

ローカルLLMの実行性能を測定する

澁谷 慎太郎

前章までに、本実験の環境設定とネットワーク状況の確認まで終わりました。本章ではいよいよ実験に入ります。

実験①… モデルpullのパフォーマンス比較

● 実験方法

モデルをpullする時間をサイズの違う複数のモデル(表1)を順番にpullして時間を計測します。

Ollamaがモデルをpullする際には、図1に示した順序で処理がされていきます。これは、単純にファイルをダウンロードするだけではないことを意味しています。計測に使用したシェル・スクリプトをリスト1に示しました。

表1
実験に使うLLM
(第1章の表2再掲)

LLM名	メーカー
Gemma3:270M	Google
Qwen3:0.6b	アリババ
Llama3.2:1b	Meta
Granite3.3:2b	IBM
Phi4-mini:3.8b	マイクロソフト

このコードでは、コマンド入力の入出力を捉えるのではなく、Ollama APIを使用してそのストリームを追う仕組みになっています。単にモデルをpullしたい場合は、

```
ollama pull <モデル名:タグ>
例) ollama pull Granite3.3:2b
```

と入力します。

● 実験結果

図2に測定結果を示します。pullの一連の流れに出てくる5つのステップごとに時間表示をしています。

図1
モデルのpullの一連の流れ

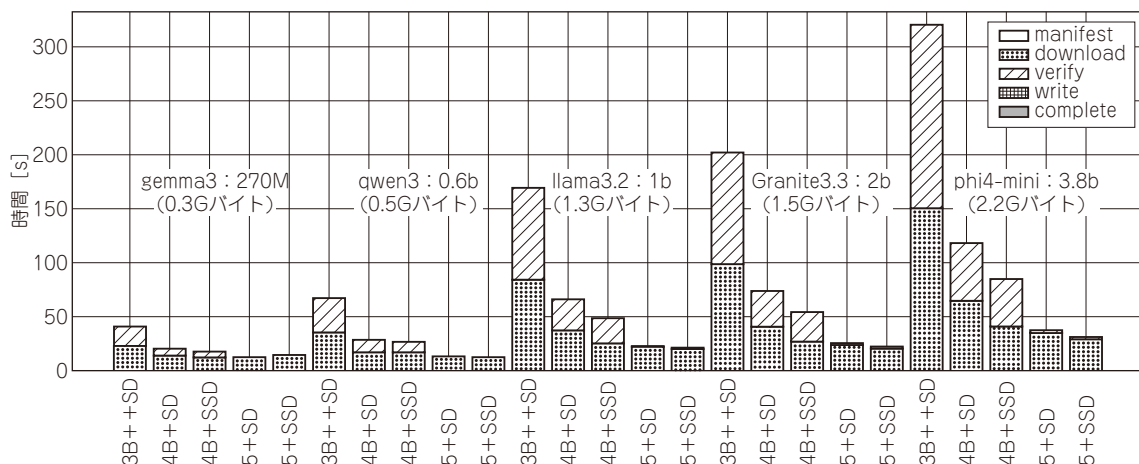
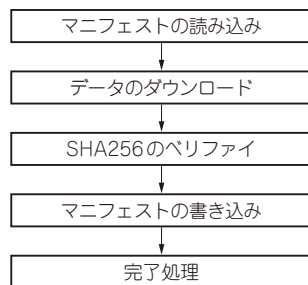


図2 各種モデルのpullにかかった時間