IA - Chat Bot

Création d'ia similaire a chat gpt en local

- IA LLAMA

IA LLAMA

Chat Bot local

Etape 1 Télécharger le projet git et le monter

Ce petit tuto a pour but de recréer un chat bot en local sur notre pc a l'aide LLAMA une ia déja près a l'emploie et aussi forte que Chat GPT

Vous pouvez suivre les etape du github suivant ou suivre les mienne

https://github.com/ggerganov/llama.cpp

Ce tuto sera pour un linux

Téléchargement des fihcier nécessaire

```
git clone https://github.com/ggerganov/llama.cpp
cd llama.cpp
make
```

Etape 2 Télécharger le Model de l'ia

Télécharger Le model iA LLAMA avec torrent et le magnet suivant :
magnet:?xt=urn:btih:b8287ebfa04f879b048d4d4404108cf3e8014352&dn=LLaMA

> ATTENTION le model d'ia fait 235Go

avec les commande suivante pour télécharger sur arch linux le téléchargeur de torrent et lancer le téléchargerment

```
sudo pacman -S transmission-cli             / sudo apt-get install transmission-cli
transmission-cli magnet:?xt=urn:btih:b8287ebfa04f879b048d4d4404108cf3e8014352&dn=LLaMA
```

stopper le processus a la fin du téléchargement

déplacer les fichier télécharger dans le dossier model qui est dans llama.cpp

```
ls ./models
```

pour que dedans on y retrouve les fichier suivant :
65B 30B 13B 7B tokenizer_checklist.chk tokenizer.model

Etape 3 installer les dépendance

installer les dépence python

```
python3 -m pip install torch numpy sentencepiece
```

Etape 4 Préparer l'ia pour le model 7B

convertir au format ggml le fichier 7B

```
python3 convert-pth-to-ggml.py models/7B/ 1
```

quantifiter en 4bit

```
python3 quantize.py 7B
```

run the inference

```
./main -m ./models/7B/ggml-model-q4_0.bin -n 128
```

Etape 4-V4 Préparer l'ia pour le model 30B

cela va prendre bcp plsu de temps car le ficherfichier est plus gros

convertir au format ggml le fichier 30B

```
python3 convert-pth-to-ggml.py models/30B/ 1
```

quantifiter en 4bit

```
python3 quantize.py 30B
```

run the inference

```
./main -m ./models/30B/ggml-model-q4_0.bin -n 128
```

Etape 4-V5 Préparer l'ia pour le model 30B

cela va prendre bcp plsu de temps car le ficherfichier est plus gros

convertir au format ggml le fichier 30B

```
python3 convert-pth-to-ggml.py models/65B/ 1
```

quantifiter en 4bit

```
python3 quantize.py 65B
```

run the inference

```
./main -m ./models/65B/ggml-model-q4_0.bin -n 128
```

Etape 5 Parler a l'ia

lancer le chat avec l'ia

```
./chat.sh -i
```