尝试使用 llama.cpp

在尝试使用模型运行 `llama.cpp` 时，你可能会遇到如下错误：

```
(py311) lzwjava@ 李智维的 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
```

此错误通常表明 `llama.cpp` 安装或模型文件本身存在问题。

一个常见的解决方法是使用 Homebrew 安装 `llama.cpp`：

```
brew install llama.cpp
```

这确保你拥有一个兼容的库版本。

以下是一些有用的资源：

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