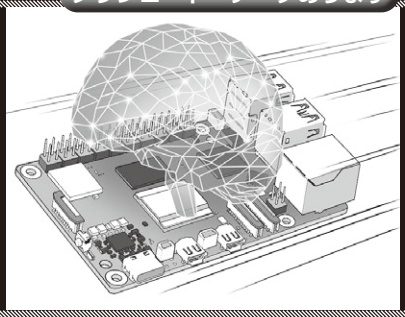
ダウンロード・データあります

リソースの限られた環境での実行性能を徹底解析！

ラズパイの限界に挑戦！ ローカルLLM 動作検証レポート



第2回 RAGを作る①…データベースに入れるデータの前準備

ご購入はこちら

澁谷 慎太郎

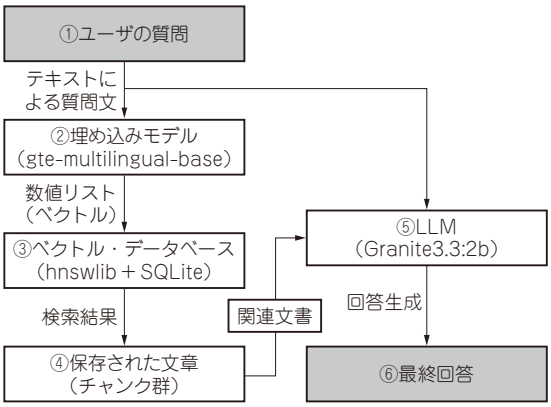


図1　今回から作成するRAGの質問から最終回答までのフロー

前回までは，ラズベリー・パイでどれくらいのサイズのLLM（Large Language Models．大規模言語モデル）がどの程度の速度で動くものかを見てきました．テストしたプロンプトも「今日の東京の天気は、どうですか？」という単純な短文だったので，モデルのサイズさえ適切に選べば，インタラクティブな回答を得ることができました．しかし，最新の情報となるとLLMごとに独特な回答が戻ってきました．

今回から，LLMが学習していない情報を補完する方法の第1弾のテストとして，RAG（Retrieval-Augmented Generation）をラズベリー・パイ5上に構築します（**図1**）．その際，ローカルLLMとしてGranite3.3[注1]を使用し，学習していないラズベリー・パイ5の商品情報をRAGが正しく回答できるかを見ます．今回は，ユーザの質問に関連する情報を集めたデータベース・ライブラリの準備を行います．

なお，本連載では，可能な限り学術的な解説を避け，平易な日常語での解説を心がけ，読者の皆さんがなんとなく理解して，使えるようになることを念頭に執筆していきます．

● RAGはローカルLLMの価値を上げる

ラズベリー・パイ5でも，モデル・サイズを選べば，ローカルでLLMを動かすことができます（連載第1回，2026年1月号）．例えば，Granite3.3:2bにラズベリー・パイ5の仕様について質問をすると，ラズベリー・パイ5はまだ公開されていないと回答されました（**リスト1**）．Granite3.3の知識カットオフは，2024年4月とされています．しかし2023年に発売開始されているラズベリー・パイ5については学習していなかったようです（**リスト1**）．さらに，回答では，2021年10月が最新とも言っています（ここではその時期は議論の対象外で，モデルがラズベリー・パイ5を学習しているかいないかが重要）．

このようにLLMは，モデルをトレーニングした時点以降の情報や，公開されていない情報は学習されていません．LLMでそれらの情報を使いたい場合は，新たに情報を補完する必要があります．

その解決策の1つとして，RAG（Retrieval-Augmented

注1：IBMが開発した軽量で高効率なオープンソース言語モデル．

リスト1　Granaite3.3:2b（RAGなし）にラズベリー・パイ5の仕様を尋ねた結果

```
日本語で質問した場合
>>> Raspberry Pi 5の主要なスペックは？
Raspberry Pi 5 は、まだ公開されておらず、完全な仕様が公表されていません。しかし、過去の発行品の傾向と現在のプロジェクションを考慮すると、
Raspberry Pi 5 が主な特徴を持つ可能性があります。（以下省略）

英語で質問した場合
>>> >>> What are the key specs of the Raspberry Pi 5?
As of my last update in October 2021, there has been no official release of a Raspberry Pi model named
"Raspberry Pi 5." The latest models in the Raspberry Pi series are the Raspberry Pi 4 Model B and the newer
Raspberry Pi Compute Module 4.（以下省略）
```