第4章 RAG効果検証から精度向上アプローチまでいろいろな方法を実験で比較

ローカルLLMを最新情報に対応させてみる

山田 晃嗣

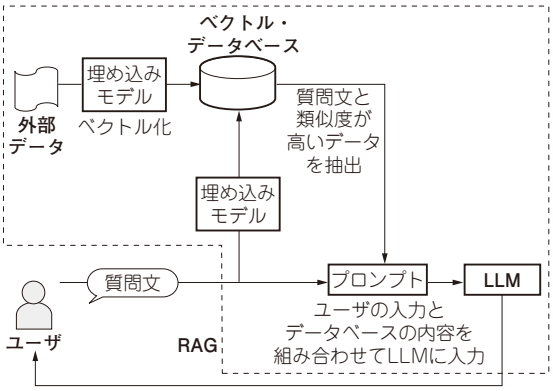

図1 RAGの概要

多くのLLM（Large Language Model，大規模言語モデル）を使用したサービスはクラウド型です．マシンの性能を気にせず利用可能である一方，最新情報や社内用語などのローカル情報は学習できていないため正しく答えられない可能性が高いです．またセキュリティの観点から社内情報など機密性の高い内容を入力することは推奨されていません．

機密性の高い内容を，セキュリティを気にせずLLMに扱わせる方法として，ローカルLLMの使用が考えられます．ローカルLLMであれば処理がローカル環境で完結するため，セキュリティ面での課題を解決できます．しかし，最新情報に対応していない点はクラウド型LLMと共通しており，ファインチューニングなどで対応するには学習コストが膨大となります．これらの課題を解決する手法として，RAG（Retrieval-Augmented Generation，検索拡張生成）（**図1**）があります．これはLLMを再トレーニングすることなく，特定の情報に対応できるようにする手法であり，LLMに学習データ以外の辞書を与えるようなものです．

そこで本稿では，ローカル環境のRAGの効果を検証するために，LLM単体の場合，LLM＋資料など外部ファイル受け渡し機能を使った場合，そしてRAGの場合について，それぞれ実装し，同じ質問をしてその応答を見てみます[注1]．

表1 クラウド型LLMとローカル型LLMの比較

	クラウド型	ローカル型
セキュリティ	×	○
コスト	△ （従量課金制が多く予測が困難）	○ （マシン購入が必要だが，その後は電力コストのみ）
性能	○	△
メンテナンス性	○ （原則不要）	×
カスタマイズ性	△	○
導入の手間	○	×

● ローカルLLMのメリットと注意すべき点

クラウド型LLMとローカル型LLMにはそれぞれにメリットとデメリットがあります．**表1**にその一部を示します．実際にLLMを使用する際には，用途や環境に合わせて適切に手法を選択することが重要です．

また，クラウド型でもローカル型でもLLMにはハルシネーション（幻覚）という問題があります．LLMを使用する際には，必ずしも正しい回答が得られるとは限らないということを常に留意した上で使用することを心がけてください．

実験環境の構築

本稿では，LLMをローカル環境で動作させます．ローカル環境で実行できるLLMの性能はPCの性能に依存します．また，GPUを搭載しているPCを使用することで高速な推論が可能となります．ただし，CPUのみでも実用的な速度で推論できる小規模なモデルも存在するため，実際に使用するモデルは環境に応じて調整することが可能です．筆者の実験環境を**表2**に示します．

注1：本稿に記載するデータは執筆時点（2025年7月）のものです．