

## 実験⑤…生成AI

中村 仁昭

話題のLLMを  
ラズパイ5で動かしてみる

## ●果たして実用度はどのくらい？

最近では生成AI、特に大規模言語モデル (Large Language Models, LLM) に注目が集まっています。

筆者は組み込み環境でさまざまなAIを試しています。今回はラズベリー・パイ5でLLMを動作させて、どのくらい実用度があるのかを実験してみます。

LLMは、2023年の段階で、ラズベリー・パイ4や Jetson Xavier NX (エヌビディア) などでも動作しています。ラズベリー・パイ5ではどのぐらいの速度で動作するのか、さらに速度向上が狙えるのかについて試してみました。

## ●動作環境

オープンソースのLLMとしては、Llama2(メタ) や、Mistral (Mistral AI) が有名です。オープンソースLLMをベースに、日本語で追加学習させた日本語LLMも複数発表されているので、さまざまなモデルを試すことができます。

今回は、ラズベリー・パイ5で日本語LLMを動作させて処理速度を中心に実験するため、SOTA (State Of The Art, 現時点での最先端レベルの性能) は狙わず、実績のあるLlama2系のモデルを採用しました。環境構築の容易さから、C++で推論させる llama.cpp で動作させてみます。

実験の前に…  
オープンソースLLMの基礎知識

## ●オープンソースなLLM「Llama2」

## ▶より少ないパラメータでChatGPT-3.5の性能に匹敵

Llama2は、2023年7月にメタ社が公開したオープンソースLLMです(図1)。より少ないパラメータでChatGPT-3.5に匹敵する性能を持つとされ、公開当初から注目されています。オープンソースなので、Llama2をベースにした日本語LLMや専用ドメインに特化したLLMなども登場しています。また、Llama2は商用利用

Discover the power of  
Llama

Democratizing access through an open platform featuring AI models, tools, and resources to give people the power to shape the next wave of innovation.

Licensed for both research and commercial use

Download models

図1(1) ChatGPT-3.5に匹敵する性能を持つオープンソースLLM「Llama2」

オープンソースなので、Llama2をベースにした日本語LLMや専用ドメインに特化したLLMなども登場している

もできるので、自社製品に組み込むことも可能です。

## ▶日本語対応LLMも開発・リリースされている

Llama2をベースにした日本語LLMは幾つかリリースされています(図2)。2023年8月に東京大学発のベンチャ企業ELYZAから発表されたELYZA-japanese-Llama-2-7bや、2023年12月に東京工業大学と産業技術総合研究所の研究チームからリリースされたSwallowが有名です。ELYZAは7B(70億パラメータ)と13B(130億パラメータ)が、Swallowは7B、13B、70Bがそれぞれリリースされています。

手元で量子化した7Bモデルを試すと、ELYZAが問題なく応答を返すのに対し、Swallowはハルシネーション(事実に基づかない情報や実際には存在しない情報を生成する現象)が多発しました。そのため筆者は現在、ELYZAをメインに使用しています。Swallowは2023年12月時点での日本語LLMでは最高性能だったので少し残念ですが、量子化が影響している可能性もあります。