

ラズパイにローカルLLM 実行環境を構築する

澁谷 慎太郎

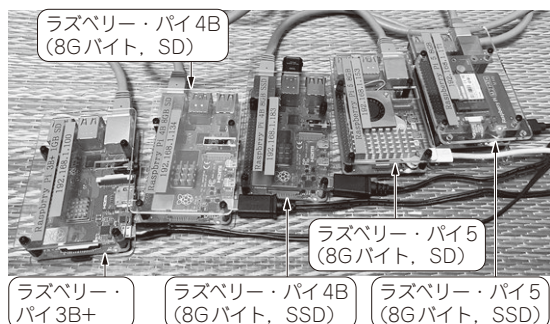


写真1 ラズベリー・パイにLLMを載せられる？実用になる？を
実験する

ラズベリー・パイ 3B+	ラズベリー・パイ 4B	ラズベリー・パイ 4B
1Gバイト + 128Gバイト SDカード	8Gバイト + 512Gバイト USB SSD	8Gバイト + 128Gバイト SDカード
ラズベリー・パイ 5	ラズベリー・パイ 5	
8Gバイト + 512Gバイト NVMe SSD	8Gバイト + 128Gバイト SDカード	

図1 テストに使う5つのハードウェア構成

ChatGPTやClaude, Gemini, CopilotなどブラウザやAPIを介して使う生成AIは、手軽に使えるのでよく使っている方も多いかもしれません。これらはクラウド・ベースのサービスである以上、入力した情報が学習に使われるかどうかにかかわらず、外部のサーバに送信されることは避けられません。そのため、特に業務で利用する際には注意が必要です。一方、ローカル言語モデルであればインターネット上にデータを流す必要がなくなります。

今回は、ローカルLLM(大規模言語モデル)をラズベリー・パイ上に構築してどれくらい実用になるのかを実験します(写真1)。

実験項目

実際のアプリケーションとして使えるようにするには、RAG(検索拡張生成)やMCP(Model Context Protocol)まで広げて検討する必要がありますが、今回はLLM単体について行います。実験項目は次の通りです。

実験① OllamaにモデルをPullするのにかかる時間

実験② Ollamaのモデルのロードから、プロンプトを入力し、回答が終わるまでの一連のプロセスにかかる時間

実験③ モデルによる日本語の理解の具合

実験環境の検討

● 5種類のラズベリー・パイで比較

今回テストに使用するのは、図1のラズベリー・パイ5/4/3を利用した5種類の構成とします。

ラズベリー・パイ5には、拡張ボードとしてAI HatやAI Kitがラインナップされていますが、これらに使用されているAIアクセラレータHailo-8Lは、CNN/画像処理向けの処理が得意で、LLMに求められるTransformerベースのAttention演算(QKVの行列積や大規模埋め込みなど)などの大規模行列演算は苦手とされています。また、Ollamaが直接サポートしていないので、今回は対象外としました。ラズベリー・パイには、標準搭載のGPUがありますが、同じ理由から使用しません。GPUには最小限のメモリ割り当てとして、最大限CPUにメモリを割り当てることにします。

● ハードウェア環境の違いで性能にどの程度差が出るか見てみる

コンピューティング環境の違いとして、大きく、CPU、メモリ、ストレージ、ネットワークがあります(表1)。CPUの違いをラズベリー・パイ5/4B/3B+で