご確認ください。
- AI が「情報が見つかりません」と回答した場合、関連するドキュメントが登録されていない可能性があります。管理者にお問い合わせください。

3.5 ファシリテータを切り替える

ファシリテータとは

ファシリテータは、AI の「回答スタイル」を用途に合わせて切り替える機能です。管理者が設定した複数のファシリテータの中から選択できます。

切り替え方

入力エリアの上部に、丸いアイコンが横一列に並んでいる場合があります。これがファシリテータの切り替えボタンです。

- アイコンをクリックすると、そのファシリテータに切り替わります
- 選択中のファシリテータは、アイコンの下に名前が表示されます

よく用意されているファシリテータの例

ファシリテータ	アイコン	特徴
AI アシスタント		正確で客観的な回答。迷ったらこれ
カスタマーサポート		親切で丁寧。トラブル相談向け
FAQ・サイト案内	☒	簡潔な回答。手早く答えを知りたいとき
社内ヘルプデスク		規程やマニュアルの根拠を示す回答
教育・研修		ヒントを出しながら考えさせるスタイル

ポイント: 同じ質問でも、ファシリテータを変えると回答のトーンや詳しさが変わります。目的に合ったファシリテータを選ぶと、より満足度の高い回答が得られます。

3.6 便利な機能

文字密度の変更

ヘッダー右側の **≡アイコン**（三本線）をクリックすると、チャット画面の文字の詰まり具合を切り替えられます。情報量が多い回答を読むときに便利です。

会話履歴のクリア

ヘッダー右側の　**アイコン**（ゴミ箱）をクリックすると、現在の会話履歴をクリアできます。確認ダイアログが表示されるので、「OK」を選択してください。

> **いつクリアすべきか**: 話題を大きく変えたいとき、または AI の回答が以前の会話に引きずられていると感じたときに、会話をリセットすると効果的です。

Widget の位置移動

チャットボタン（丸いアイコン）は **ドラッグ&ドロップ**で画面上の好きな位置に移動できます。他のコンテンツと重なって邪魔な場合に便利です。

チャット画面を閉じる

ヘッダー左端の **← ボタン**をクリックするか、チャット画面の外側をクリックすると閉じます。会話の内容はブラウザセッション中は保持されます。

3.7 よくある質問（FAQ）

Q. チャットボットが表示されません

- ブラウザの JavaScript が有効になっているか確認してください。
- 広告ブロッカーなどの拡張機能がチャットボットをブロックしている可能性があります。一時的に無効にしてお試しください。

Q. 回答までに時間がかかります

- AI の回答生成には通常 3〜10 秒程度かかります。テキスト量が多い回答や、高性能モデルを使用している場合はさらに時間がかかる場合があります。
- ネットワーク接続が不安定な場合も遅延が発生します。

Q. 音声入力が使えません

- Google Chrome の最新バージョンをご利用ください。
- HTTPS で配信されているサイトでのみ利用可能です。
- マイクの使用許可がブロックされている場合は、ブラウザのアドレスバー付近にある鍵アイコンから許可設定を変更してください。

Q. 以前の会話を見たいのですが

- Widget のチャット履歴は、ブラウザのセッション中は保持されます。ブラウザを閉じると履歴はクリアされます。
- 長期保存が必要な回答は、テキストをコピーして保存してください。

Q. AI が間違った回答をしました

- 回答の下にある　ボタンを押して、フィードバックをお送りください。改善に役立てられます。
- 質問の表現を変えて再度お試しください。
- 重要な情報については、必ず公式ドキュメントや担当者に確認してください。

Q. 対応言語は？

- 日本語をはじめ、英語、中国語など複数の言語に対応しています。質問した言語と同じ言語で回答します。

チャットボットの使い方でお困りの際は、画面上のチャットボットに「使い方を教えて」と話しかけてみてください。

第 4 章：コンテンツ管理

Knowledge Index が回答の根拠とするドキュメント（コンテンツ）を管理する方法について解説します。適切なコンテンツの登録と管理が、優れた AI 回答の土台となります。

4.1 コンテンツの登録

Knowledge Index にドキュメントを読み込ませるには、「ファイルアップロード」と「URL 入力」の 2 つの方法があります。

方法 A：ファイルアップロード

手元のファイルを直接アップロードして登録する方法です。

1. サイドメニューの「コンテンツ管理」をクリックします
2. 画面右上の「新規コンテンツ」をクリックします
3.「入力方法」で「ファイルをアップロード」を選択します
4. ファイル入力エリアをクリック（またはファイルをドラッグ＆ドロップ）してファイルを選択します※**自動でファイルタイプが判定されます**
5. 以下のメタデータを入力します
   - **タイトル**（必須）：わかりやすい名前
   - **説明**（任意）：ドキュメントの概要

- **タグ**（任意）：カンマ区切りで入力（検索での絞り込みに利用します）

6.「保存」をクリックします

方法 B：URL 入力

Web 上に公開されているページを指定して、テキストを読み込ませる方法です。

1. サイドメニューの「コンテンツ管理」をクリックします
2. 画面右上の「新規コンテンツ」をクリックします
3.「入力方法」で「URL を入力」を選択します
4. 対象の URL を入力します（例：`https://example.com/document/`）
5. タイトル、説明、タグを入力します
6.「保存」をクリックします

> **注意（URL 入力について）**- システムがアクセスできない環境（イントラネット内やログインが必要なページ）の URL は読み込めません。- 読み込み処理は 30 秒でタイムアウトします。重いページや応答が遅いページはエラーになる場合があります。

4.2 対応ファイル形式と制限

対応しているファイル形式

現在、以下の 10 種類のファイル形式に対応しています。

- **Word** (`.docx`): 業務マニュアルや手順書など。標準的な見出しと段落テキストが抽出されます。
- **Excel** (`.xlsx`): データシートや各種一覧等に。各シートとセルの値が結合して抽出されます。
- **PowerPoint** (`.pptx`): 会議資料やスライドに。スライド内のテキストボックスの内容が抽出されます。
- **PDF** (`.pdf`): 規程集や出力済みドキュメントなど。ただし、画像化された文字（スキャン画像など）は読み取れません。
- **Markdown** (`.md`): 見出し等の構造情報が正確に保持されるため、**最も RAG に適した形式**です。
- **ePub** (`.epub`): 電子書籍フォーマット。章立てされた長いドキュメントに適しています。
- **RTF** (`.rtf`): リッチテキストファイル。
- **テキスト** (`.txt`): 単純なテキストデータ。
- **CSV** (`.csv`): FAQ データなど、1 行ごとに意味が完結する構造データの読み込みに適しています。
- **HTML** (`.html`): Web ページのソースなど。

制限事項

- **ファイルサイズ**: 1 ファイルあたり最大 **50MB** まで
- **タグの数**: 1 つのコンテンツにつき最大 **10 個**まで（1 タグ最大 50 文字）
- **文字数上限**: タイトル最大 255 文字、説明最大 1000 文字

4.3 コンテンツの処理フローとステータス

登録されたコンテンツは、直ちに検索対象になるわけではありません。AI が利用可能な「ベクトル」に変換されるまで、バックグラウンドで処理が行われます。

コンテンツのステータスは以下のいずれかになります。コンテンツの詳細画面で確認できます。

ステータス	意味
UPLOADED	アップロード完了。まもなく処理が開始されます。
PROCESSING	処理中。テキスト抽出、チャンク（断片）分割、ベクトル生成を行っています。
INDEXED	**インデックス完了**。チャットの検索対象として利用可能になりました。
FAILED	処理失敗。ファイルが読み取れなかったか、その他のエラーが発生しました。

> **ヒント**: 正常にインデックスされるまで、ファイルのサイズに応じて数秒から数分かかります。一覧画面をリロードしてステータスが `INDEXED` になるのをお待ちください。

4.4 タグの設定と活用

コンテンツには任意の「タグ」を付けることができます。タグは AI の検索精度を高める上で非常に重要な役割を果たします。

タグの役割

ユーザーがチャットで質問する際、システム側（ファシリテータ設定など）で指定されたタグを持つコンテンツ**だけ**を検索対象に絞り込むことができます。これにより、「経理の質問に人事の規程が回答に混ざる」といったミスマッチを防ぐことができます。

タグ設定のポイント

- **分類軸を揃える**:「営業部」「人事部」などの「部署」や、「マニュアル」「規程」などの「文書種別」など、組織内で統一したルールを定めておくことをお勧めします。
- **付けすぎに注意**: 1 つのファイルに手当たり次第にタグを付けると、絞り込みの効果が薄れてしまいます（最大 10 個まで）。
- **後から編集可能**: タグはコンテンツの再登録なして、いつでも変更・追加できます。

4.5 チャンク（分割データ）の確認

RAG では、1 つの長いドキュメントを数百文字程度の「チャンク」と呼ばれる短い断片に分割して保存・検索します。登録したコンテンツがどう分割されたかは、画面から直接確認できます。

1. コンテンツ一覧から、対象のドキュメントをクリックします
2. **「チャンク」タブ** を選択します
3. ドキュメントがどのように分割されたか（チャンクテキストのリスト）が表示されます
4. **「分析」タブ**では、生成されたチャンクの総数や総文字数などの統計を確認できます

> **回答がおかしい時はここをチェック**チャットで期待した回答が得られない場合、チャンクタブを見てみましょう。表が崩れて分割されていたり、必要な文脈がバラバラのチャンクに分断されている場合は、元のドキュメントのレイアウトや見出しを整理して再登録することで精度が劇的に改善することがあります。

4.6 コンテンツの修正と削除

メタデータ（タイトル・説明・タグ）の編集

1. 一覧画面で対象のコンテンツをクリックし、「編集」ボタンをクリックします
2. タイトル、説明、タグを変更し「保存」をクリックします
3. ※この操作では再インデックス（ベクトル再計算）は発生しないため、すぐに完了します。

元ファイル内容の更新

Knowledge Index では、一度登録したコンテンツの中身（ファイルそのもの）を直接差し替えることはできません。ドキュメントの内容が更新された場合は、以下の手順を実施してください。

1. 更新された新しいファイルを**新規コンテンツ**として登録する
2. インデックス完了（`INDEXED`）を確認する
3. 古いコンテンツを**削除**する

再インデックスの実行

ステータスが `FAILED` になってしまった場合などに、処理をやり直すことができます。

1. コンテンツ詳細画面を開きます
2. 画面上の「再インデックス」ボタンをクリックします
3. ステータスが `PROCESSING` に戻り、バックグラウンドでの再処理が開始されます

コンテンツの削除

1. コンテンツ一覧から削除対象の行、または詳細画面を開きます
2. 「削除」ボタンをクリックします
3. 確認ダイアログで「削除」を確定します

注意: 削除したコンテンツはシステムおよび検索インデックスから完全に消去され、元に戻すことはできません。

第 5 章：RAG の仕組みと回答精度の向上策

Knowledge Index の中核は **RAG**（Retrieval-Augmented Generation：検索拡張生成）と呼ばれる技術です。この章では、RAG の基本的な動作原理を解説し、「どうすれば AI がより正確に答えてくれるようになるか」を実践的にガイドします。

5.1 RAG とは

RAG は、ユーザーの質問に対して**自社のデータを検索し、その内容を踏まえて AI が回答を生成する**仕組みです。一般的な AI チャットボットは学習済みの知識だけで回答しますが、RAG では「あなたの組織のドキュメント」を参照するため、組織固有の情報に基づいた回答が可能になります。

RAG の流れ（概要）

RAG は大きく **2 つのパイプライン** で構成されています。

■パイプライン①：コンテンツ登録時（事前準備）

ファイルアップロード → テキスト抽出 → チャンク分割 → ベクトル化 → データベース保存

1. **テキスト抽出**: アップロードされた Word・Excel・PDF・Markdown 等からテキストを取り出します。
2. **チャンク分割**: 長いテキストを扱いやすい小さな塊（チャンク）に分割します。
3. **ベクトル化（埋め込み生成）**: 各チャンクを AI（埋め込みモデル）により数値ベクトルに変換します。このベクトルが「意味」を数学的に表現したものです。
4. **データベース保存**: ベクトルとテキストを PostgreSQL + pgvector データベースに保存します。

■パイプライン②：チャット応答時（質問→回答）

ユーザーの質問 → クエリ拡張 → ベクトル検索 → 関連コンテンツ取得 → `LLM` で回答生成

1. **クエリ拡張**: 会話履歴を考慮して、検索に適したクエリを AI が自動生成します。
2. **ベクトル検索**: 質問のベクトルと、保存済みチャンクのベクトルを比較し、意味的に近いものを上位 5 件取得します。
3. **コンテキスト構築**: 検索で見つかったチャンクを整理して、AI への「参考資料」として組み立てます。
4. **LLM 応答生成**: 参考資料と質問を AI（LLM）に渡し、回答を生成します。

ポイント: RAG の回答精度は、「検索で適切なチャンクが見つかるか」と「AI がそのチャンクを正しく解釈して回答できるか」の 2 つに依存します。

5.2 コンテンツの品質と構造

RAG の精度に **最も大きな影響を与えるのはコンテンツの品質**です。AI が参照する「元データ」が曖昧であれば、回答も曖昧になります。

良いコンテンツの条件

よいコンテンツ	悪いコンテンツ
見出し・段落が整理されている	テキスト全体が一塊で構造がない
一つの話題が一つのセクションにまとまっている	複数の話題が混在している
正確で最新の情報	古い情報が残っている
冗長な表現がなく簡潔	同じ内容が何度も繰り返されている
専門用語に説明がある	略語や暗号的な用語が説明なく使われている

ファイル形式ごとの注意点

- **Word / PowerPoint**: 標準的な見出し構造や段落、スライド内のテキストボックス等を反映して抽出が行われます。
- **Excel**: シート内のセルテキストが結合されて抽出されます。表の構造よりも、説明文が含まれるセル（仕様書など）を RAG で検索する用途に向いています。
- **PDF**: 表や図の中のテキストは抽出されますが、画像内の文字（OCR が必要なもの）は取得できません。レイアウトが複雑な PDF は、テキストの順序が乱れることがあります。
- **Markdown / テキスト**: 最もクリーンにテキスト抽出できます。見出し（`#`, `##`）を使って構造化されているものが理想的です。
- **HTML**: ナビゲーションやフッターなどのノイズも含まれることがあります。本文部分だけを抜き出したファイルが望ましいです。
- **CSV**: 各行が独立したチャンクとして扱われます。FAQ データなど、行単位で意味が完結するデータに適しています。

コンテンツ量のバランス

- **少なすぎる場合**: 質問に対応するチャンクが見つからず、「情報がありません」という回答になりがちです。
- **多すぎる場合**: 似た内容のチャンクが大量にヒットし、関連度の低い情報がノイズとなることがあります。
- **推奨**: 同一トピックについて重複するコンテンツは統合し、網羅性を保ちつつ冗長性を排除するのが理想です。

5.3 タグによる検索精度の向上

タグとは

コンテンツにタグを付けることで、検索時にスコープ（検索範囲）を限定できます。タグは**ファイル単位** で最大 10 個まで設定可能です。

タグの効果

タグを活用すると、ベクトル検索の対象をフィルタリングでき、**関連性の低いコンテンツを排除**できます。

■例：IT 企業の社内ナレッジ

コンテンツ	タグ
就業規則.pdf	**人事**, **規程**
経費精算マニュアル.md	**経理**, **手続き**
AWS 運用ガイド.pdf	**インフラ**, `AWS`
営業マニュアル.md	**営業**, **手続き**

このようにタグを設計しておくと、ファシリテータごとに検索対象を絞り込めるため、「経費精算について聞いたのに AWS の情報が混ざる」といった問題を防げます。

効果的なタグ設計のポイント

1. **カテゴリ体系を決める**: 部署名、業務領域、ドキュメント種別など、組織に合った分類軸を定めましょう。
2. **粒度を揃える**:「人事」と「人事部門の年末調整」のように粒度がバラバラだと効果が薄れます。
3. **表記を統一する**:「人事」と「HR」のような揺れを避けましょう。
4. **多すぎない**: 1 ファイルに必要以上のタグを付けると、フィルタリングの効果が薄れます。3〜5 個が目安です。

5.4 チャンキングの仕組みと影響

チャンキングとは

長いドキュメントは、そのまま AI に渡すことができません。そこで、テキストを**チャンク**（小さな断片）に分割します。Knowledge Index では **RecursiveCharacterTextSplitter** という手法を使い、文章の自然な区切り（段落、改行、句読点）を見つけて分割します。

デフォルト設定

パラメータ	デフォルト値	説明
チャンクサイズ	1024 文字	1 つのチャンクの最大長
オーバーラップ	200 文字	前のチャンクとの重複部分

分割の優先順位

テキストは以下の区切り文字を順番に試して分割されます：

1. \n\n（段落区切り）
2. \n（改行）
3. 。！？（日本語の文末）
4. .！?（英語の文末）
5. （スペース）
6. 空文字（最後の手段として 1 文字ずつ分割）

オーバーラップの役割

隣り合うチャンクに 200 文字の重複を持たせることで、分割境界付近の文脈が失われるのを防ぎます。例えば、ある段落の結論が次のチャンクの冒頭に含まれるようにすることで、検索ヒット率が向上します。

チャンクサイズの影響

チャンクサイズ	メリット	デメリット
小さい（512 文字程度）	特定の質問にピンポイントでヒットしやすい	文脈が切れて情報が不完全になりがち
大きい（2048 文字程度）	十分な文脈を含むチャンクが得られる	無関係な情報も含まれやすく、ノイズが増える
デフォルト（1024 文字）	バランスが取れている	多くの用途で適切

> **推奨**: まずはデフォルト設定（1024 文字 / 200 文字オーバーラップ）で運用し、回答精度を見ながら調整してください。

5.5 LLM モデルの選択と調整

対応モデル一覧

Knowledge Index は、以下の主要 AI プロバイダーに対応しています。

■OpenAI

モデル名	特徴	推奨用途
gpt-4o	最新の高性能モデル。高速かつ高精度	一般的な用途（推奨）
gpt-4o-mini	高速・低コストのバランスモデル	コスト重視の大量問い合わせ
gpt-4	高い推論能力	複雑な質問への回答
gpt-4-turbo	GPT-4 の高速版	速度が重要な場面
gpt-3.5-turbo	低コスト・高速	シンプルな FAQ

■Anthropic

モデル名	特徴	推奨用途
claude-3-5-sonnet	高精度でバランスの良いモデル	一般的な用途（推奨）
claude-3-opus	最高精度のモデル	正確性が最重要な場面
claude-3-sonnet	速度と精度のバランス	多くの業務用途
claude-3-haiku	高速・最低コスト	大量の定型質問

■Google

モデル名	特徴	推奨用途
gemini-pro	Google の汎用モデル	一般的な用途
gemini-1.5-pro	長文コンテキスト対応	長い文書を参照する質問

温度パラメータ（Temperature）

温度パラメータは、AI の回答の「ぶれ具合」を制御します。

値	特徴	適した用途
0.0	毎回ほぼ同じ回答。最も再現性が高い	社内規程の回答、手続き案内
0.3	わずかなバリエーション	FAQ、カスタマーサポート
0.7（デフォルト）	自然な多様性	一般的な Q&A
1.0	創造的で多様な回答	ブレインストーミング、アイデア出し

> **推奨**: ナレッジ検索の用途では **0.0〜0.3** が推奨です。正確性が求められる場面では温度を下げてください。

最大トークン数（Max Tokens）

回答の長さの上限を制御します。デフォルトは 500 トークン（日本語で約 250〜400 文字相当）です。

値	適した用途
200	短い一文回答（FAQ）
500（デフォルト）	標準的な回答
1000	詳細な説明が必要な回答
2000	手順書レベルの長い回答

5.6 ファシリテータ機能

ファシリテータとは

ファシリテータは、AI の**「役割」や「振る舞い」を定義するプリセット**です。同じコンテンツに対しても、ファシリテータを切り替えることで AI の応答スタイルを変えることができます。

プリセットファシリテータ

Knowledge Index には、テナント作成時に以下の 5 つのファシリテータが自動で用意されます。

■1. AI アシスタント（デフォルト）

- **用途**: 一般的な質問やナレッジ検索
- **特徴**: 中立的・客観的なトーン、コンテキスト情報のみに基づく正確な回答
- **振る舞い**:
  - 丁寧な「です・ます」調
  - コンテキストにない情報は「情報が含まれていません」と正直に回答
  - 結論を先に述べ、簡潔にまとめる

■2. カスタマーサポート

- **用途**: お問い合わせ・トラブル対応
- **特徴**: 共感的で丁寧、段階的な解決手順を提示
- **振る舞い**:
  - まず共感と謝罪から入る
  - 状況整理 → 原因 → 解決手順 → 次のアクションの順で回答
  - 返金・補償など権限外の約束はしない

■3. FAQ・サイト案内

- **用途**: よくある質問やサイト内ガイド
- **特徴**: 迅速かつ簡潔、自己解決を促す回答
- **振る舞い**:
  - 挨拶は最小限、すぐ本題に入る
  - 中学生にもわかる平易な言葉で回答
  - 関連するマニュアルへのリンクを提示

■4. 社内ヘルプデスク

- **用途**: 社内規程やシステム操作の問い合わせ
- **特徴**: ビジネスマナーを用いた正確な回答、情報源を明示
- **振る舞い**:
  -「○○規程第○条によると」など根拠を必ず示す

  - 対象者・期限・必要なものを含めて回答
  - 情報がない場合は「担当部署に直接ご確認ください」と案内

■5. 教育・研修

- **用途**: 業務知識の学習サポート
- **特徴**: 答えを即座に与えず、思考を促すスタイル
- **振る舞い**:
  - ヒントを出しながら自律的な思考をサポート
  - 間違いを直接否定せず、別の視点を提示
  -「ここまでで疑問はありませんか？」と対話を促す

カスタムファシリテータの作成

プリセット以外にも、テナント管理者は独自のファシリテータを作成できます。

■作成手順

1. 管理画面から「テナント設定」→「ファシリテータ」タブを開く
2.「新規作成」をクリック
3. 以下の項目を設定：
   - **名前**: ファシリテータの表示名
   - **説明**: 用途の簡単な説明
   - **アイコン**: Widget での表示アイコン
   - **システムプロンプト**: AI への指示文（次のセクション参照）
4.「保存」をクリック

■効果的なシステムプロンプトのテンプレート

あなたは「[役割名]」です。[概要説明]

【行動ルール】

1．[トーン・マナーに関するルール]

2．[情報源に関するルール]

3．[回答できない場合のルール]

【回答の構造】

1．[回答の書き方のルール]

2．[フォーマットのルール]

【禁止事項】

- [やってはいけないこと]

【言語設定】

ユーザーの入力言語（質問の言語）と同じ言語で回答してください。

> **ポイント**: システムプロンプトは具体的であるほど効果的です。「丁寧に答えてください」よりも「共感を示した後、箇条書きで解決手順を提示してください」のように具体的な行動を指示しましょう。

多言語対応

すべてのプリセットファシリテータには **言語設定ルール**が含まれており、ユーザーの質問と同じ言語で回答します。英語で質問すれば英語で、中国語なら中国語で回答します。カスタムファシリテータにも同様のルールを追加することを推奨します。

5.7 会話履歴とコンテキスト

会話コンテキストの仕組み

Knowledge Index では、直近の会話履歴（最新 10 件）を自動的に AI に渡します。これにより、AI は前の質問や回答を踏まえた応答が可能です。

■例

ユーザー：有給休暇は何日もらえますか？
AI：入社 6 ヶ月後に 10 日、以降は勤続年数に応じて付与されます。

ユーザー：申請方法は？
AI：（「有給休暇の申請方法」と理解して）社内ポータルの「勤怠管理」
メニューから申請できます。申請期限は取得日の 3 営業日前です。

2 回目の質問「申請方法は？」だけでは何の申請かわかりませんが、会話履歴があるため「有給休暇の申請方法」と正しく解釈できます。

クエリ拡張

ベクトル検索の精度を高めるため、Knowledge Index はユーザーの質問を **AI が自動的に検索向けに書き換え**ます。

- **元の質問**:「それは誰に聞けばいい？」
- **拡張されたクエリ**:「有給休暇の申請方法について問い合わせる担当部署」

この拡張により、曖昧な質問でも適切なコンテンツがヒットしやすくなります。

長い会話での注意点

会話が非常に長くなると、以下のような影響が出ることがあります。

- **トークン制限**: 会話履歴が長くなると、コンテキストに使えるトークンが減り、参照できるチャンク数が制限されます。
- **話題の混在**: 1 つの会話で複数の話題を扱うと、検索結果にノイズが混ざりやすくなります。

推奨: 話題が変わったときは「新規チャット」を開始してください。会話をリセットすることで、新しい話題に集中した検索が行われます。

5.8 回答精度のチェックリスト

AI の回答がおかしいと感じたとき、以下のチェックリストに沿って原因を特定してください。

「関連する情報がありません」と言われる場合

- □ **コンテンツは登録済みですか？** → コンテンツ管理で確認
- □ **コンテンツのステータスは「INDEXED」ですか？** →「PROCESSING」や「FAILED」の場合、再インデックスを実行
- □ **質問の表現がコンテンツと大きく異なっていませんか？** →コンテンツ内で使われている用語で質問してみる
- □ **タグでフィルタリングされすぎていませんか？** → タグの設定を確認

回答が的外れな場合

- □ **複数の話題を 1 つの会話で聞いていませんか？** →新規チャットを開始
- □ **コンテンツに似た話題が多すぎませんか？** →タグを活用してスコープを絞る
- □ **コンテンツの構造は整理されていますか？** → 見出し・段落で構造化
- □ **温度パラメータが高すぎませんか？** → 0.0〜0.3 に下げてみる

回答が短すぎる / 不十分な場合

- □ **最大トークン数が小さすぎませんか？** → 500〜1000 に増やす
- □ **関連するコンテンツが不足していませんか？** →該当トピックのドキュメントを追加登録
- □ **チャンクサイズが小さすぎませんか？** → デフォルト（1024）を確認

回答が遅い場合

- □ **大量のコンテンツが登録されていませんか？** →タグによるフィルタリングを活用
- □ **高性能モデル（gpt-4, claude-3-opus）を使用していませんか？** →速度重視なら gpt-4o-mini や claude-3-haiku に変更
- □ **最大トークン数が大きすぎませんか？** → 必要十分な値に調整

5.9 ベクトル検索の仕組み（技術的な補足）

この節はより技術的な内容です。回答精度の向上には直接的に役立ちませんが、RAG の仕組みをより深く理解したい方向けの解説です。

ベクトルとは

テキストを AI の埋め込みモデルに通すと、1536 次元の数値ベクトル（数値の配列）に変換されます。意味が似た文章は、ベクトル空間上で近い位置に配置されます。

「有給休暇の申請方法」→ `[0.123, -0.456, 0.789, ... ]`**（1536 個の数値）**
「休暇を取る手続き」 → `[0.125, -0.452, 0.791, ... ]`**（近い値になる）**
「経費精算の締め切り」→ `[-0.567, 0.234, -0.123, ...]`**（異なる値になる）**

検索方法

Knowledge Index では **L2 距離（ユークリッド距離）**を使ってベクトル間の距離を計算します。距離が小さいほど「意味が近い」と判定されます。

検索時には、質問のベクトルに最も近いチャンクを上位 5 件（デフォルト）取得します。

HNSW インデックス

大量のチャンクがあっても高速に検索できるよう、**HNSW（Hierarchical Navigable Small World）**というインデックス構造を使用しています。これにより、数百万件のチャンクに対しても数ミリ秒で検索が完了します。

まとめ：回答精度を上げる 5 つの鉄則

#	鉄則	具体的なアクション
1	**良いコンテンツを入れる**	構造化された最新のドキュメントを登録する
2	**タグで整理する**	カテゴリ体系を決め、コンテンツに適切なタグを付ける
3	**ファシリテータを使い分ける**	用途に合ったファシリテータを選択・カスタマイズする
4	**パラメータを調整する**	温度を下げ（0.0〜0.3）、最大トークン数を適切に設定する
5	**会話をこまめにリセットする**	話題が変わったら新規チャットを開始する

この章の内容に関するご質問は、*Knowledge Index* のチャット *Widget* からお気軽にお問い合わせください。

第 6 章：Widget の設定と埋め込み

Knowledge Index で作成した賢い AI チャットボットは、管理画面の中だけで使うものではありません。**「Widget（ウィジェット）」機能**を使えば、あなたの会社の Web サイトや社内ポータル、Web アプリ内に簡

単に埋め込むことができます。

この章では、Widget の埋め込み方法からデザインのカスタマイズ、外部からのプログラマブルな制御方法までを解説します。

6.1 Widget の仕組みと事前準備

Widget は、Web ページの右下などに浮かんで表示されるチャット画面の UI 部品です。

事前準備（管理画面での確認）

Widget を外部サイトに設置するには、以下の 2 つの情報が必要です。

1. **テナント ID**: ダッシュボードの「テナント設定」画面などで確認できる、あなたの組織専用の識別子です。（例：`123e4567-e89b-12d3...`）
2. **システム API ベース URL / CDN URL**: Knowledge Index 本体が稼働しているサーバーのアドレスと、Widget プログラムファイル（JS）が配信されているアドレスです。通常、プラットフォーム管理者から提供されます。（例：`https://api.knowledge-index.jp/api/v1` など）

   重要: Widget 用の API リクエストを正しく処理するため、管理画面の「コンテキスト・プロンプト」設定（チャットベースモデルや各種 API キー等）が完了している必要があります。

6.2 Widget の埋め込み方法

自社サイトへの埋め込みは非常にシンプルです。用途に合わせた 2 つの方法を紹介します。

方法 A：HTML への直接埋め込み（最も簡単）

会社のホームページ（WordPress などの CMS や、静的 HTML）に設置する場合の一般的な方法です。ページの`</body>` タグの直前に以下の数行のコードを追加するだけです。

```
<!-- Knowledge Index Chat Widget -->
<script>
  window.RAG_CHAT_CONFIG = {
    tenantId: "ここにテナント ID を入力してください",
    apiBaseUrl: "https://api.knowledge-index.jp/api/v1", // 提供された API URL
    template: "standard", // デザインテンプレート
    launcherIcon: "chat"  // 右下のアイコン
  };
</script>
<script src="https://cdn.knowledge-index.jp/widget.js" async defer></script>
```

`tenantId` の部分を、あなたのテナント ID に書き換えてください。

方法 B：Next.js や React アプリへの組み込み（開発者向け）
自社開発の Web アプリに組み込む場合は、コンポーネントとしてマウントする方法が便利です。App Router 環境での `layout.tsx` などの共通レイアウトに配置します。

```
// components/ChatWidget.tsx
'use client';
import { useEffect } from 'react';

export default function ChatWidget() {
  useEffect(() => {
    // 既存のスクリプトがない場合のみ追加
    if (!document.getElementById('rag-chat-widget-script')) {
      // 設定オブジェクトをグローバルスコープに定義
      window.RAG_CHAT_CONFIG = {
        tenantId: process.env.NEXT_PUBLIC_WIDGET_TENANT_ID,
        apiBaseUrl: process.env.NEXT_PUBLIC_API_URL,
        template: 'support',
        launcherIcon: 'bot2'
      };

      const script = document.createElement('script');
      script.id = 'rag-chat-widget-script';
      script.src = process.env.NEXT_PUBLIC_WIDGET_CDN_URL;
      script.async = true;
      document.body.appendChild(script);
    }
  }, []);

  return null;
}
```

環境変数（`.env`）から ID や URL を読み込むことで、安全にかつ動的に切り替えることができます。

6.3 デザインテンプレートとアイコンのカスタマイズ

Web サイトの雰囲気や用途に合わせて、Widget の見た目を変えることができます。

テーマテンプレートの選択 (`template`)
チャット画面全体のカラーリングや雰囲気を設定します。

- **`standard`**: デフォルト。汎用的で清潔感のあるデザイン。

- **support**: コーポレートサイトや、一般のお客様向けのお問い合わせ（カスタマーサポート）窓口に適したフォーマルなデザイン。
- **faq**: 色味を抑えたシンプルなデザイン。FAQ ページや規程集などに最適。
- **helpdesk**: 社内ポータルや情シス向けのヘルプデスクに適した、少し堅めのデザイン。
- **training**: 社内研修や e-ラーニング環境に馴染む、柔らかなデザイン。

ランチャーアイコンの選択 (launcherIcon)
画面右下に常に表示される「チャット開始ボタン」のアイコンを設定します。

- **ciel**: 公式マスコットキャラクター（Ciel）
- **chat**: シンプルな吹き出し（最も標準的）
- **robot**: 四角いロボットの顔
- **bot2**: 丸みのあるロボットの顔
- **support**: ヘッドセットのアイコン（お問い合わせ窓口向け）
- **question**: ハテナマーク（よくある質問への誘導向け）
- **sparkle**: キラキラマーク（AI っぽさの演出向け）
- **logo**: 丸いロゴプレースホルダー

  実装時は window.RAG_CHAT_CONFIG オブジェクトの中の template および launcherIcon プロパティの文字列を書き換えてください。

カスタムランチャーアイコンのアップロード（ダッシュボードから設定）
プリセット一覧にはないオリジナルの画像（会社ロゴや自社キャラクターなど）をランチャーボタンとして表示したい場合は、**管理画面のダッシュボードから独自の画像ファイルをアップロード**できます。

■設定手順

1. ダッシュボード（/dashboard）にログインします。
2.「**カスタムランチャーアイコン**」セクションを見つけます。
3.「**画像アップロード**」ボタンをクリックし、使用したい画像ファイルを選択します。
4. アップロードが完了すると、プレビューが表示されます。
5. **埋め込みコードの変更は不要です。**アップロード後、Widget を設置したページをリロードするだけで自動的に反映されます。

■対応フォーマットと推奨サイズ

項目	詳細
対応フォーマット	PNG・JPEG・GIF・SVG・WebP
推奨サイズ	**64×64 px 以上の正方形画像**（長方形のファイルも指定可能ですが、円形のボタン内に収まるよう自動トリミングされます）

■プリセットアイコンに戻す　カスタムアイコンを削除してプリセットの状態に戻したい場合は、「**プリセットに戻す**」ボタンをクリックしてください。直ちにカスタム画像が削除され、埋め込みコードに設定した `launcherIcon` のプリセットが表示されます。

> **ヒント**: ホスティング専用 URL や埋め込みコード経由で表示した Widget には、常に最新の設定が反映されます。管理画面での変更後はページのリロードを行ってください。

6.4 外部からのプログラマブルな制御（JavaScript API）

「Web サイト上のボタンをクリックしたときにチャット画面を開きたい」といった要望に応えるため、**グローバル関数 `window.ragChat()` API** が用意されています。

Widget 控制コマンド

ページ内で以下の JavaScript 関数を実行することで、Widget を外部から操作できます。

コマンド	解説	実行例
`open`	チャット画面を強制的に開きます。	`window.ragChat('open');`
`close`	チャット画面を強制的に閉じます。	`window.ragChat('close');`
`toggle`	閉じている時は開き、開いている時は閉じます。	`window.ragChat('toggle');`

活用例：サイト内のボタンでチャットを開く

「製品サポートへ問い合わせる」といった目立つボタンを本文中に配置し、それをクリックしたときに直接チャット Widget を起動させる実装例です。

```
<button onclick="window.ragChat('toggle')">
  AI チャットボットに質問する
</button>
```

Knowledge Index の公式サポートページにある「AI チャットボットに質問」ボタンも、この機能を利用して実装されています。サイトの導線を設計する際に大変便利な機能です。

次のステップ: Widget の設置が完了したら、実際にスマートフォンと PC 両方のブラウザからアクセスし、正しく表示されるか、質問と回答ができるかを確認してください。

第 7 章：ユーザー管理

Knowledge Index では、安全な運用を実現するため、利用者の役割に応じた詳細な「ロール（権限）」管理機能を提供しています。

7.1 ロールの種類と権限

システム上には大きく分けて 4 つのロールが存在します。

ロール	想定される利用者	主な権限
PLATFORM_ADMIN(プラットフォーム管理者)	Knowledge Index プラットフォームの運営事業者	全機能のフルアクセス。全テナントの管理、プラン変更、全ユーザーの管理が可能。
TENANT_ADMIN(テナント管理者)	各テナント（組織）の導入・運用責任者	自テナント内の設定（LLM モデル、API キーの登録）、自テナント内のユーザー追加・削除、全コンテンツの管理が可能。
OPERATOR(運用者)	コンテンツを実際に追加・更新する担当者	テナント設定にはアクセス不可。コンテンツの登録・編集・削除、チャット、分析画面の閲覧が可能。
AUDITOR(監査者)	内部監査や利用状況のモニタリング担当者	全データの**閲覧のみ（Read Only）**が可能。コンテンツの追加や設定変更は不可。

7.2 ユーザーの追加と管理

テナント管理者は、自テナントに所属するユーザーを追加・削除することができます。

新規ユーザーの作成手順

1. 左メニューから「ユーザー管理」をクリックします。
2. 画面右上の「新規ユーザー追加」をクリックします。
3. 以下の情報を入力します：
   - メールアドレス（ログインに使用）
   - 表示名（3 文字以上の英数字とアンダースコア）
   - パスワード（8 文字以上、大文字小文字数字を含む）
   - ロール（OPERATOR または AUDITOR）
4. 「保存」をクリックします。

ユーザーの削除と権限変更

不要になったアカウントを削除したり、ロールを変更したりする場合は、一覧から対象ユーザーの「編集」または「削除」を選択してください。削除したユーザーは元に戻すことができないため注意が必要です。

第 8 章：分析・モニタリング

AI チャットボットがどのように利用されているかを把握し、改善につなげるための機能群です。

8.1 利用統計の確認

左メニューの「統計・分析」メニューからは、テナント全体でのチャットボットの利用状況をグラフや数値で確認できます。

確認できる主な指標

- **総質問数（クエリ数）**: 指定期間内にユーザーがチャットで質問した総回数。
- **アクティブユーザー数**: チャットを利用したユニークユーザーの数。
- **トークン消費量**: LLM の API で消費された文字数の目安。課金の基本となる指標です。
- **よく質問されるトピック**: クラスタリングされた上位の質問傾向を分析します。

これらの指標を使って、「どの部署で多く使われているか」「追加すべきコンテンツは何か」を判断するヒントを得ることができます。

8.2 会話履歴の分析

管理者は、ユーザーと AI との実際のやり取り（チャット履歴）を閲覧することができます。

- どのような質問に対して AI が「情報が見つかりません」と回答したかを確認し、不足している資料を追加します。
- ユーザーからのフィードバック（ ボタン、 ボタン）が付いた回答を抽出し、 がついた内容のソースを確認してチャンクサイズやタグ付けを修正します。

8.3 監査ログ (Audit Log) の確認

システム内で「誰が・いつ・何を変更したか」の証跡（ログ）を提供する機能です。

監査ログの特徴

- コンテンツの追加・削除、API キーの設定変更、ユーザーの追加など、設定・データに関するすべての変更操作が記録されます。
- 情報漏洩対策やコンプライアンス要件への対応に利用します（テナント管理者は自テナントのログのみ閲覧可）。
- 監査担当者（AUDITOR ロール）は、ログの検索・フィルタリング・CSV エクスポートが可能です。

第 9 章：課金・決済

Knowledge Index の利用料金は、ご契約いただいた各プランの「**基本料金（月額・前払い）**」と、AI モデルのトークン消費量に応じた「**生成 AI API 利用料（従量課金・後払い）**」で構成されます。

9.1 ライセンスと無料トライアル

新規にテナントを利用登録された場合、**最初の 14 日間は Knowledge Index の月額基本料金が無料となるトライアル期間**が適用されます。機能の制限なくすべての機能をお試しいただけます。

ただし、RAG 機能等で実際に生成 AI を利用して発生した「**API 利用料（従量課金）**」については、無料トライアル期間中であっても実費として課金が発生いたします。※お客様ご自身で取得した生成 AI の API キーをシステムに登録した場合は、システム側からの API 利用料の請求は免除されます。

9.2 トークン消費（従量課金）の仕組み

RAG では、1 回の「質問」に対してバックグラウンドで複数の AI 通信が発生します。

1. **クエリ拡張**: ユーザーの質問を検索用に書き換えるコスト（小）
2. **埋め込み生成**: 質問文をベクトル（数値）に変換するコスト（極小）
3. **回答生成**: 見つかった参考資料（コンテキスト）と質問をプロンプトにまとめて AI に回答を作らせるコスト。（中〜大）

特に、③で使われる**コンテキストの量（＝チャンクの文字数）**がトークン消費の大部分を占めます。長いドキュメントを大量に読み込ませて 1 回の回答に含めるほど、費用が高くなります。利用する各 AI モデル（GPT-4o や Claude 等）の 1 トークンあたりの単価は料金ページをご確認ください。

9.3 お支払い方法と請求サイクル

利用に当たっては事前の決済登録が必要です。Knowledge Index では「オンライン決済（クレジットカード）」と「オフライン決済（銀行振込）」を用意しています。

クレジットカード決済（Stripe）の場合

1. ダッシュボードから「料金・サブスクリプション」へ進み、決済クレジットカード情報を登録します。
2. 登録完了後、コンテンツ管理などの全機能が利用可能になります。
3. **請求タイミング（アニバーサリー課金）**:
    - 14 日間の無料トライアルが終了した翌日が、正式な「サブスクリプション更新日（課金開始日）」となります。
    - 以降、毎月の更新日に当月分の「基本料金」と、蓄積された「API 利用料」が合算されて自動決済されます。
    - ただし、一週間ごとに API 利用料を精査し、金額が「**500 円（設定閾値）**」を超過した場合は、その時点で API 利用料のみが即時請求されます。

オフライン決済（銀行振込）の場合

クレジットカードのご利用が難しい法人様向けに、請求書払い（銀行振込）によるオフライン決済を提供しています。

- **預り金（デポジット）制度**: 銀行振込を利用する場合、事前の貸倒対策として**利用開始前に一律 30,000 円 のご入金**が必要となります。

- お振込み確認後にサービスが利用可能になり、発生した毎月の基本料金や API 利用料は、この預り金から自動で充当（差し引き）されていきます。
- サービス退会（解約）時に未払金がある場合は精算を行い、残額はお客様ご指定の口座へ返金いたします（振込手数料はお客様負担）。

9.4 請求書と明細の確認

過去 15 ヶ月分の請求情報（PDF の請求書および領収書）は、すべて管理画面からダウンロード可能です。明細には、LLM モデルごとのトークン消費量内訳などの詳細が記載されます。

9.5 請求・サブスクリプションの管理操作

管理画面の「請求管理」を開き、[請求・サブスクリプション管理画面] から現在の契約プランの確認や変更、および Stripe ポータルを通じた支払情報の管理が行えます。

9.5.1 支払情報・明細の確認 (Stripe ポータル)

「支払情報・明細の確認 (Stripe)」ボタンをクリックすると、Stripe 社が提供するカスタマーポータル画面へ遷移します。

この画面では以下の操作が可能です。- クレジットカード情報の追加・削除・変更 - 過去の請求履歴の一覧確認 - 領収書・請求書の表示とダウンロード

9.5.2 プラン変更

「プラン変更」ボタンをクリックすると、プランおよび支払いサイクル（月払い・年払い）を変更するためのダイアログが表示されます。

- **プランの選択**: ベーシック、スタンダード、プロから選択可能です。
- **支払いサイクルの選択**: 年払いを選択することで、月額換算で 15% お得にご利用いただけます。
- **日割り計算（Proration）について**:
  - プランの変更は **即時適用** されます。
  - 変更に伴う差額（現在プランの未使用分と新プランの日割り料金の相殺）はシステムにより自動計算されます。
  - 変更のタイミングで即座に決済が発生することはありません。差額分は**次回の請求（更新日）**にて自動的に合算・調整されます。

9.5.3 解約（退会）・アカウント削除

本サービスの利用を終了し、解約（退会）する場合は以下の手順でご対応ください。

[解約申請メニュー]

1. **解約申請の手順**
   - 画面右上の「歯車アイコン（設定メニュー）」をクリックします。
   - メニューから**「解約申請」**を選択し、画面の案内に従って手続きを行ってください。

- 解約手続き後も、すでに支払い済みの現在の契約期間の終了日までは継続してサービスをご利用いただけます。

2. **アカウント（テナント）の完全削除について**
   - 契約期間が終了するとアカウントは停止されますが、テナント内の全情報（アップロードしたファイル、会話履歴、ユーザー情報など）を物理的に直ちに完全削除したい場合は、別途サポート窓口までご連絡ください。

第 10 章：インフラストラクチャ概要

Knowledge Index は、エンタープライズの厳格な要件に応えるため、セキュアで可用性の高いクラウド基盤上に構築されています。本章は、企業のセキュリティ担当者からインフラ要件についてよく寄せられる質問にお答えするための概要です。

10.1 Microsoft Azure 東日本リージョンでの運用

本システムの全インフラストラクチャおよびお客様のデータは、**Microsoft Azure の「Japan East（東日本）」リージョンに限定**して構築・維持されています。データが海外リージョンに保存されたり、越境移転されたりすることはありません。

主要コンポーネント構成

- **アプリケーション基盤**: Azure Container Apps 上のスケーラブルなコンテナとしてフロントエンドおよびバックエンド API が動作します。KEDA によるオートスケールが有効化されており、アクセス急増時にも安定したレスポンスを提供。
- **データベース基盤**: Azure Database for PostgreSQL Flexible Server を採用し、`pgvector` 拡張機能を用いて高速なベクトル検索（HNSW インデックス）を実現。マルチ AZ 配置により高可用性を確保しています。
- **オブジェクト記録**: 登録された元ドキュメント（PDF 等）のバイナリデータは、暗号化の上で Azure Blob Storage に保存されます。

10.2 セキュリティとデータ分離

1. **テナント分離**: データベースレベルでテナントごとの RLS（Row-Level Security: 行レベルセキュリティ）とロジック分離を実施しています。他の組織のデータへアクセスすることは物理システム的に不可能です。
2. **データの暗号化**:
   - 保管時（At-Rest）: Azure Storage および PostgreSQL の保存データは、256 ビット AES 暗号化で保護されます。
   - 転送時（In-Transit）: すべての Web 通信（API/フロントエンド）は TLS 1.2 以上で暗号化されています。
3. **認証設計**: システム内部の認証基盤には、短寿命の JWT トークンを採用し、セッション乗っ取りのリスクを最小化。また、各生成 AI（OpenAI 等）と通信する際の API キー自体も、システム内では強力に暗号化された状態で保存されます（システムのコードを見ても平文のキーは抽出不可能）。
4. **CORS 設定**: Widget 呼び出し用の API は、厳格な CORS（Cross-Origin Resource Sharing）設定で

保護しつつ、外部サイトへ安全にボットを提供できる設計としています。

10.3 モデルプロバイダー（OpenAI 等）へのデータ送信について

Knowledge Index は、内部でベクトル生成およびチャット応答のために AI 企業（OpenAI 社、Anthropic 社、Google 社等）の API を使用します。

AI 企業側のポリシー保証：API 経由で送信されたお客様のドキュメントデータ（質問文・検索結果テキスト）は、各社の明確な規約に基づき **AI の学習データとしては一切利用されません**。通信後、一定期間で破棄されます。※ 例外として、各 AI プロバイダーのオプトインを意図的に実施しない限り学習利用は禁止されています。

第 11 章：バッチ処理と自動化

Knowledge Index プラットフォームの裏側では、課金やデータの整理を目的としたバッチプログラム（バックグラウンド処理）が動作しています。

本章は、プラットフォーム運用担当者や、システムがどのようなタイミングでデータ処理を行っているかを知りたいユーザー向けの技術情報です。

11.1 主な定期実行（Cron）バッチ

システムは以下のバッチジョブ（Python スクリプトなど）を一定期間ごとに自動実行しています。

定期請求およびメータリング集計バッチ (`billing_batch.py`)

- **実行タイミング:** 日次または設定された各スケジュール間隔
- **機能:** 各テナントの「API モデル別のトークン消費数（従量課金分）」を定期的に集計し、Stripe の該当サブスクリプションに対して使用量を送信（メータリング報告）します。また、一週間の API 利用料が一定の閾値（500 円等）を超過した場合の即時請求の発火や、アニバーサリー課金（加入日ベース）更新時の決済サイクルの連動を行います。

日次イベント処理 (`daily_event_processor.py`)

- **実行タイミング:** 毎日深夜
- **機能:** `scheduled_events` テーブルに登録されている様々な日次スケジュールタスクを消化します。たとえば、「無料トライアル終了前のアラートメール送信」などを処理します。

11.2 自動データパージ（テナント解約時の処理）

利用テナントがサービスを「退会・解約」した場合のデータの流れです。

論理削除から物理削除へ

1. **退会手続き直後**: テナントを「論理削除」し、ログインとすべての API アクセスを遮断します。システム上には復元可能な形で一定期間保存されます。

2. **パージバッチの実行**: 解約から 30 日（※運用ポリシーに基づく）が経過したのち、専用のメンテナンスバッチが起動します。このバッチによって、PostgreSQL 上のテキストやベクトルデータ、および Blob Storage 上の実ファイルが **物理的に完全消去**されます。消去後のデータ復元はできません。

第 12 章：運用・メンテナンス

Knowledge Index を安定して長期間運用するための、システムメンテナンスに関する機能群です。（プラットフォーム管理者向け）

12.1 メンテナンスモードの設定と解除

システム全体のアップグレードや緊急対応の際、すべてのテナントのアクセスを一時的に遮断する「メンテナンスモード」を利用できます。

1. プラットフォーム管理者でログインし、システム管理画面にアクセスします。
2.「メンテナンスモードを有効にする」スイッチをオンにします。
3. すべてのエンドユーザーおよびテナント管理者の画面には「現在メンテナンス中です」という画面が表示されます。

合言葉（Bypass Code）によるバイパス
メンテナンスモード中でも、運用担当者だけがシステムの動作確認を行うための仕組みがあります。設定された URL パラメータ（例：`?bypass=your_secret_code`）付きでアクセスすることで、通常通りシステムにログインし、検証作業を行うことができます。

12.2 システム API キーと環境変数の管理

Knowledge Index では、各種システムの挙動（バッチ処理の有無、Stripe 連携等）を環境変数および設定ファイル (`billing_config.json` 等) で制御しています。

特に重要なのは「システム API キー」です。これはダッシュボード等のフロントエンドが、安全にバックエンド API と通信するための内部的な認証トークンです。定期的なキーのローテーションが推奨されます。

12.3 スケジュールイベント管理とリマインダー

「ScheduledEvent」という仕組みにより、システムが非同期に行うべきタスク（イベント）がデータベース上で一元管理されています。

- 無料トライアルの終了通知
- お支払いの失敗通知
- インデックス処理のタイムアウト警告

これらのイベントがスタックしていないか、管理者は定期的にログやデータベースの状況を監視し、必要に応じて手動でイベントを再実行するなどの運用を行います。

第 13 章：トラブルシューティング

日常的な運用の中で問題が発生した場合の、よくある原因と解決策をまとめました。

13.1「情報が見つかりません」とばかり回答される

- **コンテンツが登録されていない、または FAILED になっている**
  - 【対策】「コンテンツ管理」画面で、目的のドキュメントが `INDEXED` になっているか確認してください。
- **質問文の表現が不足している**
  - 【対策】ドキュメント内で使われている正式な用語を使って、もう一度詳しく質問してみてください。
- **ファシリテータやタグのフィルタリングが強すぎる**
  - 【対策】検索範囲が必要以上に狭められていないか確認してください。

13.2 チャット機能自体がエラーで停止する

- **API キーが無効、または未設定**
  - 【対策】テナント管理者が、各 AI プロバイダーの API キーを正しく登録し、「有効」にしているか確認してください。
- **API プロバイダー側の障害、または制限**
  - 【対策】OpenAI や Anthropic 等のクレジット残高不足や、各サービスのシステム障害が発生していないか、それぞれのステータスページを確認してください。

13.3 コンテンツのインデックス処理が `FAILED` になる

- **ファイルサイズが大きい (50MB 以上)**
  - 【対策】ファイルを分割して再度アップロードしてください。
- **パスワードロックのかかった PDF や壊れたファイル**
  - 【対策】パスワードを解除したファイルをアップロードするか、PDF をテキスト等で保存し直してください。
- **URL 指定したページが取得できない**
  - 【対策】30 秒のタイムアウトに引っかかっているか、システムの背後からアクセスできない（社内ネットワークや要ログイン）ページです。

13.4 Web サイトに Widget が表示されない

- **埋め込みスクリプトの誤り**
  - 【対策】HTML または JS コンポーネントに貼り付けた `tenantId` や API URL に誤りがないか確認してください。
- **CORS（クロスオリジン通信）エラー**
  - 【対策】ブラウザの開発者ツールを開き、赤字のエラー（Console 欄）が出ていないか確認してくだ

さい。API 側でドメインが許可されていない可能性があります。プラットフォーム管理者に連絡してください。

13.5 パスワードを忘れた / ログインできない

- **パスワードを忘れてしまった場合**
  - 【対策】ログイン画面にある「パスワードをお忘れですか？」のリンクをクリックし、登録しているメールアドレスを入力してください。パスワードリセット用のリンクが記載されたメールが送信されます。
- **正しいパスワードを入れても「認証に失敗しました」と表示される場合**
  - 【対策】アカウントがロックされているか、テナント管理者によって利用が一時停止されている可能性があります。まずは自組織のテナント管理者へお問い合わせください。テナント管理者自身がログインできない場合はサポート窓口へご連絡ください。

13.6 サポートへの連絡方法について

マニュアルおよび本トラブルシューティングを確認しても問題が解決しない場合は、各種サポートをご利用ください。

- **カスタマーサポート窓口 (メール)**
  - `support@example.com` 宛に、具体的なエラー内容または管理画面の「監査ログ (Audit Log)」から出力したデータを添付の上、ご連絡ください。
- **担当者への直接のお問い合わせ（個別契約の場合）**
  - プロプランをご契約中、または導入時に専任サポート担当者がついている場合は、個別にお伝えしている直通のサポートチャット（Slack や Chatwork 等）または電話番号へご連絡ください。

第 14 章 LINE 公式アカウント連携ガイド

本章では、外部自動化ツール（Make / Zapier 等）を使用して、Knowledge Index と「LINE 公式アカウント」を連携させる手順を解説します。この連携により、ユーザーが普段使い慣れた LINE アプリから直接 AI ボットに質問を投げかけ、回答を受け取ることが可能になります。

14.1 連携の概要と前提条件

外部ツール（本ガイドでは **Make** を例に解説）を利用して、LINE サーバーと Knowledge Index API の間の中継（メッセージの中継ぎ）を行います。

処理の流れ: 1. ユーザーが LINE 上でテキストメッセージを送信 2. LINE Messaging API から Make へ Webhook（受信通知）が飛ぶ 3. Make が Knowledge Index API (`POST /api/v1/chats`) にテキストを投げる 4. Knowledge Index の回答を Make が受け取り、LINE の Reply API を使ってユーザーへ返す

コストについて: LINE からの「ユーザーからのメッセージに対する自動返信（Reply）」は、何度繰り返しても LINE 公式アカウントの月間無料配信数（メッセージ数）を消費せず、**実質 0 円**で運用可能です（コミュニケーションプランの場合）。ただし、外部ツール（Make 等）のタスク実行数が増加するとツール側の有料プランが必要になる場合があります。

14.2 Step 1: LINE 公式アカウントとチャネルの開設

まずは、受け皿となる LINE 側の公式アカウントと、API 連携用の「チャネル（Channel）」を開設します。

1. **LINE 公式アカウントの作成**
   - LINE Business ID にログイン（または新規登録）します。
   - 「LINE 公式アカウントの作成」より、アカウント名（ボットの名前）や業種などを入力してアカウントを作成します。
2. **LINE Developers コンソールへの登録**
   - LINE Developers コンソールにアクセスし、先ほどの Business ID でログインします。
   - 「新規プロバイダー」を作成（会社名などを入力）し、その中に「新規チャネル作成」ボタンから **Messaging API** を選択してチャネルを作成します。
   - 作成時に、先ほど作った LINE 公式アカウントと紐付けます。
3. **必要な認証キーの取得**
   - 作成したチャネルの設定画面を開きます。
   - **「チャネル基本設定」**タブで**チャネルシークレット**（`Channel Secret`）を控えます。
   - **「Messaging API 設定」**タブで**チャネルアクセストークン**（`Channel Access Token`）を「発行」ボタンから生成して控えます。
   - 同じく「Messaging API 設定」タブ内の「Webhook の利用」を **オン**にしておきます。（※のちほど Make で生成される URL をここに登録します。）
4. **自動応答のオフ設定**
   - 公式アカウントマネージャー（管理画面）の「応答設定」に行き、「あいさつメッセージ」と「応答メッセージ」を**オフ** にし、「Webhook」を **オン**に変更します。（AI 以外が勝手に返信するのを防ぐためです）

14.3 Step 2: Knowledge Index 側の準備

連携させたい自社のテナント（データ）に対する API キーを発行します。

1. Knowledge Index の管理画面にテナント管理者としてログインします。
2. 「API キー管理」メニューを開き、ボット連携用の新しい API キー（例: `LINE-Bot-Key`）を発行し、この文字列を控えておきます。

14.4 Step 3: Make での連携シナリオ作成

Make（旧 Integromat）を利用して、実際のメッセージ中継シナリオを作成します。

1) LINE Webhook の設定

1. Make にログインし、「Create a new scenario」をクリックします。

2. 最初のモジュールとして **LINE** アプリを選び、`Watch Events`（Webhook の受信）アクションを選択します。
3. 接続（Connection）の追加画面で、Step 1 で取得した `Channel Access Token` を入力して接続を確立させます。
4. 設定を保存すると「Webhook URL」が表示されますので、それをコピーします。
5. **LINE Developers コンソール** の「Messaging API 設定」画面に戻り、コピーした URL を `Webhook URL` に貼り付けて「検証」ボタンを押し、成功することを確認します。

2) Knowledge Index API への送信（HTTP Request）

1. Make のシナリオで次のモジュールとして **HTTP** を追加し、`Make a request` を選びます。
2. 以下の内容を設定します。
   - **URL**: `https://{あなたの KI のドメイン}/api/v1/chats/default/messages` ※ `default` の部分はユーザー ID などに置き換えることも可能ですが、汎用設定の場合は固定で構いません。
   - **Method**: `POST`
   - **Headers**:
     - Key: `X-API-Key` / Value: `（Step 2 で取得した KI の API キー）`
     - Key: `Content-Type` / Value: `application/json`
   - **Body type**: `Raw`
   - **Content type**: `JSON（application/json）`
   - **Request content**:

     ```
     {
       "content": "{{1.events[].message.text}}"
     }
     ```

     ※ `{{1.events[].message.text}}` の部分は、LINE モジュールから受け取ったユーザーの入力テキストを動的にマッピングします。
   - **Parse response**: `Yes`

3) LINE への返信（Reply Message）

1. 最後のモジュールとして再度 **LINE** アプリを追加し、`Send a Reply Message` アクションを選びます。
2. 以下の内容を設定します。
   - **Connection**: `（1）で作成した接続`
   - **Reply Token**: `{{1.events[].replyToken}}`（LINE モジュールから受け取ったトークンをマッピング）
   - **Messages**:
     - **Type**: `Text`
     - **Text**: `{{2.data.message.content}}`（HTTP モジュールから返ってきた KI の回答テキストをマッピング）

4) シナリオの稼働

1. 左下の「Run once」をクリックしてシナリオを待機状態にします。
2. スマホの LINE から、作成した公式アカウント宛に「○○について教えて」とテスト質問を送ります。
3. KI の回答が LINE 上に返信されれば成功です。シナリオを「ON（稼働状態）」に切り替えます。

これで連携構築は完了です。LINE からいつでも Knowledge Index のナレッジを利用できるようになります。

第 15 章：テナント管理者ガイド

テナント管理者（TENANT_ADMIN）として、組織の Knowledge Index 環境を構築し、日々の運用を円滑・安全に行うための実践的なガイドラインです。他の章（機能別の操作説明）と併せて、運用のベストプラクティスとしてご活用ください。

15.1 初期セットアップと環境構築フロー

システムに初めてログインしたテナント管理者は、以下のフローで組織用の環境を整えます。

1. **AI モデルと API キーの検討・設定**
   - デフォルトで提供されるシステム共通の API を利用するか、自社（自テナント）の契約による独自 API キーを利用（BYOK: Bring Your Own Key）するかを決定します。
   - 独自の API を利用する場合は、「テナント設定」＞「AI モデル設定」から各種プロバイダー（OpenAI, Anthropic 等）の API キーを登録し、使用するモデルを選択してください。
2. **初期ユーザー（管理者・運用者）の招待**
   - 「第 7 章：ユーザー管理」を参考に、初期のコンテンツ登録作業を行うメンバー（OPERATOR）や、共同管理者（TENANT_ADMIN）のアカウントを発行します。
3. **初期ナレッジの投入**
   - 組織の基本となる規程類や業務マニュアルなどをシステムにアップロードし、インデックス化を完了させます（「第 4 章：コンテンツ管理」参照）。
4. **ウィジェット（チャット画面）のデザイン設定**
   - コーポレートカラーやロゴに合わせて、チャットアイコンやカラーテーマを設定します（「第 6 章：Widget 連携」参照）。

15.2 ユーザーロールの運用指針

システムには以下のロールがありますが、組織規模に応じて権限を適切に分離することがセキュリティの要となります。

- **TENANT_ADMIN（テナント管理者）**：1〜3 名程度に限定し、API キーの管理やプラン変更、ユーザー追加など特権的な操作を担当します。
- **OPERATOR（運用者）**：各部署で実際にマニュアルなどのドキュメントを更新する担当者に付与します。コンフィグ変更はできませんが、コンテンツの追加・削除やチャット履歴の分析が可能です。
- **AUDITOR（監査者）**：情報システム部門やコンプライアンス部門など、コンテンツの改ざんや不正アクセスを確認する担当者に付与します。設定変更は行なえず、監査ログや統計の閲覧のみが可能です。

15.3 継続的な AI 精度の向上（運用サイクル）

Knowledge Index を導入後、AI の回答精度を高めていくための PDCA サイクル（運用プロセス）です。

1. **会話履歴の分析（Check）**

- 管理画面の「会話分析」メニューから、ユーザーが実際にどのような事を聞いているかを定期的に確認します。
- 特に AI が**「情報が見つかりません」と答えている質問**や、 / 評価で　**が押されている回答**を重点的にピックアップします。

2. **原因の特定（Analyze）**
   - ドキュメント自体が存在しないのか、専門用語が多すぎてユーザーの質問意図（クエリ）とマッチしていないのかを分析します。
3. **ナレッジの改善（Act）**
   - **コンテンツがない場合**: 必要な情報をまとめた PDF やテキストを作成し、新規登録します。
   - **検索に引っかからない場合**: 既存のドキュメントの表現を見直すか、同義語を併記して再アップロード（再インデックス）します。

15.4 セキュリティ監視と監査ログの活用

テナントの安全性を保つため、管理者は定期的に以下のセキュリティチェックを実施することが推奨されます。

- **監査ログ（Audit Log）の定期確認**:
  – 月に 1 回程度、「監査ログ」画面から直近のアクセス履歴などを CSV でエクスポートし、異常な操作（不自然な時間帯の大量ダウンロードや、見覚えのないユーザー追加など）がないか確認します。
- **利用状況（トークン消費）の監視**:
  –「統計・分析」ダッシュボードからトークン消費量を確認し、予算管理（または想定以上の激しい利用による異常検知）を行います。

Hints: 各画面の具体的なボタンの操作や仕様の詳細は、本マニュアルの該当する機能章（第 4 章、第 7 章、第 8 章など）を併せてご参照ください。

第 16 章：付録 (Appendix)

A. 主要な技術用語集

用語	意味
RAG	Retrieval-Augmented Generation（検索拡張生成）。自社データを検索し、その「参考資料」をもとに AI に回答を作らせる仕組み。
LLM	Large Language Model（大規模言語モデル）。GPT-4 や Claude 3 などの生成 AI の心臓部。
ベクトル（埋め込み / Embedding）	文章の意味を数百〜数千次元の数値配列に変換したもの。「意味の近さ」を計算できるようにする技術。
チャンク (Chunk)	長い文章を短く（例えば 1000 文字などに）切り分けた断片。RAG ではチャンク単位で検索を行う。
pgvector	PostgreSQL データベース上でベクトル検索（類似度計算）を高速に行うための拡張機能。

用語	意味
トークン	LLM がテキストを処理する際の最小単位。英語なら 1 単語≒ 1〜2 トークン、日本語なら 1 文字≒ 1〜3 トークン程度。課金の基準となる。

B. 対応 LLM プロバイダーとモデル一覧表

※ 本一覧は *2025* 年時点のものです。システムアップデートにより順次追加・変更される可能性があります。

OpenAI

モデル名	主な用途・特徴
gpt-4o	最高性能。速度と精度のバランスが良く、ほぼ全ての用途で推奨。
gpt-4o-mini	非常に高速で低コスト。大量の定型質問向け。
gpt-4-turbo	長文対応と高い推論能力。

Anthropic

モデル名	主な用途・特徴
claude-3-5-sonnet	高いコーディング能力と正確な文章読解力。推奨モデルの一つ。
claude-3-opus	最高精度の推論能力。複雑な分析が必要な場合に適する。
claude-3-haiku	最軽量クラスで瞬時に応答を返す。

Google

モデル名	主な用途・特徴
gemini-1.5-pro	超大容量のコンテキストウィンドウ（〜200 万トークン）を持ち、大量の背景情報を処理可能。
gemini-1.5-flash	速度重視の軽量モデル。

本書の内容に関するご質問、およびサービスの導入サポートにつきましては、サポート窓口までお問い合わせください。