


# Fine-tuning d'un LLM

# Présentation

- Personnaliser un LLM
- Comment fine-tuner ?
- Travaux pratiques 
- Références

# Personnalisier ein LLM

- Prompt Engineering
- Retrieval-Augmented Generation (RAG)
- Fine-tuning
- Full training

# Prompt engineering



**You**

A "whatpu" is a small, furry animal native to Tanzania. An example of a sentence that uses the word whatpu is:

We were traveling in Africa and we saw these very cute whatpus.

To do a "farduddle" means to jump up and down really fast. An example of a sentence that uses the word farduddle is:

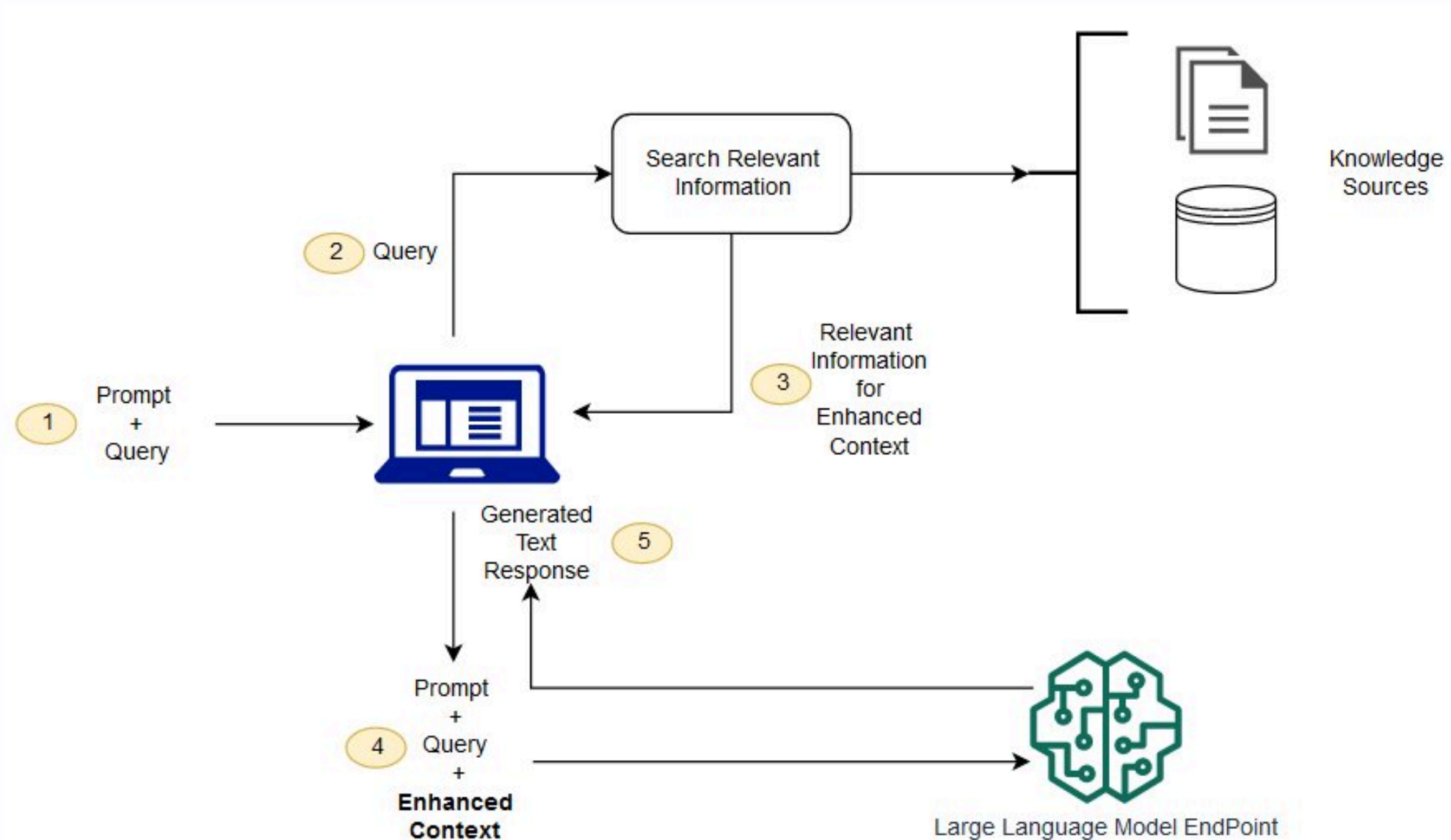


**ChatGPT**

We were so excited about the concert that we started to farduddle in the front row, much to the amusement of the band.



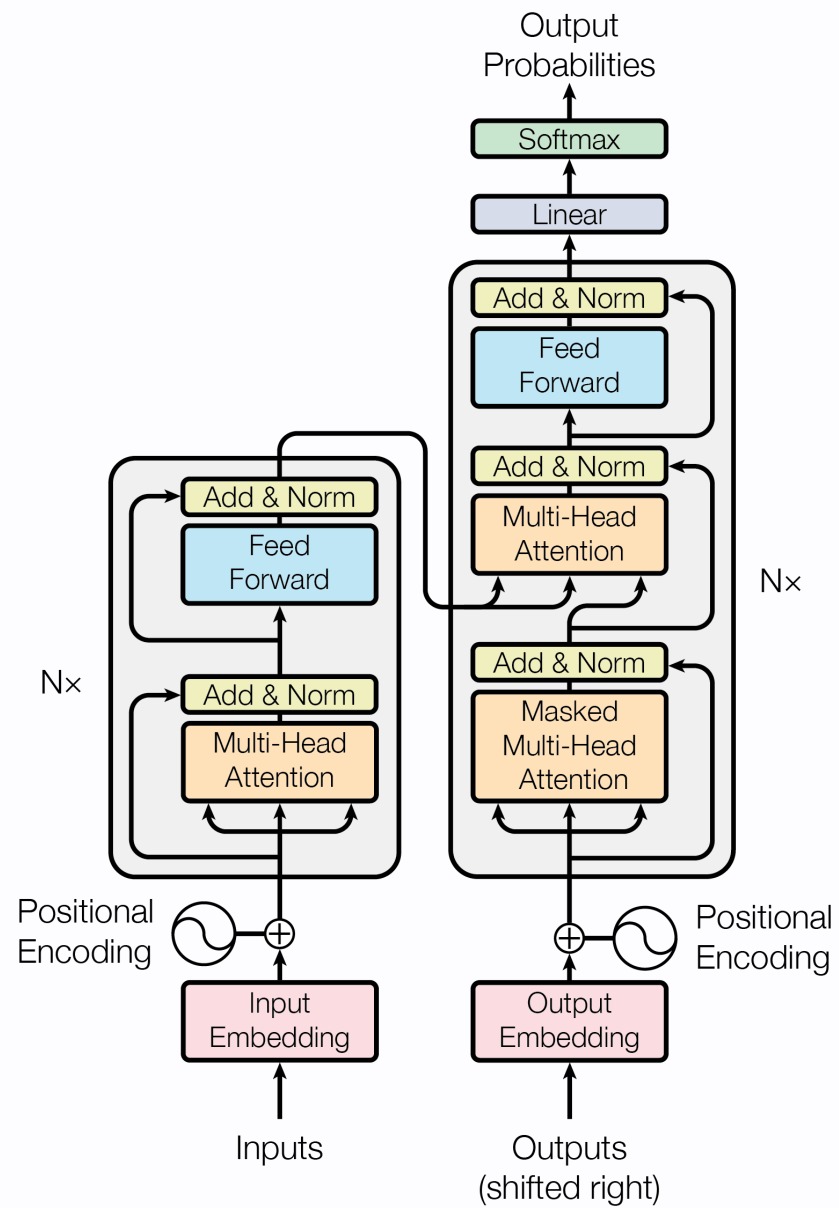
# Retrieval-Augmented Generation (RAG)



# Comment fine-tuner ?

On part avec :

- un **LLM**
- un **dataset** spécifique à notre domaine ou tâche
- une méthode d'**évaluation**



# Qu'est-ce qu'on modifie ?

LLama 7B, 13B, 70B c'est :

7 000 000 000

13 000 000 000

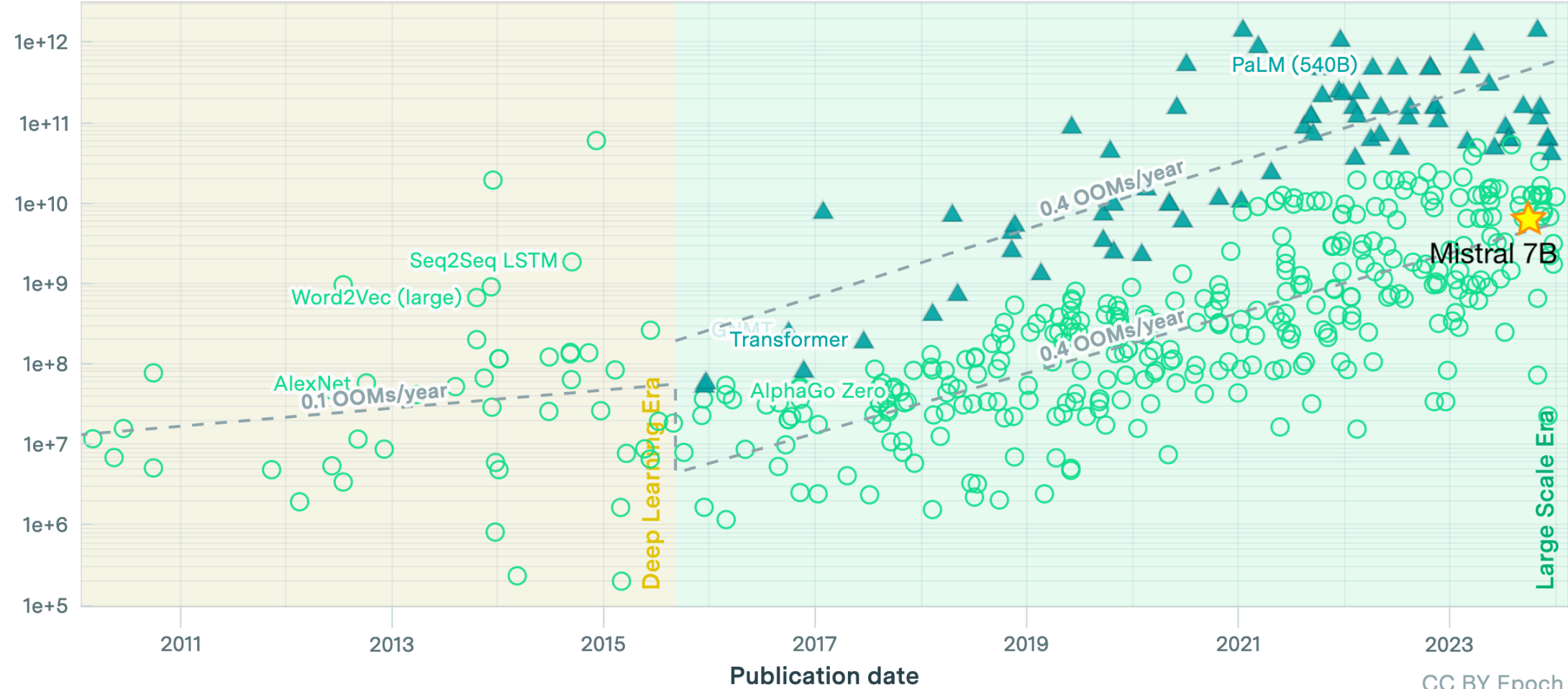
70 000 000 000

paramètres !



# Model Size of Notable machine learning Systems Over Time

Number of trainable parameters



CC BY Epoch

# ML CO2 Impact

***Quantifying the Carbon Emissions of Machine Learning***

<https://mlco2.github.io/impact>

# Apprentissage

L'apprentissage d'un LLM est très coûteux :

- Calcul (FLOPS)
- Mémoire

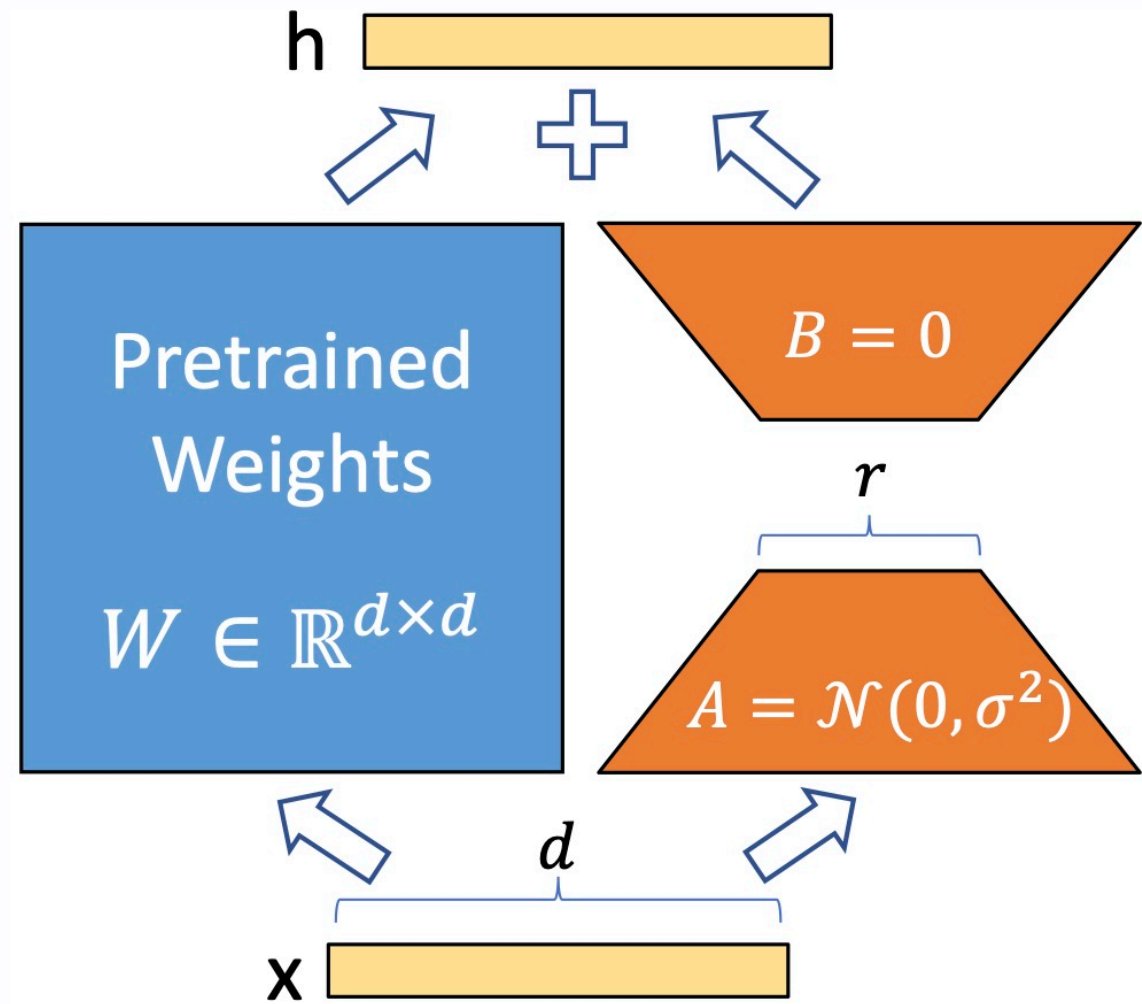


# PEFT !

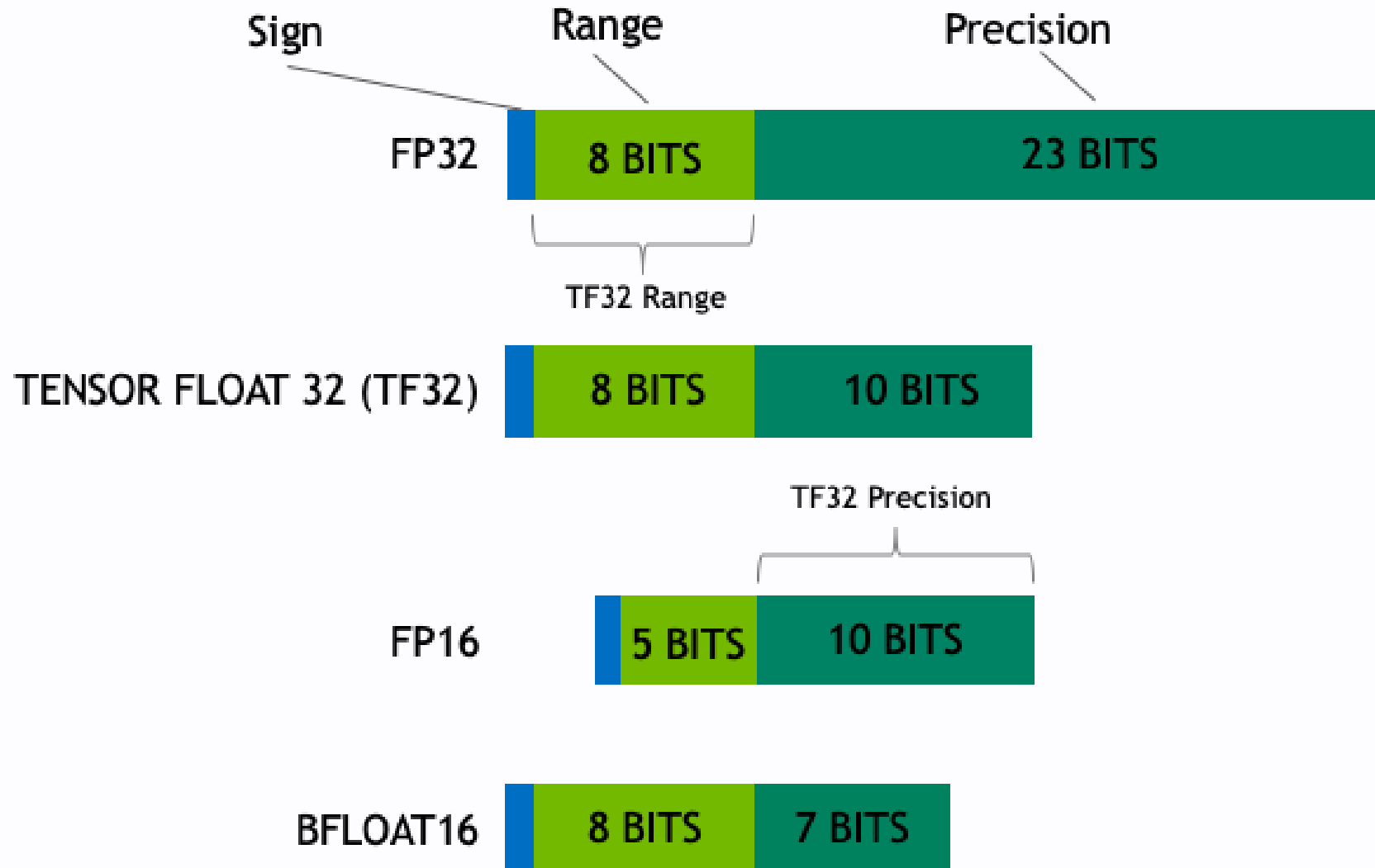
## Parameter Efficient Fine-Tuning

- méthodes sélectives
- méthodes de reparamétrisation (LoRA)
- méthodes additives

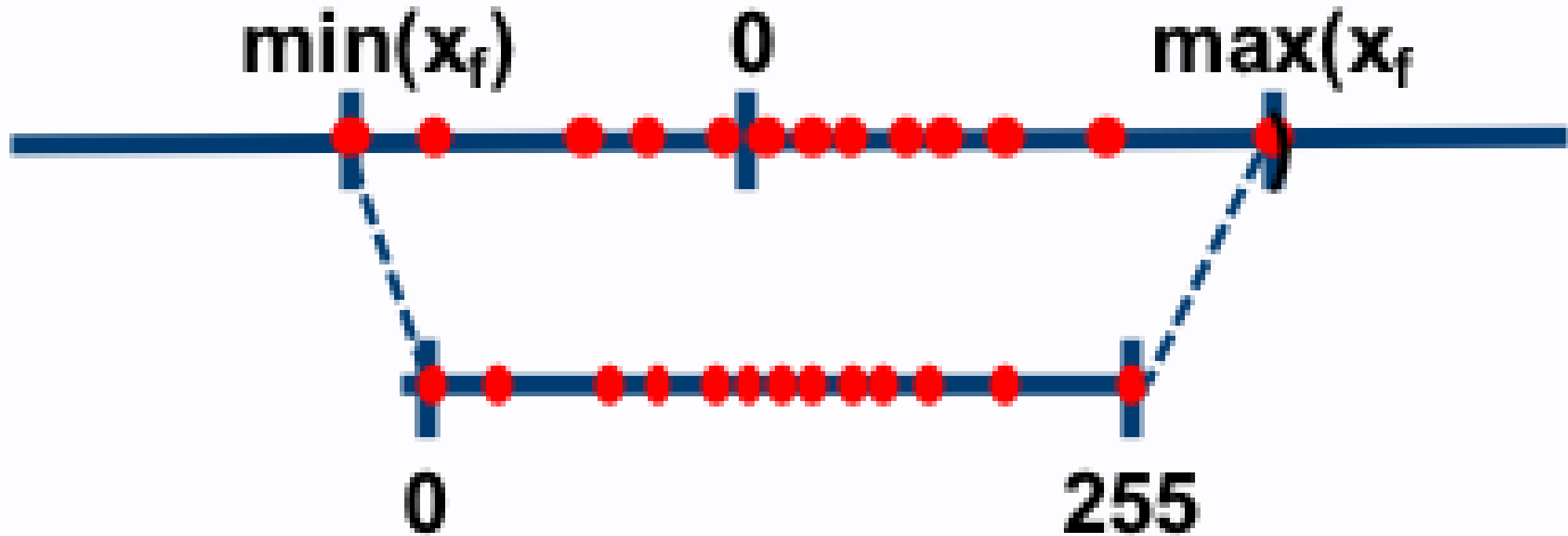
# LoRA: Low-Rank Adaptation



# Quantification (1/2)



# Quantification (2/2)



# Méthodes d'alignement

- Reinforcement Learning by Human Feedback (RLHF)
- Direct Preference Optimization (DPO)



# TP

Notebook de fine-tuning  
Supervision du fine-tuning  
Notebook d'inférence (vanilla)

# Références

# Ressources

- Cours de NLP de Hugging Face
- Article de Maxime Labonne
- Article de HelixML
- Explication de la quantification

# Outils

## Infra

- Google Colab
- GPU dédié sur runpod.io
- Inference sur octoai.cloud

# Librairies

- Apprentissage & inférence
  - [Hugging Face libs](#)
  - [Unsloth](#)
- Supervision de l'apprentissage
  - [Weights and Biases \(wandb\)](#)
- Quantification
  - [bitsandbytes](#)
  - [quanto \(Hugging Face\)](#)