RAGを設定して検索精度を上げる

社内LAN上のPCからOpen WebUIでチャットできるようになったが、業務に活用するにはもう一工夫したい。そこで登場するのが「RAG」（検索拡張生成）だ。

RAGはLLMの生成能力に、外部知識ベースからの検索を組み合わせたアーキテクチャのこと。LLMはカットオフ（学習データの締め切り日）があり、それ以降の情報は持っていない。また、社内の契約書・規定・製品情報といった企業固有の知識は最初から学習されていない。これらの情報をRAGに登録しておくことで、「○○の契約書の内容は？」といった質問に正確に答えられるようになる。

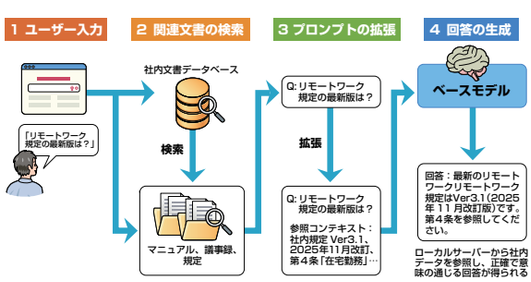
RAGの処理の流れは次のようになる。

①ユーザーが質問（クエリ）を入力
②事前に登録したドキュメントから関連情報を検索・抽出
③抽出した情報をコンテキストとしてLLMに供給、回答を生成

これらの処理により、LLMの事前学習だけではカバーできない最新情報や企業固有の知識に基づいた正確な回答が得られる。また、検索元のドキュメントが明示されるため回答の根拠が追跡可能となり、ハルシネーション（AIが事実と異なる情報を生成する現象）の抑制にも寄与する。

GoogleのNotebookLMは、このRAGの手法で広く知られているが、外部に機密性の高いデータをアップロードできない企業も多いだろう。ローカルLLMサーバー、Open WebUI、RAGを組み合わせれば、社内で完結できるというわけだ。

RAGの仕組み



社内規定に合致したデータをAIに参照させることで、生成される回答の精度が向上する

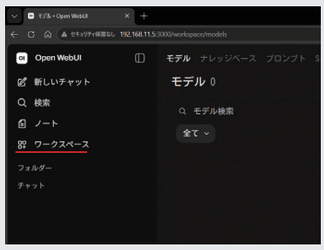
作成したローカルLLMサーバーのパフォーマンス

今回構築したローカルLLMサーバーにて、モデルに26Bの「Gemma 4 26B-A4B-it AWQ INT4」を用いた際の性能を、自作のベンチマークテストで測った結果は**表7**のとおり。

4並列の同時リクエストであっても、1人あたり約39tok/sを維持し、実効スループットは155.8tok/sを達成、多くのユーザーが「十分速い」と感じる水準となった。利用者10人規模でも、一般的なチャット・文書要約中心であれば、現実的に共用できる性能といえる。

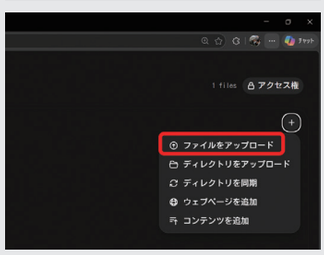
ナレッジベースの作成とRAGの利用

1 ナレッジベースの作成していく。最初に「ワークスペース」を選択



2 作成パネルが表示されたら、各項を次の画面のように設定し、「ナレッジベース生成」を押す

3 右上の＋ボタンを押してコレクションを追加する。今回は「ファイルをアップロード」でCSVファイルを登録する



参考 アップロードしたRAG用CSVファイルの中身の一部。今回は各カテゴリに10種ずつの質問と回答を定義した、架空のデータを用意した

category	question	answer
サービス	提供サービスを教えて	AIボット・DataVision・データ分析・MLトレーニング・カスタム開発最新版
会社情報	本社はどこにありますか	東京都渋谷区神南1-2-3
製品_DataVision	DataVisionのバックアップ体制は	毎日自動バックアップ・90日間保持・手動復元も可
製品_DataVision	DataVisionの価格は	リアルタイムデータ可視化・予測分析・ダッシュボード共有
決済・請求	領収書の発行は	オンライン申請でPDF自動発行・再発行は1人まで無償
製品_AIBot	AIボットの価格プランは	月3,000円〜/月9,800円〜/月24,800円〜
製品_DataVision	DataVisionの接続先データソースは	MySQL・PostgreSQL・BigQuery・Snowflake・Google Sheets連携
製品_DataVision	DataVisualizationのリアルタイム対応は	1秒以下で更新レイテンシでリアルタイム対応
決済・請求	請求書払い可能か	可能・請求書払い専用の請求アドレスに、当月締め翌月末日払い、要審査
利用規約	データアクセス方法が、	アカウント設定>データエクスポートから
サービス	提供サービスを教えて	AIボット・DataVision・データ分析・MLトレーニング・カスタム開発最新版
製品_DataVision	DataVisionのAPI連携は	GraphQL RESTフルサポート・Webhook通知機能あり
会社情報	代表取締役の名前は何ですか	山田太郎氏
研修プログラム	研修料金は	1コース5万円〜 / 年間パスポート18万円

4 「+新しいモデル」を押し、新しいモデルの作成画面で、モデル名、モデルID、基本モデル、説明を記載。ナレッジベースを選択し、保存して作成すれば準備は完了

5 RAGの使い方は、「セッション全体で有効にする」方法と、「特定の質問にのみ使う」方法がある。今回は、モデルを選んでセッション全体でRAGを有効にした

表7 ローカルLLMサーバーのパフォーマンス

並列数	個人	実効	並列効率
1	50.17tok/s	50.2tok/s	100%
2	42.86tok/s	85.7tok/s	85.40%
3	38.08tok/s	114.2tok/s	75.90%
4	39.02tok/s	155.8tok/s	77.60%

日本AMD https://www.amd.com/ja/

TEXT：芹澤 正芳（p.2 ～ p.3）、西川 和久（p.3 ～ p.8）
編集：PCWatch編集部
発行：株式会社インプレス