



**HAL**  
open science

## Déployer des LLMs en local

Yannis Bendi-Ouis, Xavier Hinaut

► **To cite this version:**

Yannis Bendi-Ouis, Xavier Hinaut. Déployer des LLMs en local. Dataquitaine, Mar 2024, Bordeaux, France. hal-04850952

**HAL Id: hal-04850952**

**<https://hal.science/hal-04850952v1>**

Submitted on 20 Dec 2024

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

*Ínria*

Déployer des  
LLMs en Local



*Inria*

# Yannis Bendi-Ouis

Doctorant dans l'équipe Mnemosyne

Encadré par Xavier Hinaut




Email :

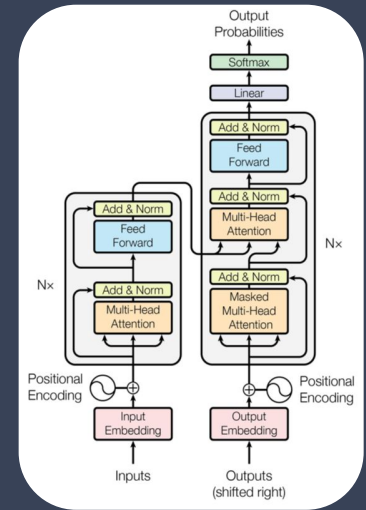
[yannis.bendi-ouis@inria.fr](mailto:yannis.bendi-ouis@inria.fr)

Site web :

[www.naowak.fr](http://www.naowak.fr)

## Qu'est ce qu'un LLM ?

- > ChatGPT 
- > LLM : Large Language Model (Gros modèle de langage)
- > Génère du texte
- > Basé sur une architecture : Transformers
- > Beaucoup de paramètres/calculs



# Pourquoi faire du local ?

- > Pas besoin de connexion internet
- > Force l'utilisation de modèles plus petits et moins énergivores
- > Protection des données et de la vie privée
- > Décentralisation du pouvoir (toute les données ne vont pas au même endroit)
- > Réduction des risques de biais volontaires (propagande, manipulation de masse)
- > Modèles ouverts (étudiables, on peut observer ses tendances/biais, maîtrise ses MAJ)
- > Possibilités de finetune (contrôle du process de A à Z).

01

# De nombreux modèles

Du choix pour tous

# Hugging Face



# Hugging Face

**Tasks** Libraries Datasets Languages Licenses  
Other

Filter Tasks by name

Multimodal

Image-Text-to-Text

Visual Question Answering

Document Question Answering

Computer Vision

Depth Estimation Image Classification

Object Detection Image Segmentation

Text-to-Image Image-to-Text

Image-to-Image Image-to-Video

Unconditional Image Generation

Video Classification Text-to-Video

Zero-Shot Image Classification

**Models** 559,111 Filter by name

new Full-text search

Sort: Trending

xai-org/grok-1

Text Generation · Updated 1 day ago · 1.28k

CohereForAI/c4ai-command-r-v01

Text Generation · Updated 1 day ago · 13k · 685

xai-org/grok-1

NousResearch/Hermes-2-Pro-Mistral-7B

Text Generation · Updated 5 days ago · 3.88k · 242

google/gemma-7b

Text Generation · Updated 21 days ago · 269k · 2.46k

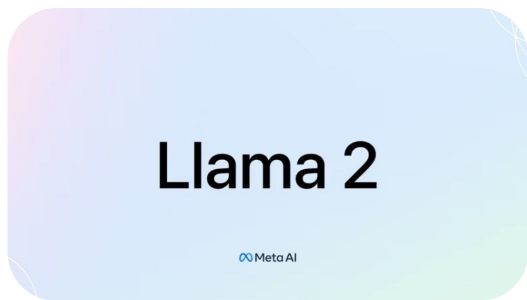
stabilityai/sv3d

Image-to-Video · Updated 1 day ago · 180

cagliostrolab/animate-xl-3.1

Text-to-Image · Updated 2 days ago · 8.71k · 155

# Llama 2 & Mistral



7B, 13B, 70B



Mistral 7B

Mixtral 8x7B

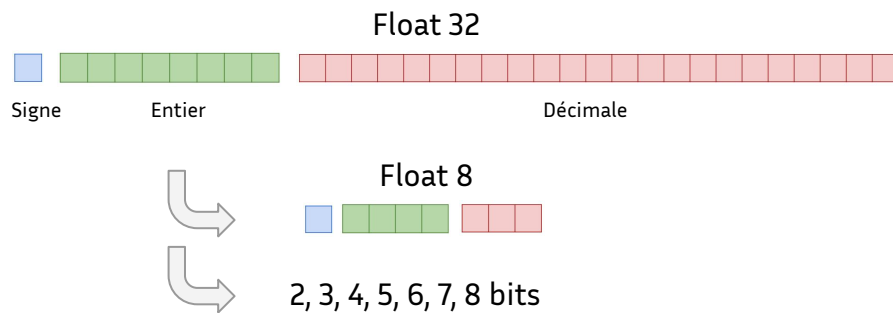




# Llama 2 & Mistral

Rank ▲	🌐 Model ▲	☆ Arena Elo ▲	Organization ▲	License ▲
1	<a href="#">GPT-4-1106-preview</a>	1251	OpenAI	Proprietary
1	<a href="#">GPT-4-0125-preview</a>	1249	OpenAI	Proprietary
1	<a href="#">Claude 3 Opus</a>	1247	Anthropic	Proprietary
4	<a href="#">Bard (Gemini Pro)</a>	1202	Google	Proprietary
4	<a href="#">Claude 3 Sonnet</a>	1190	Anthropic	Proprietary
5	<a href="#">GPT-4-0314</a>	1185	OpenAI	Proprietary
7	<a href="#">GPT-4-0613</a>	1159	OpenAI	Proprietary
7	<a href="#">Mistral-Large-2402</a>	1155	Mistral	Proprietary
8	<a href="#">Qwen1.5-72B-Chat</a>	1146	Alibaba	Qianwen LICENSE
8	<a href="#">Claude-1</a>	1145	Anthropic	Proprietary
8	<a href="#">Mistral Medium</a>	1145	Mistral	Proprietary
12	<a href="#">Claude-2.0</a>	1126	Anthropic	Proprietary
12	<a href="#">Mistral-Next</a>	1123	Mistral	Proprietary
12	<a href="#">Gemini Pro (Dev API)</a>	1118	Google	Proprietary
13	<a href="#">Claude-2.1</a>	1115	Anthropic	Proprietary
13	<a href="#">Mixtral-8x7b-Instruct-v0.1</a>	1114	Mistral	Apache 2.0

# Quantification



> Taille d'un modèle quantifié

7B ~ 30 Go → 4 bits → ~ 4 Go

> RAM / VRAM nécessaire

7B ~ 30 Go → 4 bits → ~ 7 Go

02

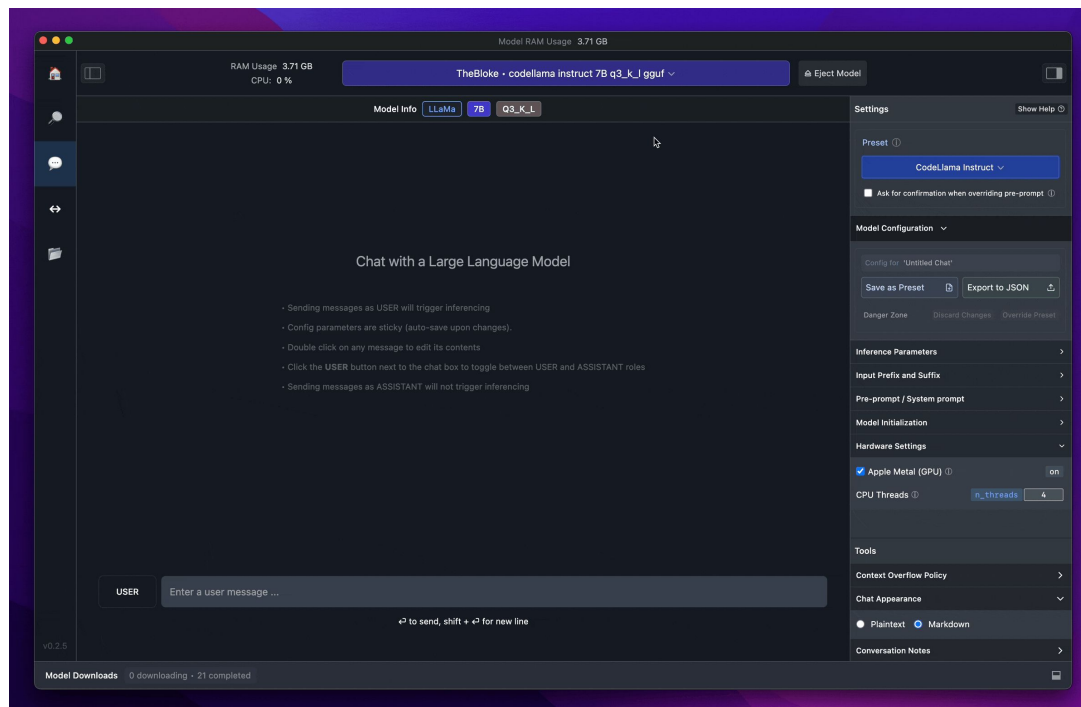
# LM Studio

Une interface graphique pour du local

# LM Studio



- > Logiciel propriétaire
  - > Gratuit
  - > Facile
- 
- > Aucune connaissance en programmation requise
  - > Installation d'un modèle quantifié à partir de HuggingFace
  - > Configuration Automatique / Manuelle



# 03

## Ollama

Intégration facile dans le terminal  
et plus encore

# Ollama



- > Open source
- > Gratuit
- > Facile

- 
- > Installation et utilisation facile dans un terminal
  - > Lancement facile d'un serveur
  - > Plusieurs interfaces graphiques disponibles
  - > Connexion facile à d'autres applications

A screenshot of a terminal window with a dark background. The window title is 'tobiasmann@Tobias-MBP--'. The terminal shows the following commands and their outputs:

```
ollama rm falcon:40b
ollama run falcon:40b
ollama run mistral:7b-instruct-q8_0
ollama run mistral:7b-instruct-q8_0
clear
ollama run mistral:7b-instruct-q8_0
```

The command 'ollama run mistral:7b-instruct-q8\_0' is highlighted with a pink background.

# 04

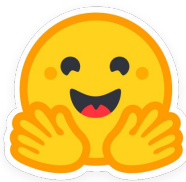
## Transformer & Llama.cpp

Du développement en python

# Transformers

- > Open source
- > Gratuit
- > Facile mais développement python

- 
- > Implémentation Pytorch & Tensorflow
  - > Mis à jour très régulièrement
  - > Accepte énormément d'architecture de LLM



## Hugging Face

```
>>> from transformers import AutoModelForCausalLM, AutoTokenizer
>>> device = "cuda" # the device to load the model onto

>>> model = AutoModelForCausalLM.from_pretrained("mistralai/Mixtral-8x7B-v0.1")
>>> tokenizer = AutoTokenizer.from_pretrained("mistralai/Mixtral-8x7B-v0.1")

>>> prompt = "My favourite condiment is"

>>> model_inputs = tokenizer([prompt], return_tensors="pt").to(device)
>>> model.to(device)

>>> generated_ids = model.generate(**model_inputs, max_new_tokens=100, do_sample=True)
>>> tokenizer.batch_decode(generated_ids)[0]
"The expected output"
```



# Llama.cpp

- > Open source
  - > Gratuit
  - > Facile mais développement C++/python
- 
- > Permet l'utilisation de modèle quantifié (GGUF)
  - > Mis à jour très régulièrement
  - > Accepte énormément d'architecture de LLM



**r/LocalLLaMA**

Communauté sur Reddit

# 05

## vLLM

Déploiement à grande échelle

# vLLM

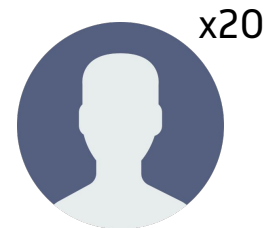
- > Librairie open-source
- > Fonctionne sur les GPU NVIDIA
- > Optimise les calculs et permet de traiter plusieurs utilisateurs en simultan 



A100



Mixtral



~ 20 tokens/seconde

# Merci !

Suivez-nous sur [www.inria.fr](http://www.inria.fr)