

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。

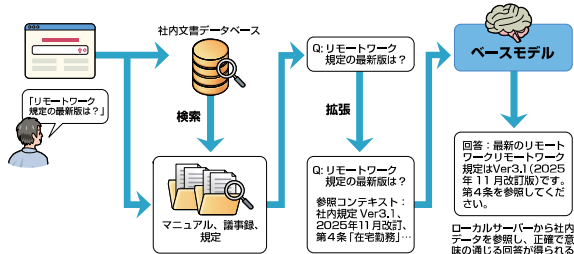
RAGの処理の流れは次のようになる。

- 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

- 1 ユーザー入力
- 2 関連文書の検索
- 3 プロンプトの拡張
- 4 回答の生成



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



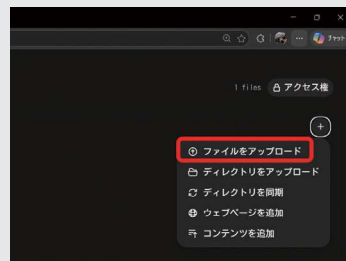
2

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	サービスとは何ですか?	AIサービスとは、AI技術を活用して、様々な業務を自動化・効率化するサービスです。
製品情報	最新の製品情報は何ですか?	最新の製品情報は、当社のウェブサイトからご確認ください。
お問い合わせ	お問い合わせの受付時間は何時間ですか?	お問い合わせの受付時間は、毎日午前9時から午後5時までです。
お問い合わせ	お問い合わせの受付方法はありますか?	お問い合わせの受付方法は、メール、電話、Webチャットがあります。
お問い合わせ	お問い合わせの受付場所はありますか?	お問い合わせの受付場所は、当社の本社、支店、営業所があります。
お問い合わせ	お問い合わせの受付窓口はありますか?	お問い合わせの受付窓口は、当社のカスタマーサポートセンターです。
お問い合わせ	お問い合わせの受付窓口の電話番号はありますか?	お問い合わせの受付窓口の電話番号は、03-XXXX-XXXXです。
お問い合わせ	お問い合わせの受付窓口のメールアドレスはありますか?	お問い合わせの受付窓口のメールアドレスは、support@xxxx.co.jpです。
お問い合わせ	お問い合わせの受付窓口のWebチャットはありますか?	お問い合わせの受付窓口のWebチャットは、当社のウェブサイトからご利用いただけます。

4

「+新しいモデル」を押して、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

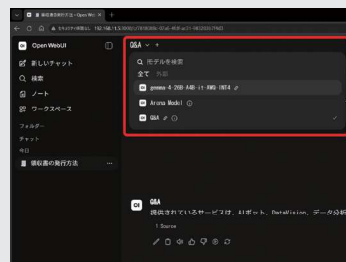


表7 ローカルLLMサーバーのパフォーマンス

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%