Vous devez télécharger llama.cpp, voici un lien pour télécharger llama.cpp :

https://github.com/ggerganov/llama.cpp/releases/tag/b1606

./main -m ./models/VOTRE\_MODELE --repeat\_penalty 1.0 --color -c 7000 -i -r "User:" -f prompts/Mission/prompt\_initial\_roanne\_shorter.txt

Tapez cette commande dans le terminal en voulant dans le répertoire de llama.cpp et en remplaçant VOTRE\_MODELE par le nom de votre modèle que vous aurez au préalable téléchargé et déposé dans le répertoire “models” de llama.cpp.

Voici les modèles que nous avions téléchargé sur Hugging Face :

* French-Alpaca-7B-Instruct-quantized-q8\_0-beta.gguf
* NeuralBeagle14-French-Aplaca.Q4\_0.gguf

Les autres modèles que nous avions testés étaient de type “chat” ce qui ne correspondait pas à nos attentes.