Essayer llama.cpp

Lorsque vous essayez d'exécuter `llama.cpp` avec un modèle, vous pourriez rencontrer une erreur comme celle-ci :

```
(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
```

Cette erreur indique généralement un problème avec l'installation de `llama.cpp` ou avec le fichier du modèle lui-même.

Une solution courante consiste à installer `llama.cpp` via Homebrew :

```
brew install llama.cpp
```

Cela garantit que vous disposez d'une version compatible de la bibliothèque.

Voici quelques ressources utiles :

- Modèles GGML sur Hugging Face
- Dépôt GitHub de llama.cpp
- Dépôt GitHub de ggml
- Ollama
- Ollamac