llama.cpp を試す

`llama.cpp` でモデルを実行しようとすると、以下のようなエラーが発生することがあります：

```
(py311) lzwjava@Zhiweis-MacBook-Air llama.cpp % ./main -m models/7B/Phi-3-mini-4k-instruct-q4.gguf
main: build = 964 (f3c3b4b)
main: seed  = 1737736417
llama.cpp: loading model from models/7B/Phi-3-mini-4k-instruct-q4.gguf
error loading model: unknown (magic, version) combination: 46554747, 00000003; is this really a GGML file?
llama_load_model_from_file: failed to load model
llama_init_from_gpt_params: error: failed to load model 'models/7B/Phi-3-mini-4k-instruct-q4.gguf'
main: error: unable to load model
```

このエラーは通常、`llama.cpp` のインストールまたはモデルファイル自体に問題があることを示しています。

一般的な解決策は、Homebrew を使用して `llama.cpp` をインストールすることです：

```
brew install llama.cpp
```

これにより、互換性のあるバージョンのライブラリがインストールされます。

以下は役立つリソースです：

- Hugging Face GGML Models
- llama.cpp GitHub リポジトリ
- ggml GitHub リポジトリ
- Ollama
- Ollamac