嘗試 llama.cpp

當嘗試使用模型運行 `llama.cpp` 時，您可能會遇到如下錯誤：

```
(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
```

此錯誤通常表示 `llama.cpp` 安裝或模型文件本身存在問題。

一個常見的解決方案是使用 Homebrew 安裝 `llama.cpp`：

```
brew install llama.cpp
```

這確保您擁有兼容的庫版本。

以下是一些有用的資源：

- Hugging Face GGML 模型
- llama.cpp GitHub 倉庫
- ggml GitHub 倉庫
- Ollama
- Ollamac