



**HAL**  
open science

## LLMs Local: Pourquoi et Comment ?

Yannis Bendi-Ouis, Xavier Hinaut

► **To cite this version:**

Yannis Bendi-Ouis, Xavier Hinaut. LLMs Local: Pourquoi et Comment ?. Midis de la Bidouille à l'Inria, Nov 2023, Bordeaux, France. hal-04851146

**HAL Id: hal-04851146**

**<https://hal.science/hal-04851146v1>**

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

# Inria

LLMs local

Pourquoi et  
comment ?



LLAMA 2

CHATGPT

MISTRAL AI



**LLAMA 2**



**CHATGPT**



**MISTRAL AI**

# Sommaire

01. Qu'est-ce qu'un Transformer
02. Des modèles open-sources
03. Le prompt : la création d'agents
04. LLaMA.cpp et la communauté open-source

Puis un peu de pratique !

*Inria*

# Pourquoi faire du local ?



# Pourquoi faire du local ?

- > Pas besoin de connection internet

# Pourquoi faire du local ?

- > Pas besoin de connection internet
- > Plus petite consommation d'énergie

# Pourquoi faire du local ?

- > Pas besoin de connection internet
- > Plus petite consommation d'énergie
- > Protection des données et de la vie privée



# Pourquoi faire du local ?

- > Pas besoin de connection internet
- > Plus petite consommation d'énergie
- > Protection des données et de la vie privée
- > Décentralise le pouvoir (toute les données ne vont pas au même endroit)

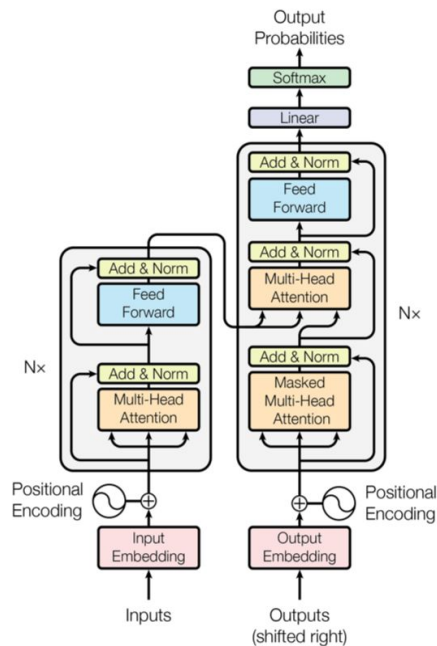
# Pourquoi faire du local ?

- > Pas besoin de connection internet
- > Plus petite consommation d'énergie
- > Protection des données et de la vie privée
- > Décentralise le pouvoir (toute les données ne vont pas au même endroit)
- > Réduction des risques de biais volontaires (propagande, manipulation de masse)

01

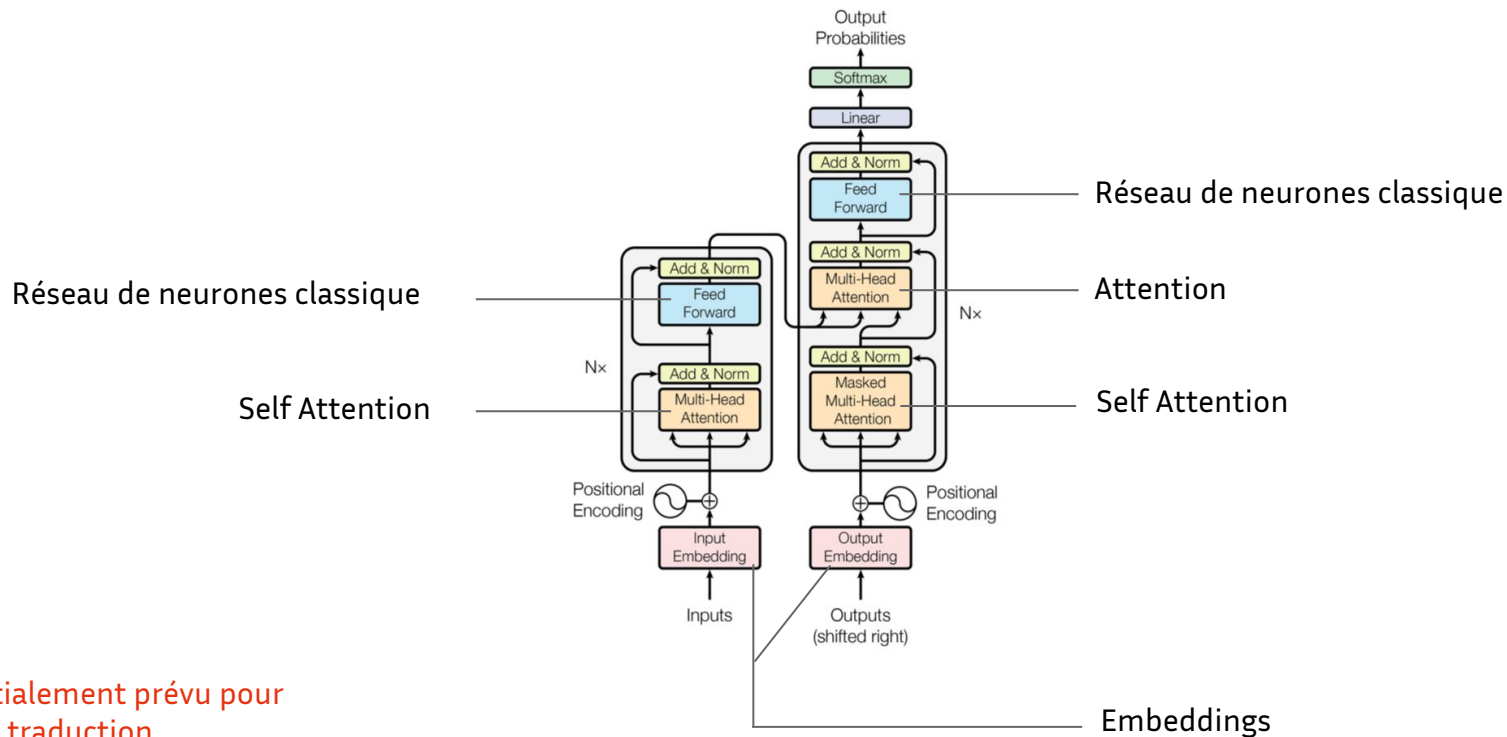
Qu'est ce qu'un  
Transformers ?

# L'architecture d'un Transformers



\* Initialement prévu pour de la traduction

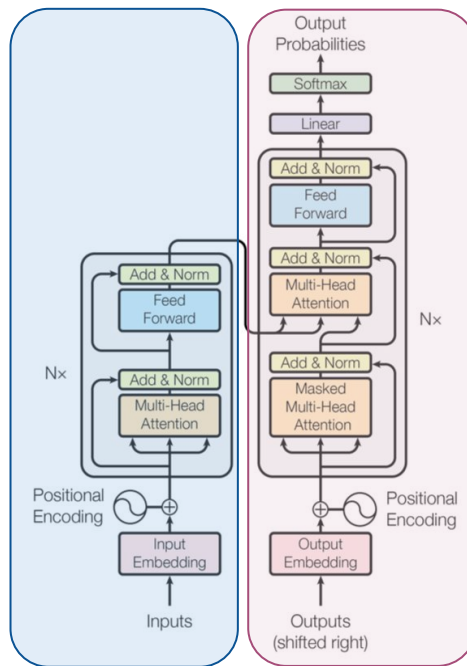
# L'architecture d'un Transformers



\* Initialement prévu pour de la traduction

# Encodeur et Décodeur

Encodeur



Décodeur

\* Initialement prévu pour de la traduction

# Encodeur

- > Son but est de compresser les données qu'il reçoit, de manière à ce qu'elle puisse être ensuite décodé par le décodeur.

# Encodeur

- > Son but est de compresser les données qu'il reçoit, de manière à ce qu'elle puisse être ensuite décodée par le décodeur.
- > Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo



0.12	0.42	0.89	1.56	2.11	0.08	1.95	1.12	0.88	2.02
------	------	------	------	------	------	------	------	------	------



# Encodeur

- > Son but est de compresser les données qu'il reçoit, de manière à ce qu'elle puisse être ensuite décodée par le décodeur.
- > Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo



0.12	0.42	0.89	1.56	2.11	0.08	1.95	1.12	0.88	2.02
------	------	------	------	------	------	------	------	------	------

# Décodeur

- > Son but est de décoder des données issues de l'encodeur.

# Encodeur

- > Son but est de compresser les données qu'il reçoit, de manière à ce qu'elle puisse être ensuite décodée par le décodeur.
- > Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo



0.12	0.42	0.89	1.56	2.11	0.08	1.95	1.12	0.88	2.02
------	------	------	------	------	------	------	------	------	------

# Décodeur

- > Son but est de décoder des données issues de l'encodeur.
- > On a remarqué que si on ne lui donne pas de données issues de l'encodeur, mais qu'on lui donne le début d'une séquence, il est capable de la continuer.

# Encodeur

- > Son but est de compresser les données qu'il reçoit, de manière à ce qu'elle puisse être ensuite décodée par le décodeur.
- > Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo



0.12	0.42	0.89	1.56	2.11	0.08	1.95	1.12	0.88	2.02
------	------	------	------	------	------	------	------	------	------

# Décodeur

- > Son but est de décoder des données issues de l'encodeur.
- > On a remarqué que si on ne lui donne pas de données issues de l'encodeur, mais qu'on lui donne le début d'une séquence, il est capable de la continuer.
- > Dans notre cas (texte), il est capable de deviner le prochain mot d'une phrase.

# Encodeur

- > Son but est de compresser les données qu'il reçoit, de manière à ce qu'elle puisse être ensuite décodée par le décodeur.
- > Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).

Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo



0.12	0.42	0.89	1.56	2.11	0.08	1.95	1.12	0.88	2.02
------	------	------	------	------	------	------	------	------	------

# Décodeur

- > Son but est de décoder des données issues de l'encodeur.
- > On a remarqué que si on ne lui donne pas de données issues de l'encodeur, mais qu'on lui donne le début d'une séquence, il est capable de la continuer.
- > Dans notre cas (texte), il est capable de deviner le prochain mot d'une phrase.
- > C'est chatGPT.

02

# Des modèles, toujours plus de modèles

# Toujours plus gros, toujours plus fort !

Modèle	Architecture	Nombre paramètres
GPT (2017)	Décodeur	120.000.000 (120M)
BERT (2018)	Encodeur	340.000.000 (340M)
GPT-2 XL (2019)	Décodeur	1.500.000.000 (1.5B)
GPT-3 (2020)	Décodeur	175.000.000.000 (175B)
GPT-4 (2023)	Décodeur + MOE	1.760.000.000.000 (8 * 220B)

# Autant avec moins ?

Modèle	Architecture	Nombre paramètres
GPT (2017)	Décodeur	120.000.000 (120M)
BERT (2018)	Encodeur	340.000.000 (340M)
GPT-2 XL (2019)	Décodeur	1.500.000.000 (1.5B)
GPT-3 (2020)	Décodeur	175.000.000.000 (175B)
GPT-4 (2023)	Décodeur + MOE	1.760.000.000.000 (8 * 220B)
LLaMA-2 (2023)	Décodeur	7B - 14B - 70B
Mistral (2023)	Décodeur	7.000.000.000 (7B)
GPT-3.5-turbo (2023)	Décodeur	20.000.000.000 (20B)

# Autant avec moins ?

Modèle	Architecture	Nombre paramètres
GPT (2017)	Décodeur	120.000.000 (120M)
BERT (2018)	Encodeur	340.000.000 (340M)
GPT-2 XL (2019)	Décodeur	1.500.000.000 (1.5B)
GPT-3 (2020)	Décodeur	175.000.000.000 (175B)
GPT-4 (2023)	Décodeur + MOE	1.760.000.000.000 (8 * 220B)
LLaMA-2 (2023)	Décodeur	7B - 14B - 70B
Mistral (2023)	Décodeur	7.000.000.000 (7B)
GPT-3.5-turbo (2023)	Décodeur	20.000.000.000 (20B)

> Et il en existe encore plein ! Notamment si on compte les version finetuned.



# Des modèles de base open-source



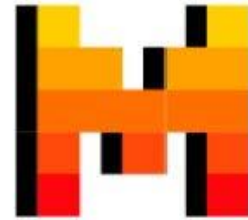
Falcon  
(7B, 40B, 180B)

Emirats Arabes Unis  
Juin 2023



Llama 2  
(7B, 13B, 70B)

Etats-Unis  
Juillet 2023



Mistral  
(7B)

France  
Septembre 2023

# 03

## Le prompt : la création d'agents

# Format de prompt

**system**

**user**

**assistant**

**user**

**assistant**

...

# Format de prompt

**system**

---

**user**

**assistant**

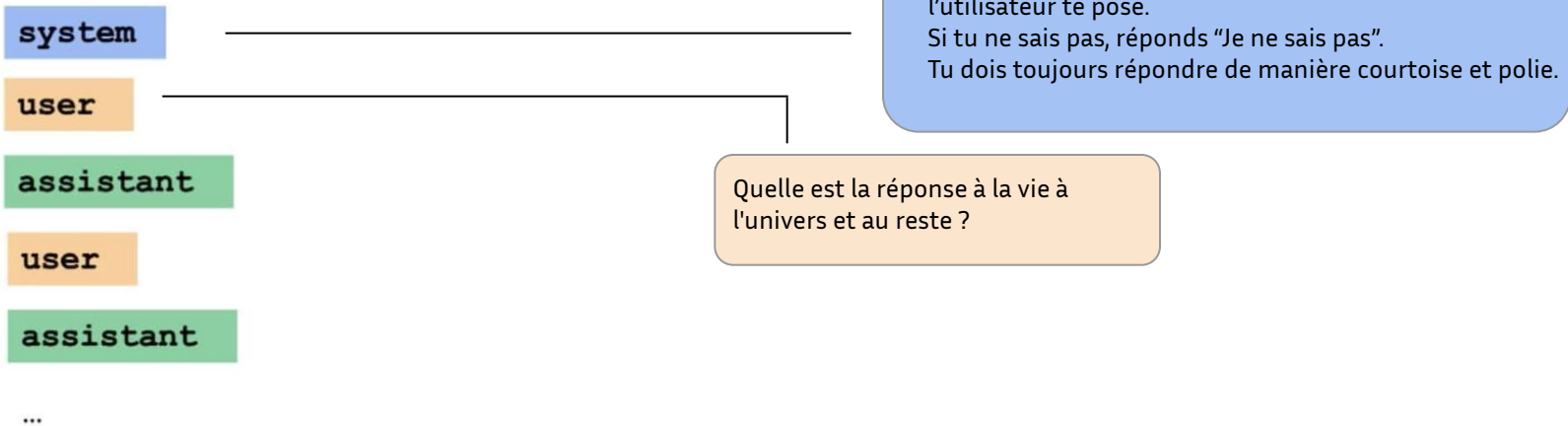
**user**

**assistant**

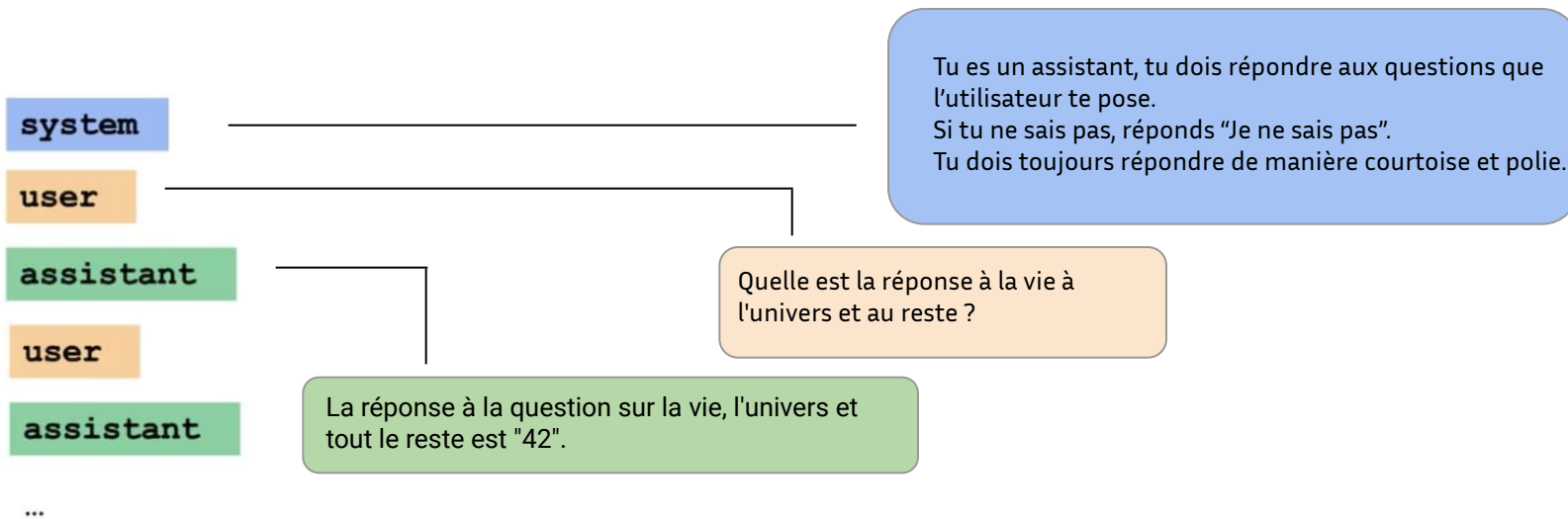
...

Tu es un assistant, tu dois répondre aux questions que l'utilisateur te pose.  
Si tu ne sais pas, réponds "Je ne sais pas".  
Tu dois toujours répondre de manière courtoise et polie.

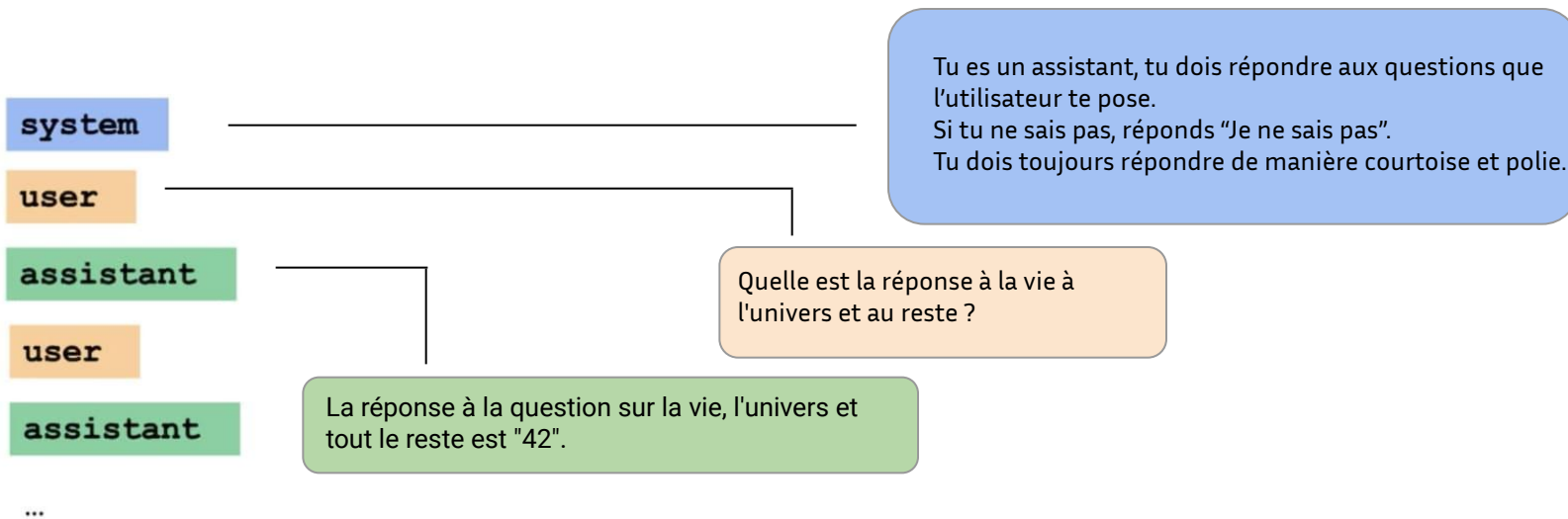
# Format de prompt



# Format de prompt

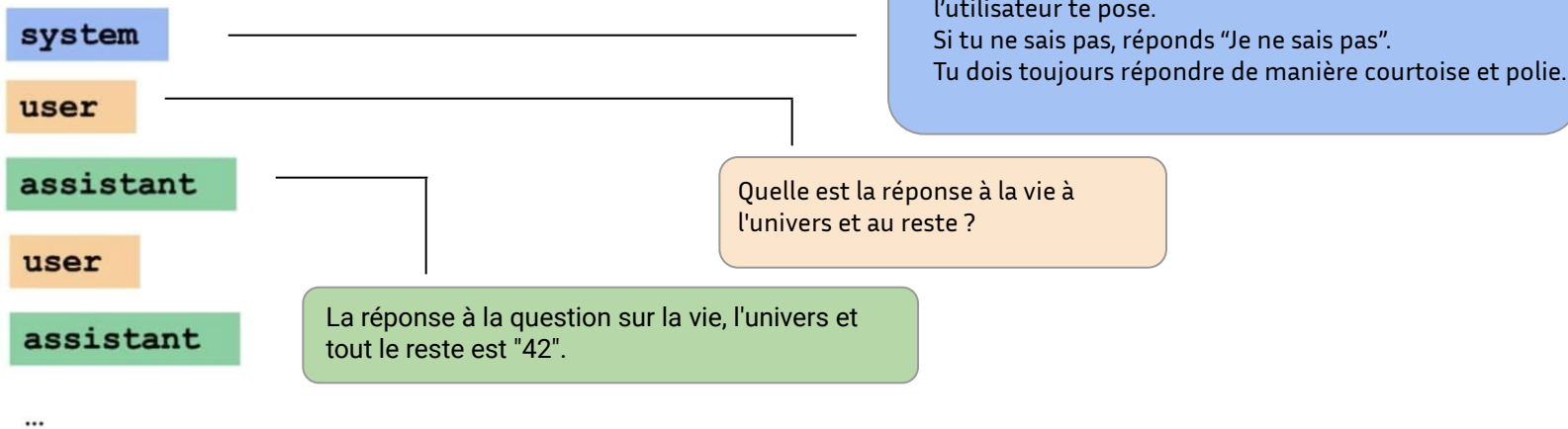


# Format de prompt



```
<|im_start|>system{system message}<|im_end|>  
<|im_start|>user{user_message}<|im_end|>  
<|im_start|>assistant
```

# Format de prompt



```
<|im_start|>system{system message}<|im_end|>  
<|im_start|>user{user_message}<|im_end|>  
<|im_start|>assistant
```



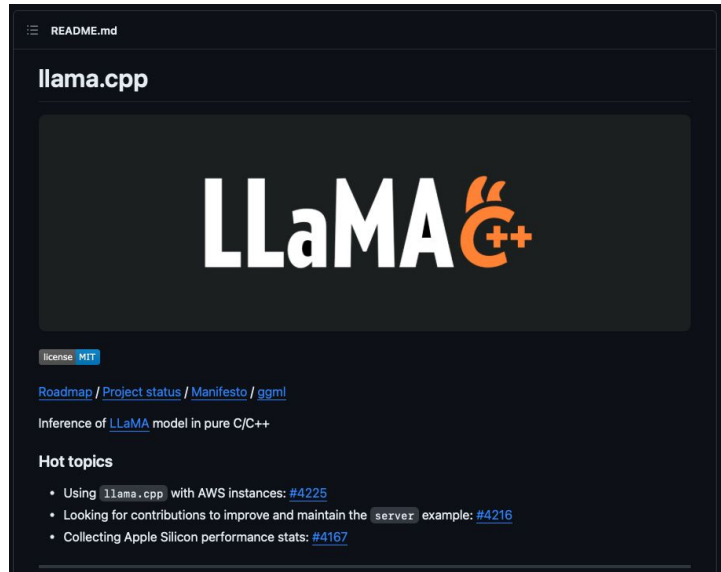
04

# LLaMA.cpp

Une communauté open-source

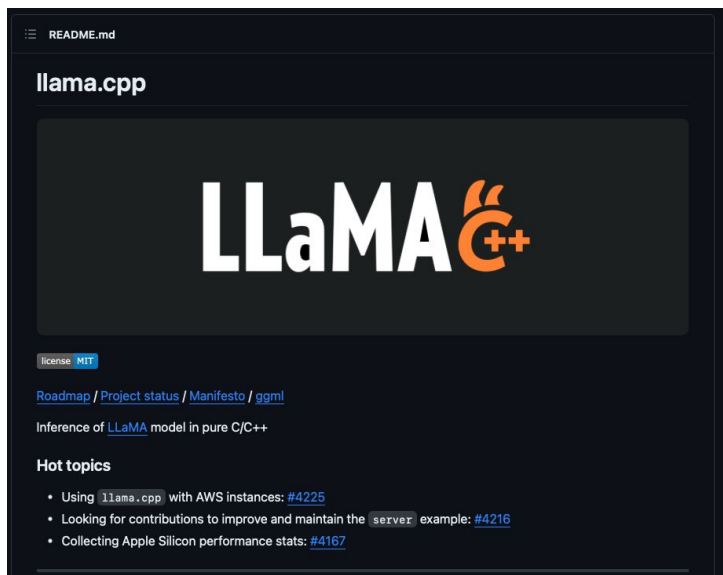
# Le projet d'une soirée : LLaMA.cpp

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

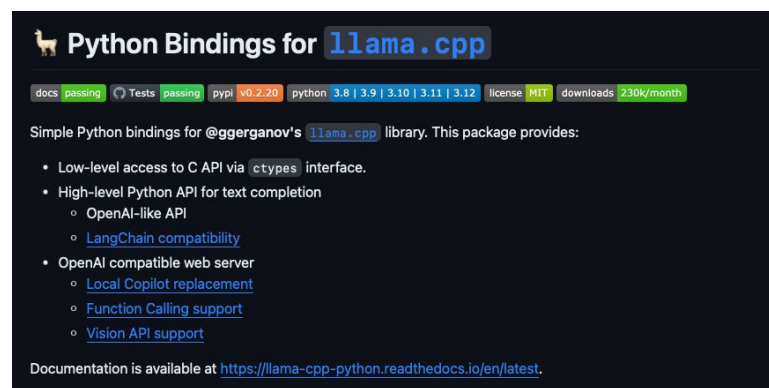


# Le projet d'une soirée : LLaMA.cpp

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

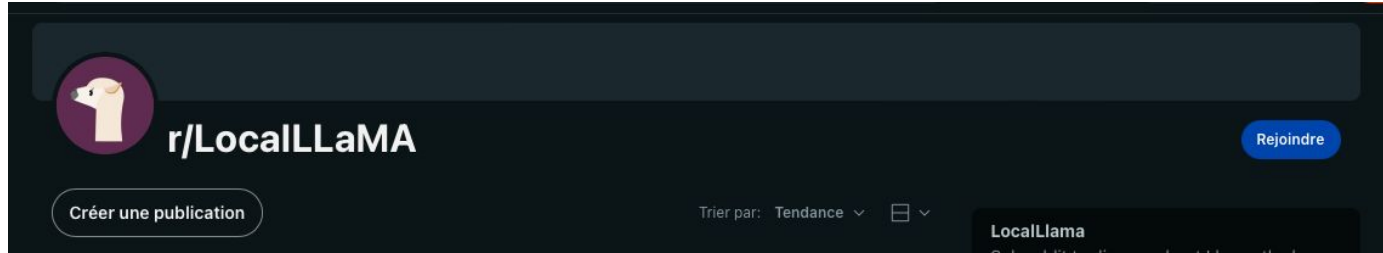


> <https://github.com/abetlen/llama-cpp-python>

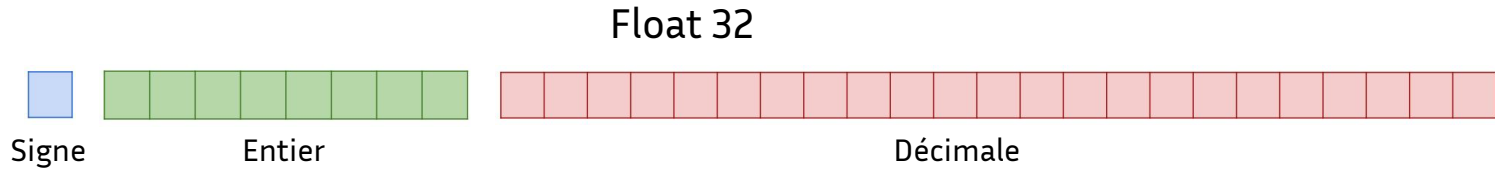


# Un thread Reddit : r/LocalLLaMA

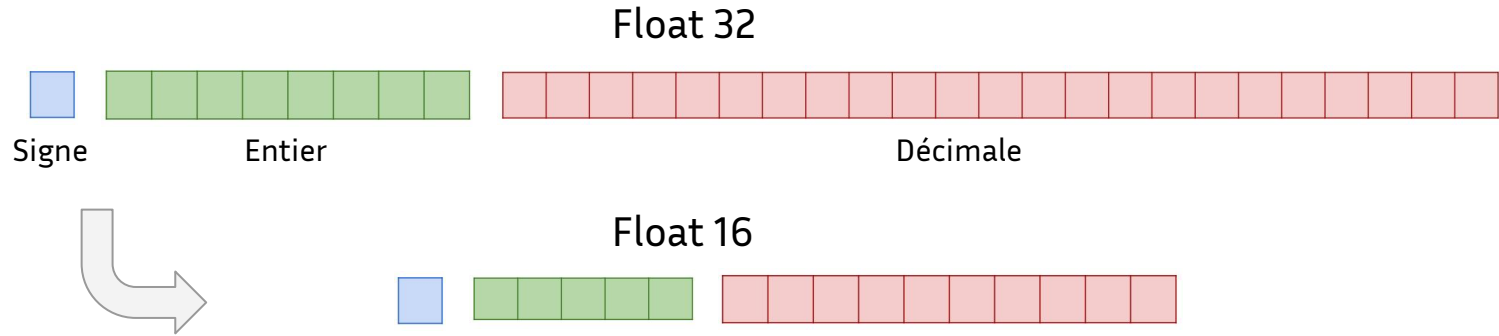
> <https://www.reddit.com/r/LocalLLaMA/>



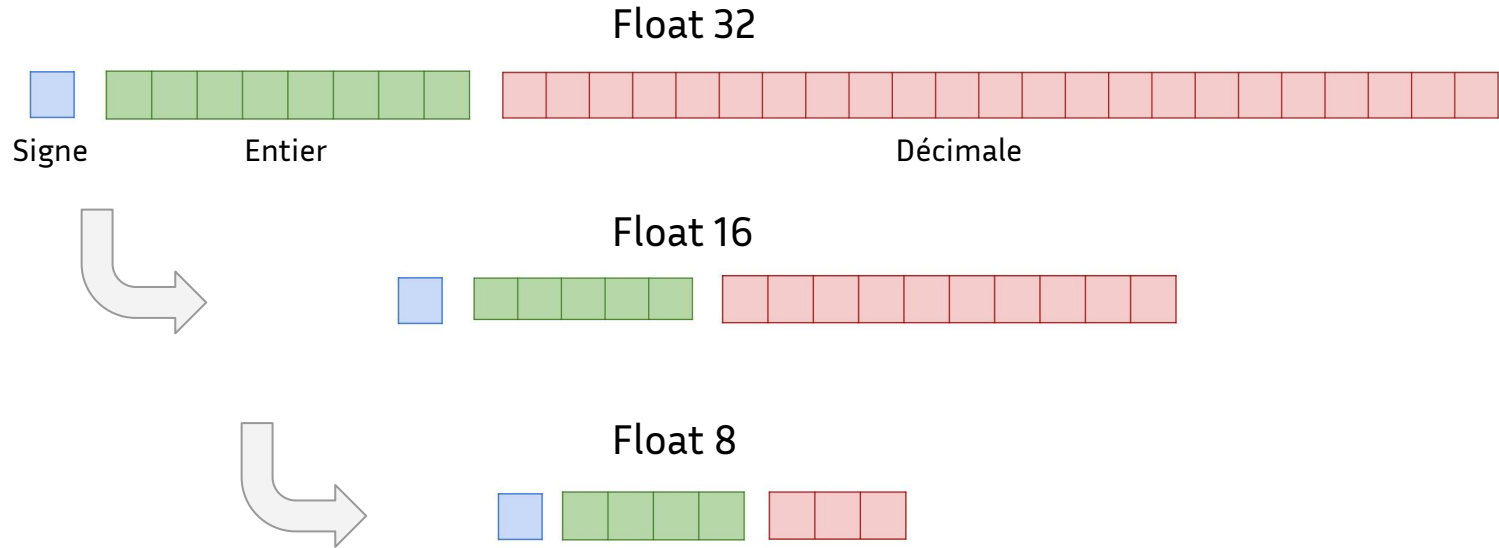
# Quantification



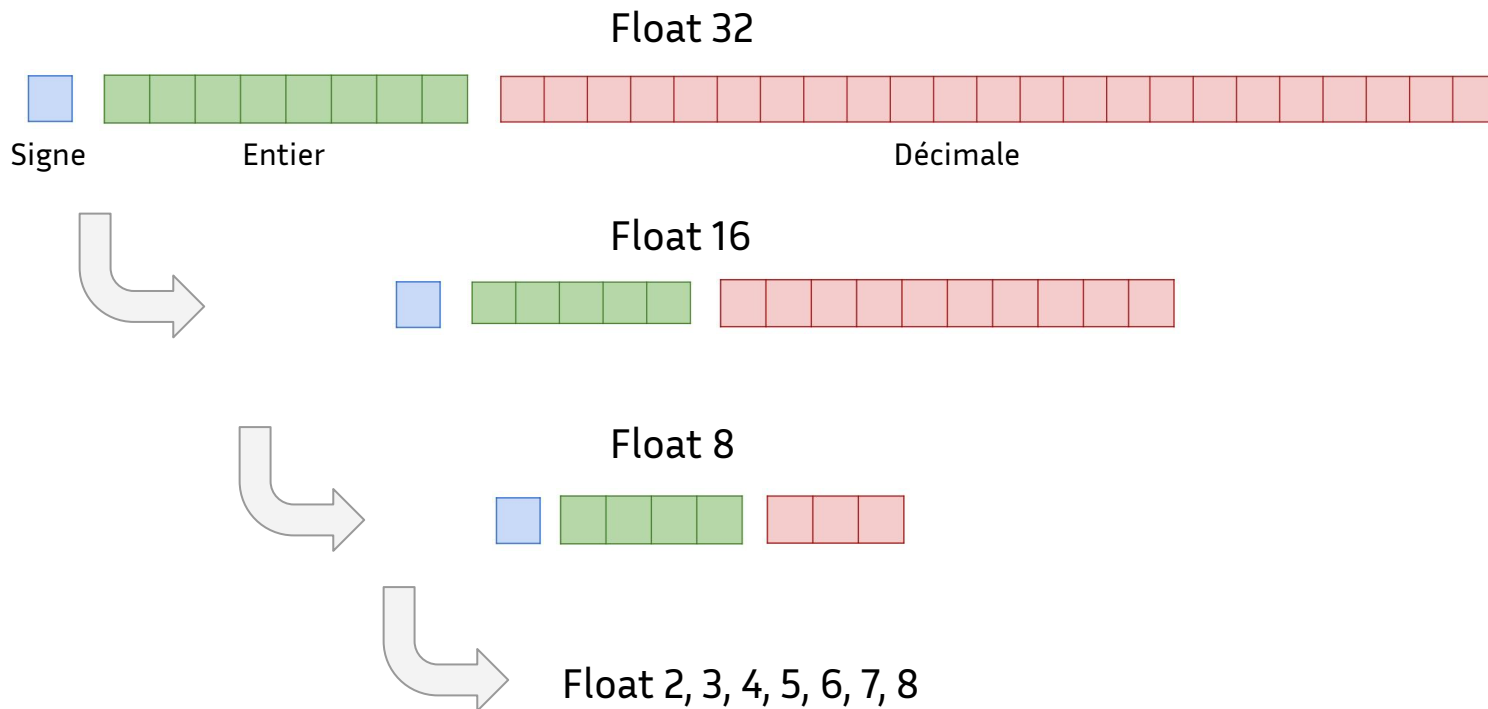
# Quantification



# Quantification



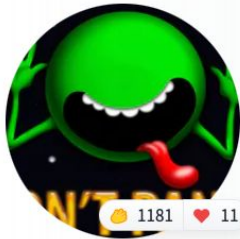
# Quantification





# Le github des modèles d'IA : HuggingFace

> <https://huggingface.co/TheBloke>



1181 11

**Tom Jobbins** PRO  
TheBloke

Follow

10272 followers · 8 following

TheBlokeAI TheBloke

## Collections 1

### Recent models: last 100 repos, sorted by creation date >

The last 100 repos I have created. Sorted by creation date descend...

TheBloke/OpenOrca-Zephyr-7B-AWQ

Text Generation · Updated about 9 hours ago

TheBloke/OpenOrca-Zephyr-7B-GGUF

Updated about 10 hours ago · 4

TheBloke/OpenOrca-Zephyr-7B-GPTQ

Text Generation · Updated about 9 hours ago · 1

TheBloke/Poro-34B-GPTQ

Models 2893

# C'est parti pour la pratique !

<https://github.com/Naowak/Local-LLMs>