LLMs Local : Pourquoi et Comment ?

Yannis Bendi-Ouis, Xavier Hinaut

LLMs local

-

Pourquoi et comment ?

Yannis Bendi-Ouis


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 !

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)

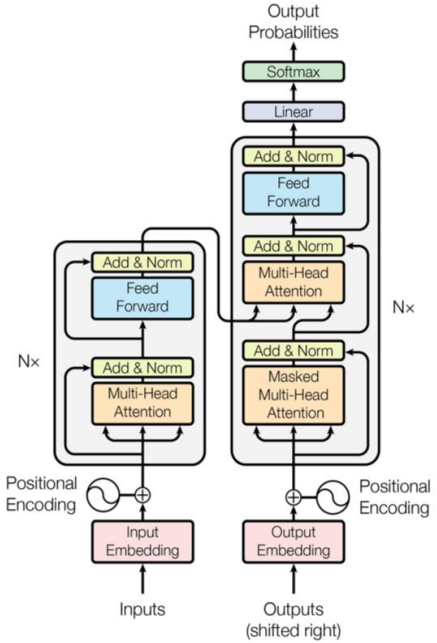
Pourquoi faire du local ?

- 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

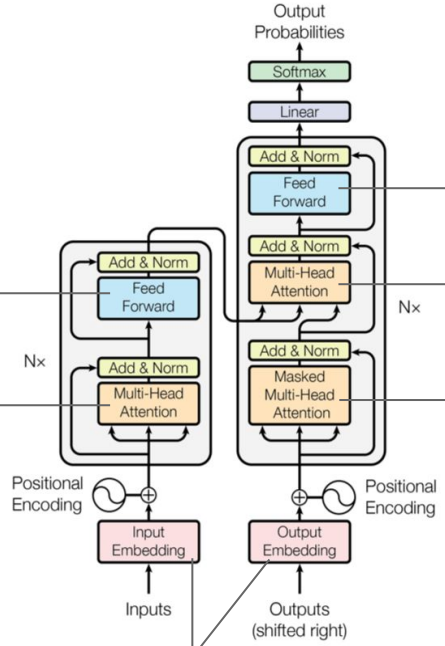
Qu’est ce qu’un Transformers ?

L’architecture d’un Transformers



* Initialement prévu pour de la traduction

L'architecture d'un Transformers



Réseau de neurones classique

Réseau de neurones classique

Attention

Self Attention

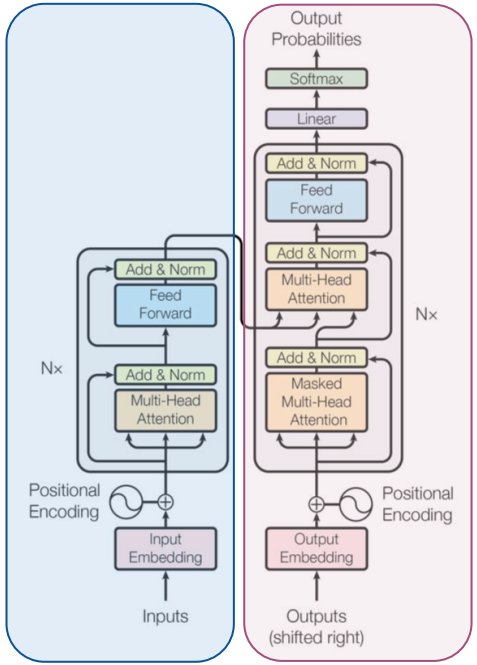
Self Attention

Embeddings

* 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 décoder par le décodeur.

Encodeur

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

> Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).

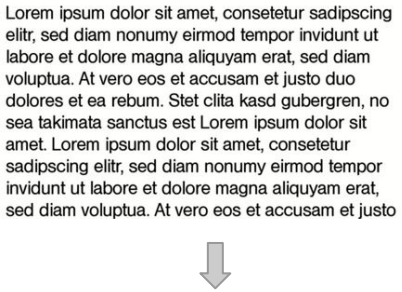
Lorem ipsum dolor sit amet, consetetur sadipscing elitr, sed diam nonumy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem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

> Il peut être capable de transformer un texte en un simple vecteur (modèle BERT).



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 décoder 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 décoder 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 décoder 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

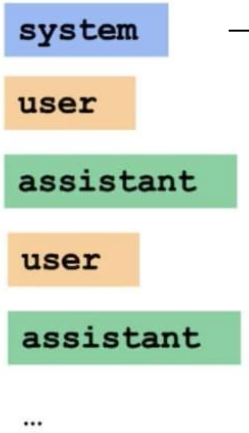
03

Le prompt : la création d'agents

Format de prompt

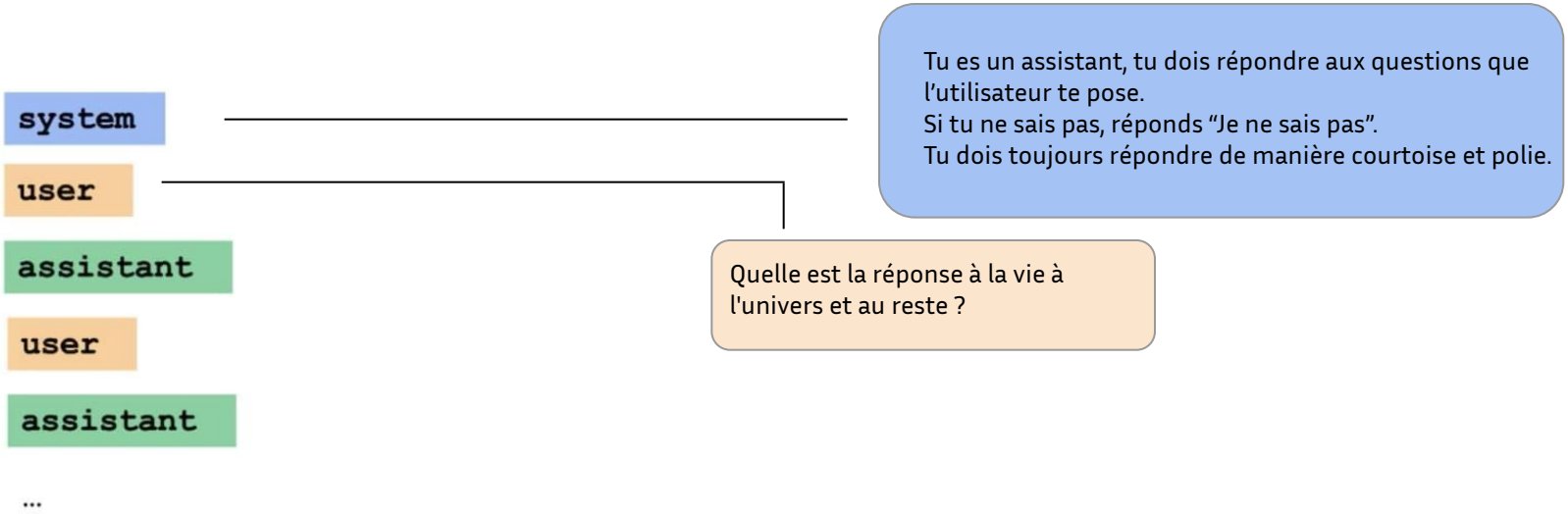
```
system
user
assistant
user
assistant
...
```

Format de prompt

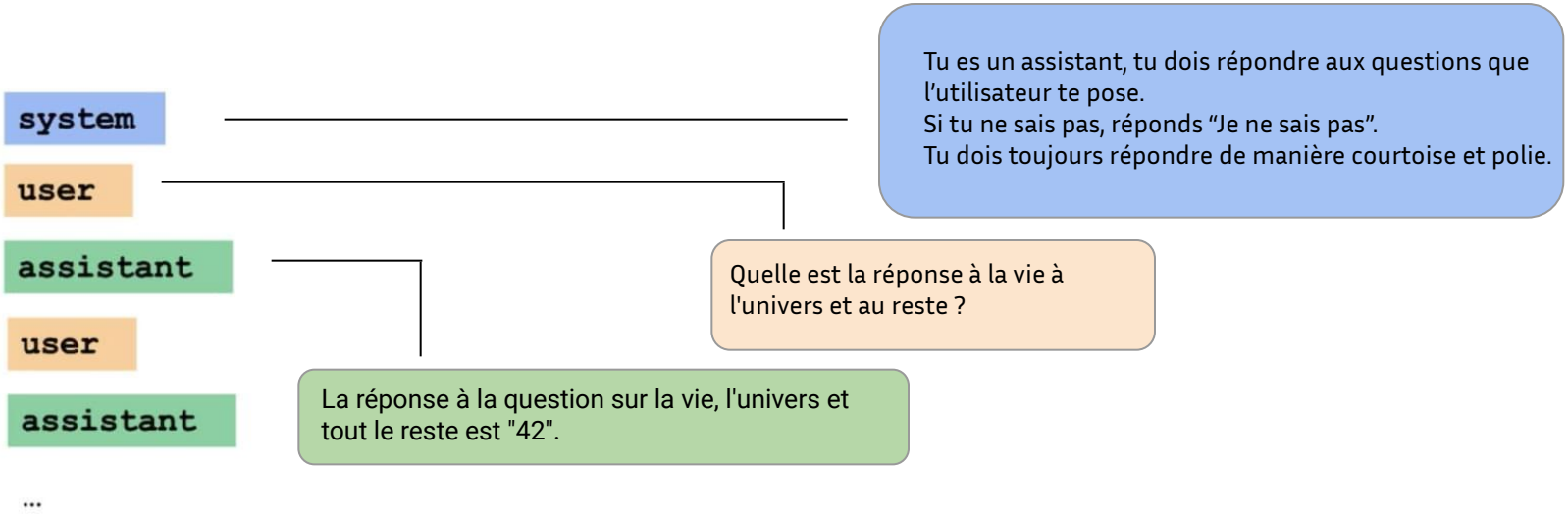


Tu es un assistant, tu dois répondre aux questions que l'utilisateur te pose.
Si tu ne sais pas, réponds "Je ne sais pas".
Tu dois toujours répondre de manière courtoise et polie.

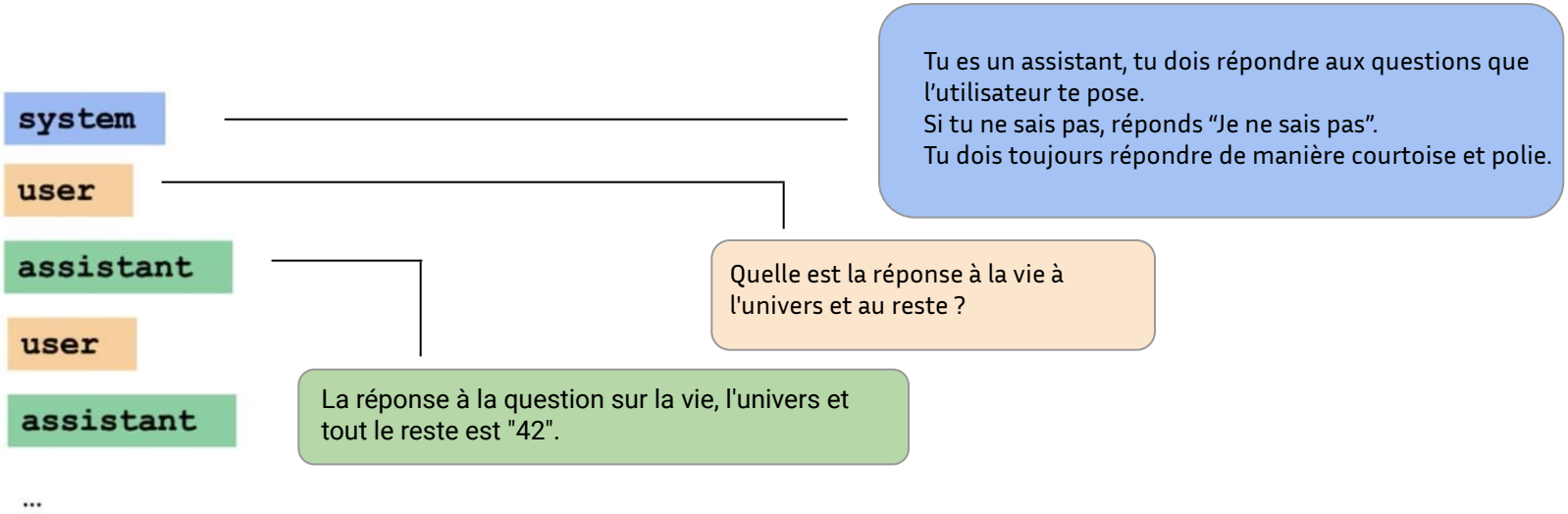
Format de prompt



Format de prompt

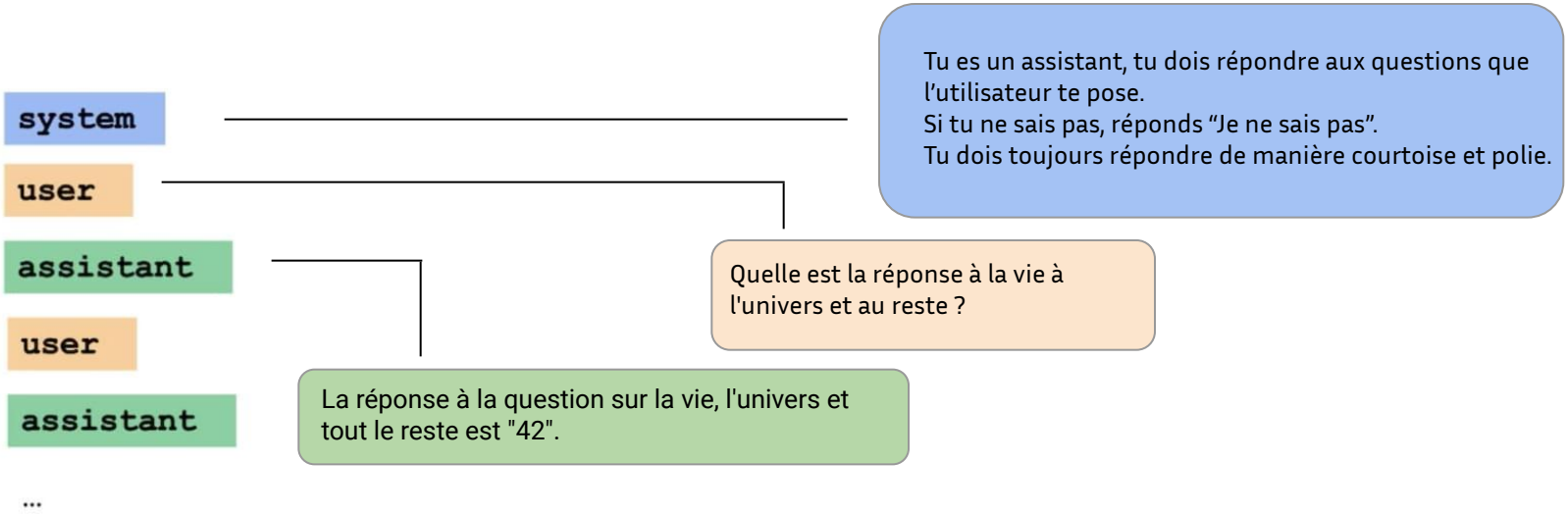


Format de prompt




```
<|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
```

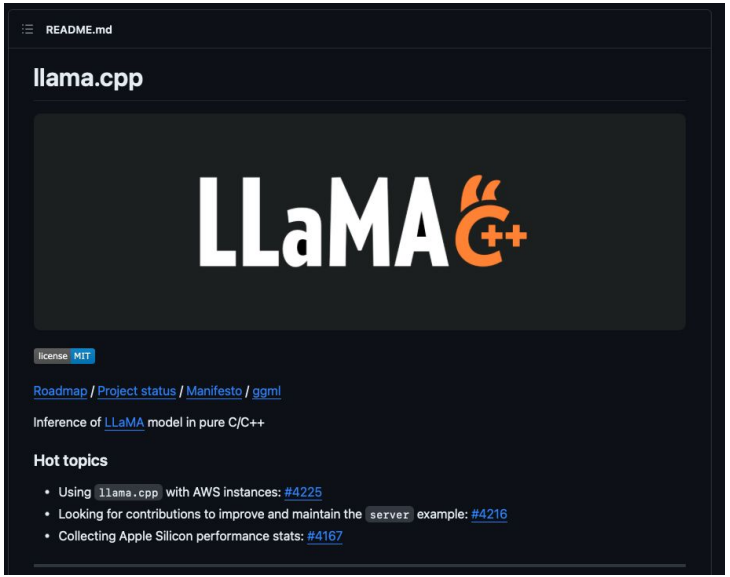
04

LLaMA.cpp

Une communauté open-source

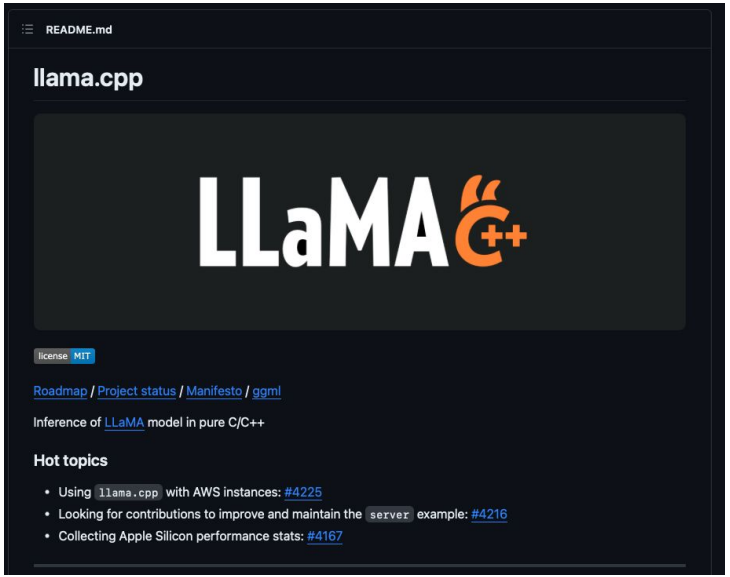
Le projet d'une soirée : LLaMA.cpp

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

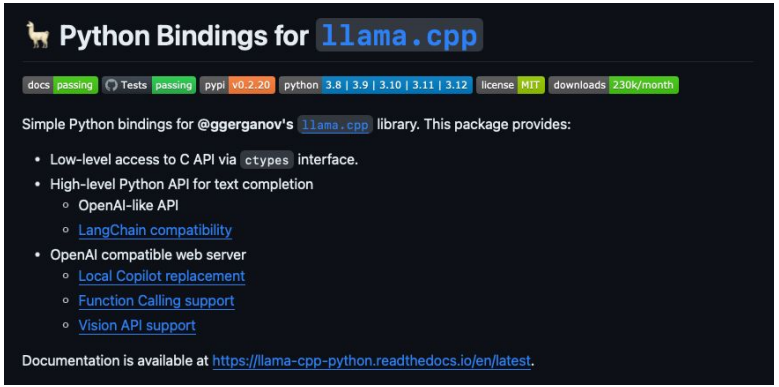


Le projet d'une soirée : LLaMA.cpp

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

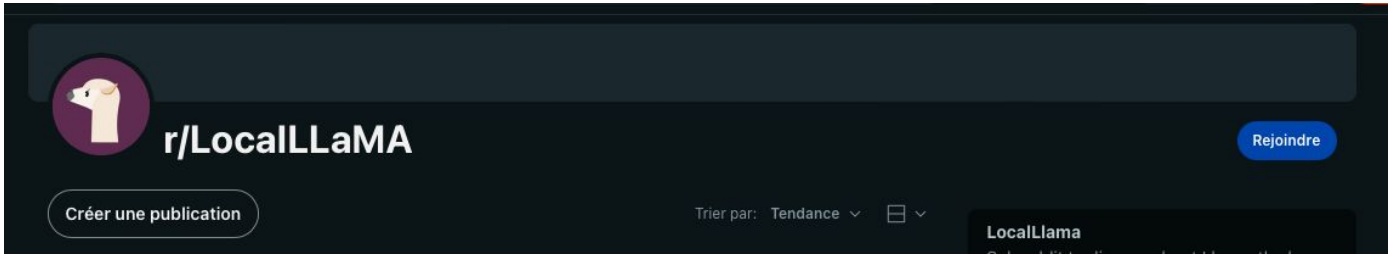


> https://github.com/abetlen/llama-cpp-python

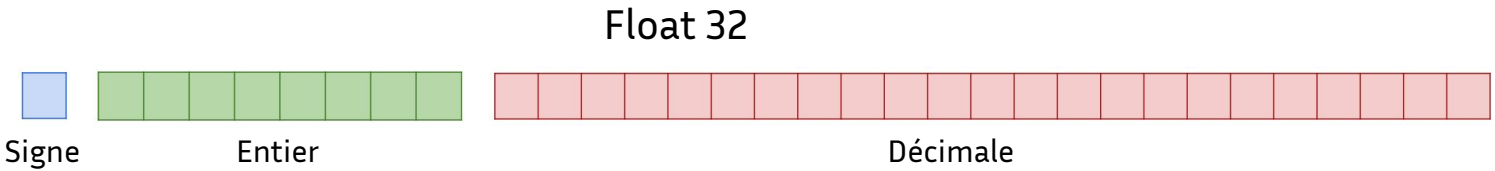


Un thread Reddit : r/LocalLLaMA

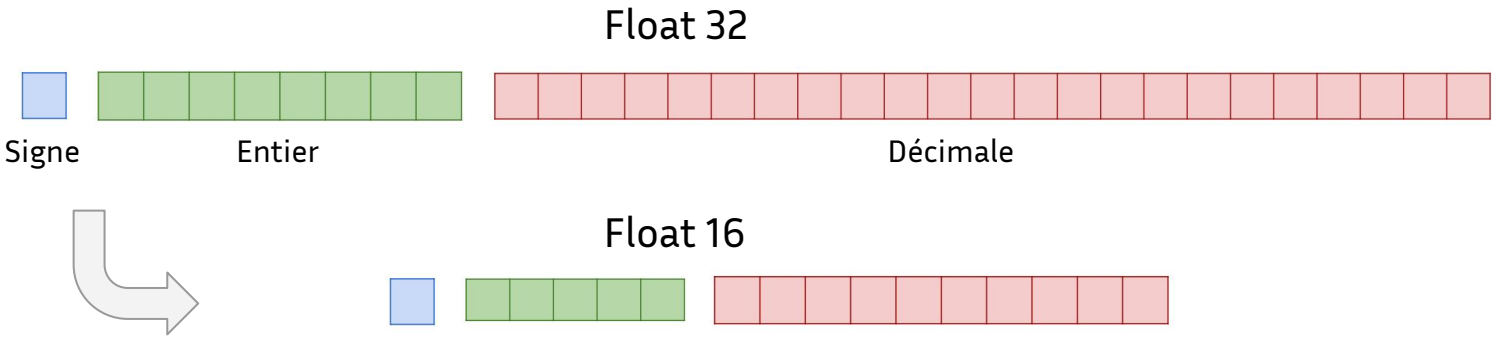
> https://www.reddit.com/r/LocalLLaMA/



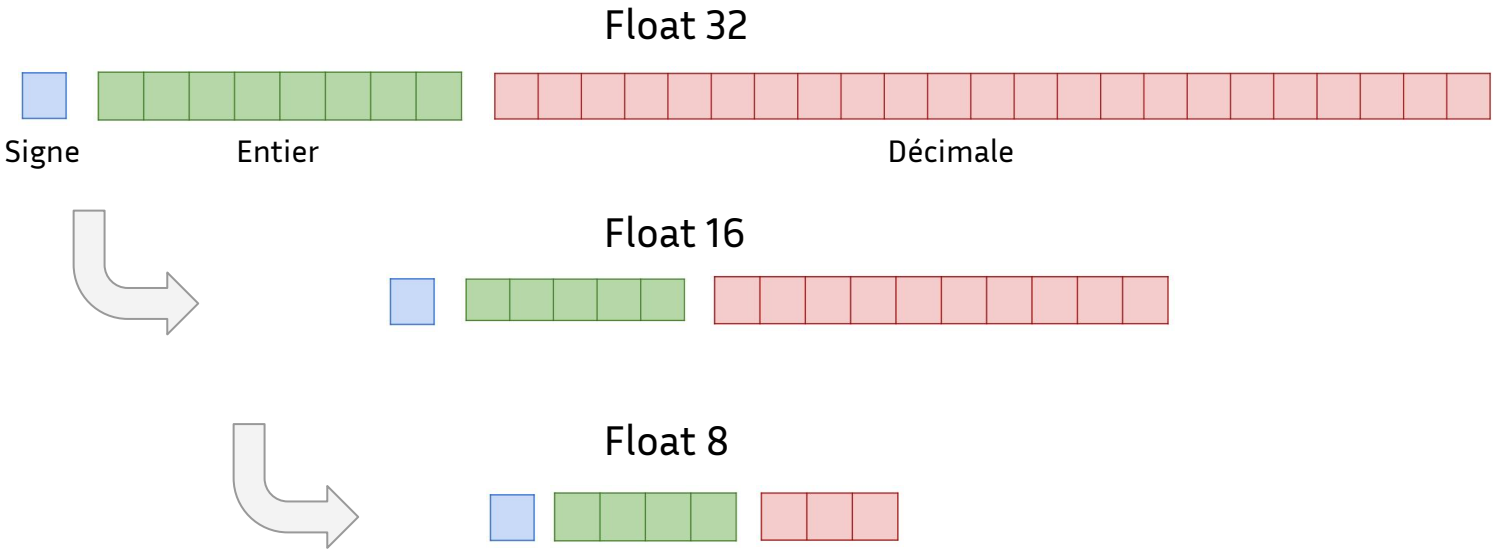
Quantification



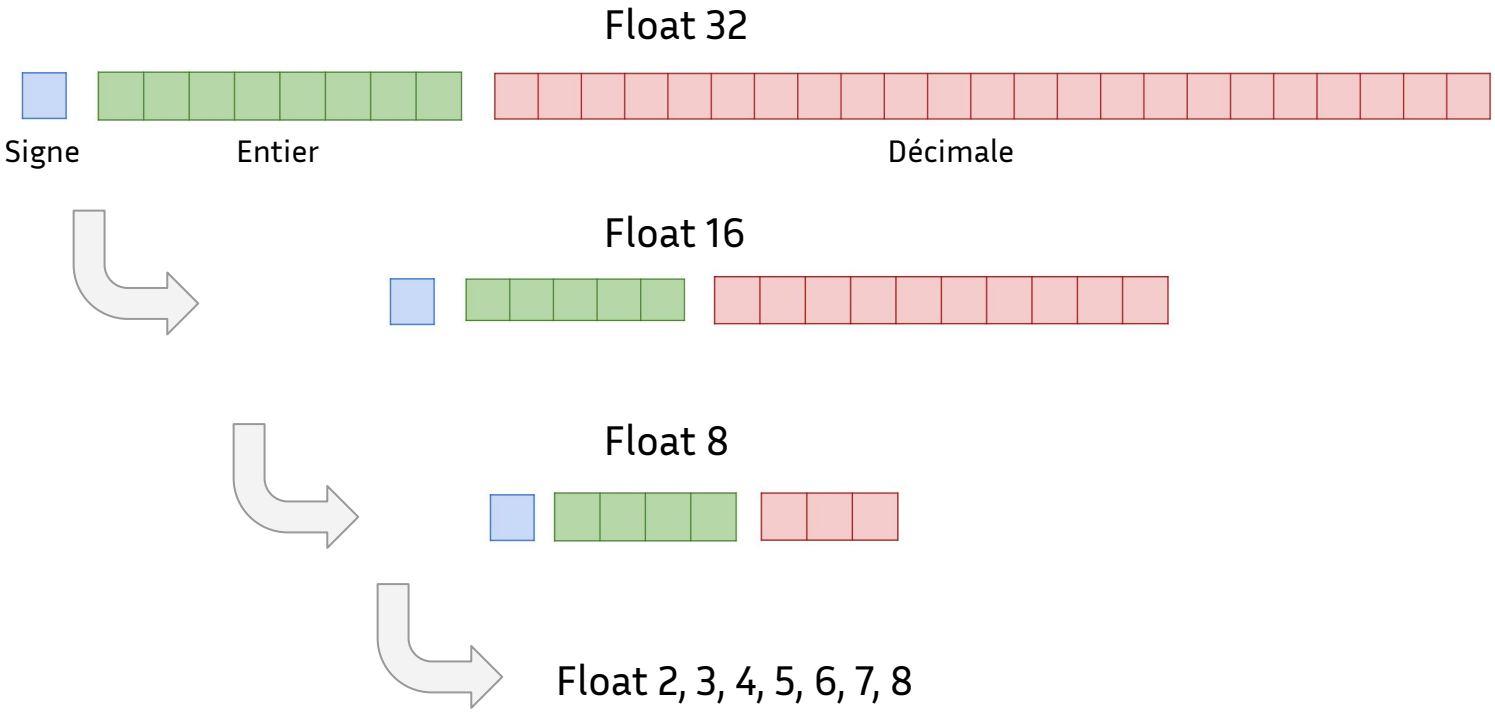
Quantification



Quantification

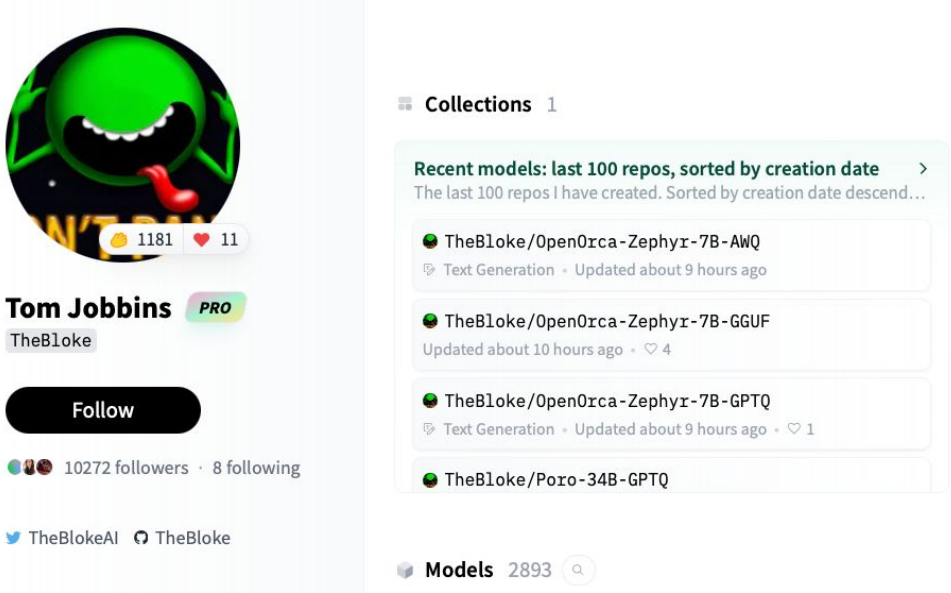


Quantification



Le github des modèles d'IA : HuggingFace

> https://huggingface.co/TheBloke



C’est parti pour la pratique !

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