

RAGの動作イメージと今注目される理由

佐藤 聖

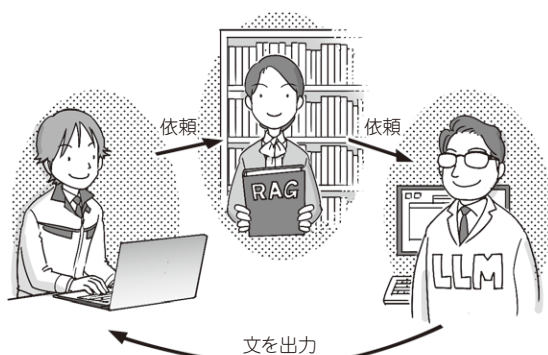


図1 構築するローカルのRAGのイメージ

本章では、企業でも注目され導入が進んでいるRAG (Retrieval-Augmented Generation) に注目します。この章では、まずRAGとは何か、RAGの必要性について説明します。

RAGの動作イメージ

● 例えると図書館と司書のコンビ

RAGは、まるで巨大な図書館と司書のコンビのように動作します(図1)。まず、利用者(クエリ)が「このテーマの本を教えてほしい」と司書(Retrieval)に尋ねると、司書は膨大な蔵書(ベクトル・データベース)から関連性の高い本(ドキュメントやチャンク)を素早く取り出します。次に、その本を元に研究者(Generation/LLM)が要点をまとめたり、引用しながら新たな文章(回答)を執筆したりします。

● LLMとの違い

従来のAIは、情報の海から手当たり次第に文章を紡ぐ無秩序な書き手のようでした。専門性の高い分野の回答が得られにくく、回答されたとしても内容についてのファクト・チェックが必要でした。

RAGは、司書が適切な資料を選び出すように、参

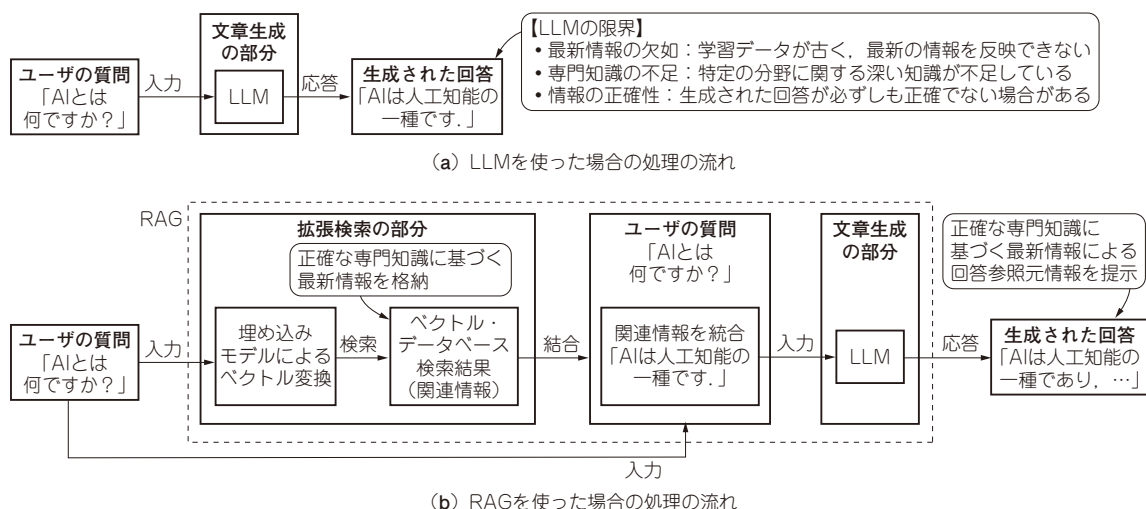


図2 LLMとRAGの比較

LLMの限界をRAGが解決