Trying llama.cpp

When attempting to run `llama.cpp` with a model, you might encounter an error like this:

```
(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
```

This error typically indicates an issue with the `llama.cpp` installation or the model file itself.

A common solution is to install `llama.cpp` using Homebrew:

```
brew install llama.cpp
```

This ensures you have a compatible version of the library.

Here are some useful resources:

- Hugging Face GGML Models
- llama.cpp GitHub Repository
- ggml GitHub Repository
- Ollama
- Ollamac