第1章 直近の出来事も社内ローカル情報も生成精度UP

RAG注目の理由＆仕組み

山田 薫

Retrieval-Augmented Generation（RAG）とは情報検索機能を大規模言語モデル（LLM：Large Language Model）などの言語モデルと連携させる仕組みです．これを活用することで，ChatGPTなどの汎用のLLMサービスでも学習していない知識，例えば直近の出来事や社内ルールなどローカルな情報も使って，テキストを生成させることができます（**図1**）．2022年のChatGPTの公開以降から徐々に話題となっており，2023年によく耳にするようになりました．

本稿では，RAGの概要，具体的な構成や処理方法と徐々に深掘りしていきます．

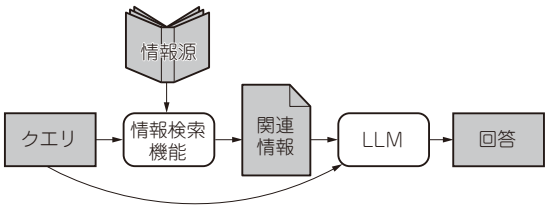

図1　RAGの概要図
四角，本型，書類型のオブジェクトはデータを，角丸の四角いオブジェクトは処理内容を示す．ユーザからのクエリを元に情報検索を行い，検索された関連文章とクエリをLLMに渡すことによって，LLMが情報源の知識に基づいたテキストを生成できる

RAGの始まりはChatGPTよりも前

RAGは，2020年にLewisらによって提案[1]されました．ChatGPTの登場が2022年なので，それより前に提案されていました．この後，世の中に論文の内容が浸透したことに加えて，ChatGPTなどの高性能なLLMが数多く登場したことで，昨今大きく話題となったと考えられます．

RAGは2023年によく耳にするようになりました．例えば大手クラウド・サービス・プロバイダの例を挙げると，Amazon Web Services（AWS）ではAmazon Bedrockのナレッジ・ベース，Microsoft AzureではAzure OpenAI On Your Data，Google CloudではVertex AI Agent Builderといったサービスが登場しており，これらでRAGを構築し活用することができます．

RAGとは…LLMへのプロンプトを補完する仕組み

通常，LLMの入力には指示や質問などのテキスト（プロンプト）を与えます．このプロンプトの中に検索によって得られた情報を加える仕組みがRAGです（**図1**）．RAGは，日本語では検索拡張世代，検索拡張生成と訳されることもありますが，RAGという表現が一般的になってきています．

RAGが今注目の理由

LLMは通常，学習によって知識を獲得するので，LLM単独でも学習で得られた知識に基づくテキストを生成することができます．ではなぜRAGが今，注目されているのか，その理由を説明します．

● 理由①…学習はコストがかかる

LLMに限らずニューラル・ネットワークを使用したモデルに言えることですが，モデルは学習フェーズと推論フェーズを持ちます．学習フェーズでは，大量のデータセットを使って学習を行い，モデルの重み（パラメータ）を更新します．このパラメータ更新によってLLMは知識を獲得します．一方で，チャットボットやQAシステムなどの実際にモデルを活用するときは推論フェーズと言います．推論フェーズでは学習フェーズで学習済みのモデルを使用し，パラメータを更新するような学習を行いません．RAGはこの推論フェーズでの仕組みです．

RAGは推論フェーズの仕組みにも関わらず，あたかも学習で新しい知識を獲得しているかのように直近の情報やローカル情報に基づくテキストを生成します．そのため，学習によってこれらの知識を獲得させるよりもコストを削減できる可能性があり，注目されています．