Probando llama.cpp

Al intentar ejecutar `llama.cpp` con un modelo, es posible que te encuentres con un error como este:

```
(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
```

Este error generalmente indica un problema con la instalación de `llama.cpp` o con el archivo del modelo en sí.

Una solución común es instalar `llama.cpp` usando Homebrew:

```
brew install llama.cpp
```

Esto asegura que tengas una versión compatible de la biblioteca.

Aquí tienes algunos recursos útiles:

- Modelos GGML en Hugging Face
- Repositorio de GitHub de llama.cpp
- Repositorio de GitHub de ggml
- Ollama
- Ollamac