Université Paris Cité

CM Master 2 AMSD / MLSD

Comprendre et utiliser le NLP en entreprise : De la théorie à la pratique

Enseignant : Hervé Trinh
Mail : trinh.herve@gmail.com

Université Paris Cité

Sommaire

I. Introduction aux LLMs

II. Optimisation des LLMs

Introduction aux LLMs

- **2 types de LLMs :**
- BERT / RoBERTA models (100M – 300M) , T5 models (220M, 770M, 3B)
- "Very" LLMs => 100+ B paramètres

- **1 modèle pour résoudre toutes les tâches de NLP ?**

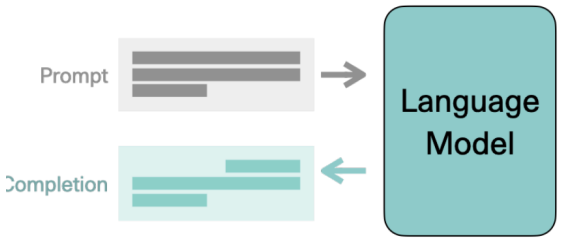


Image credit: Jay Alammar

Introduction aux LLMs

- **Rappel : BERT (Bidirectional Encoder Representations from Transformers)**

Pre-training objectives:

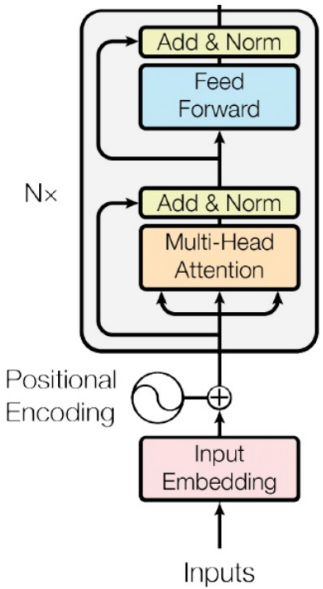
masked language modeling + next sentence prediction

MLM => Masquer **15%** tokens aléatoires et prédire les mots masqués :

Je m'**<mask>** Hervé et je suis **<mask>** scientist

NSP => Prédire si la 2e phrase de la paire suit logiquement la 1ere (supprimé dans RoBERTa / CamemBERT)

Introduction aux LLMs

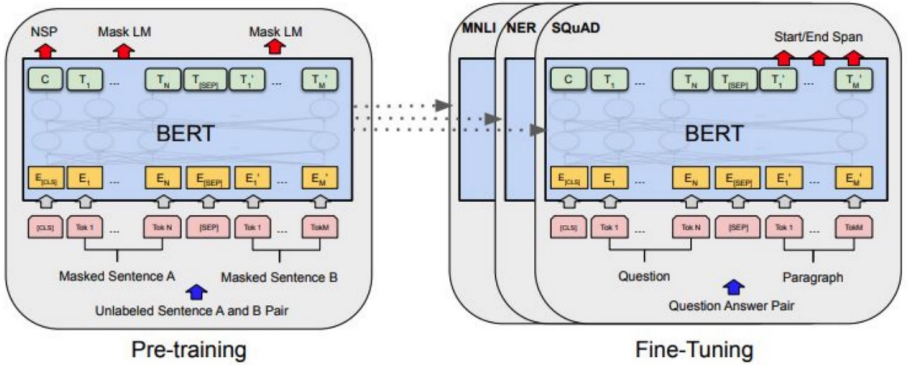


BERT-base : 12 couches, 768 hidden size, 12 attention heads, 110M parameters

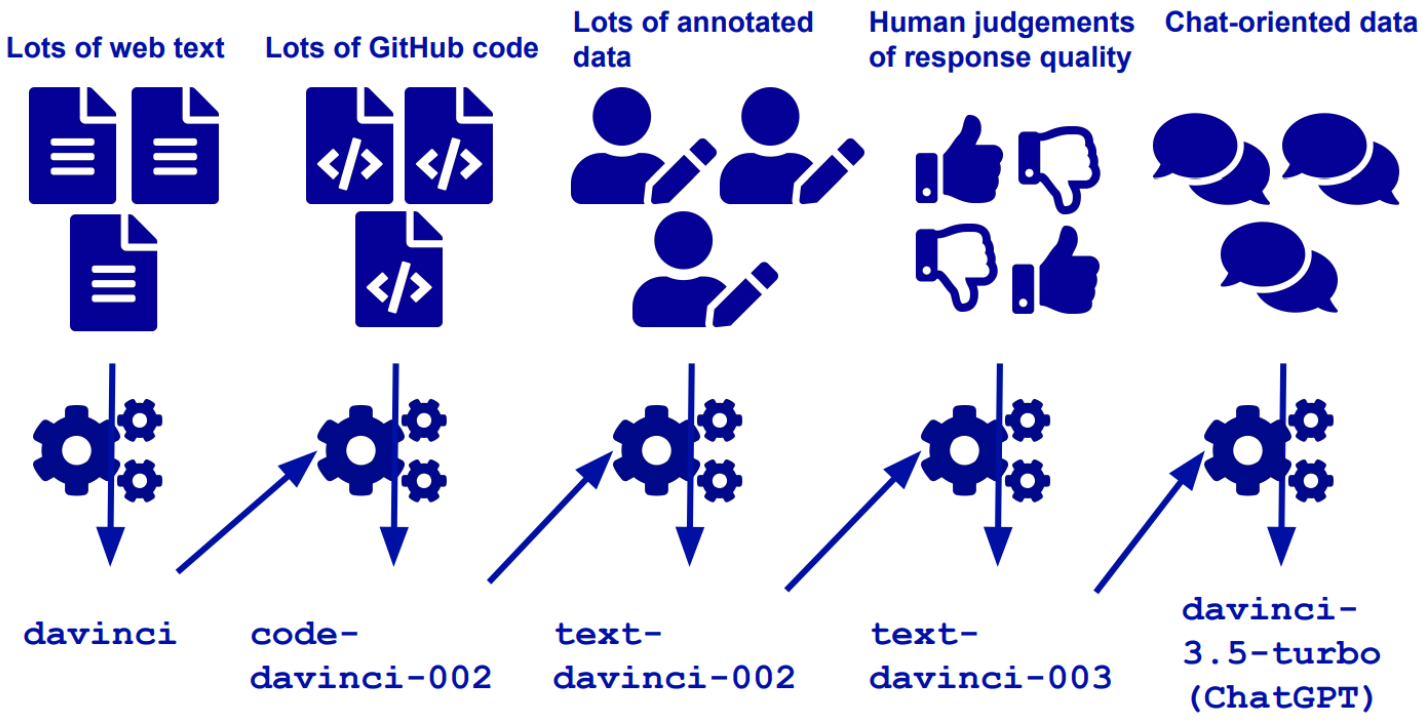
BERT-large : 24 couches, 1024 hidden size, 16 attention heads, 340M parameters

Training corpus : Wikipedia (2.5B) + BooksCorpus (0.8B)

Max sequence size: 512 tokens



Introduction aux LLMs



Introduction aux LLMs

Rouge

- **Métrique pour le résumé de texte**

$$\text{ROUGE-1 Recall} = \frac{\text{unigram matches}}{\text{unigrams in reference}}$$

$$\text{ROUGE-1 Precision:} = \frac{\text{unigram matches}}{\text{unigrams in output}}$$

$$\text{ROUGE-1 F1:} = 2\,\frac{\text{precision x recall}}{\text{precision + recall}}$$

Bleu score

- **Métrique pour la traduction**

Optimisation des LLMs

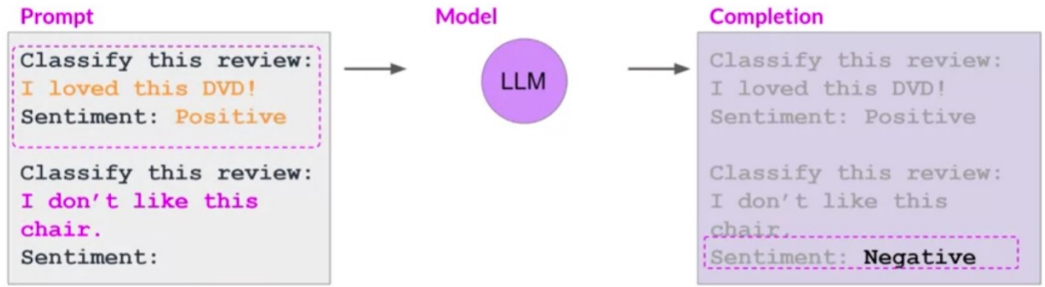
Prompt Engineering

- Ecrire des instructions claires
- Demander au LLM d'incarner un rôle
- Utiliser des délimiteurs pour distinguer les instructions et l'input
- Spécifier des étapes
- Donner des exemples (one shot / few shot learning)
- Spécifier la longueur désirée de l'output
- Spécifier au LLM de répondre avec des citations
- ...

https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions

Optimisation des LLMs

- **Exemple d'un one shot learning**

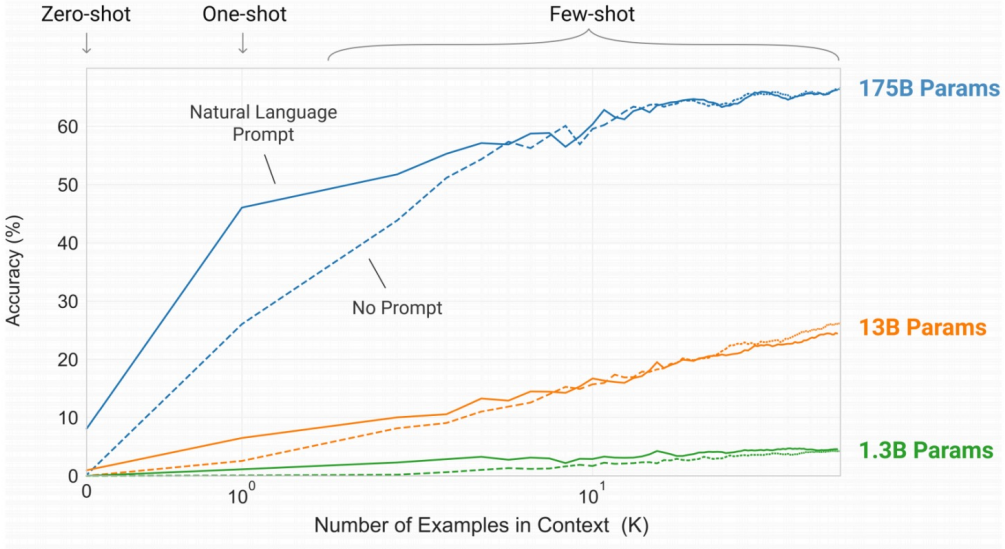


One-shot or Few-shot Inference

Inconvénient :
Consomme de la place pour le contexte@

Optimisation des LLMs

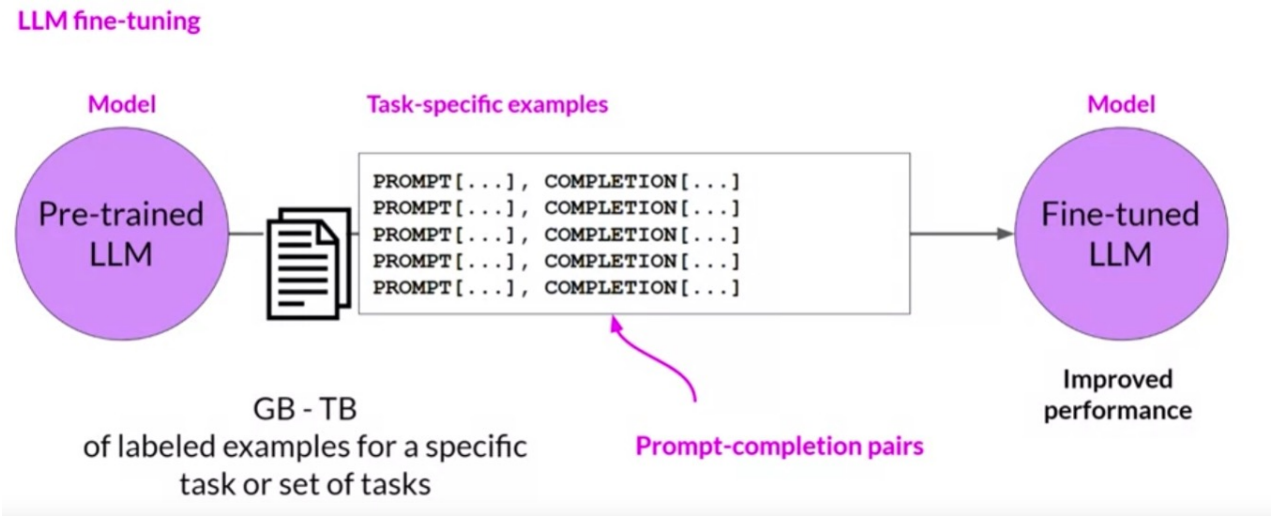
- **Few-shot learning**




```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← examples
3  peppermint => menthe poivrée        ← examples
4  plush girafe => girafe peluche      ← examples
5  cheese =>                           ← prompt
```

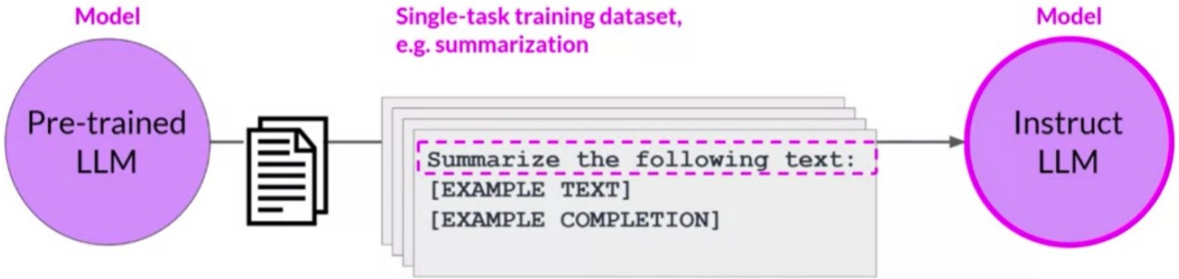
Optimisation des LLMs

- **Instruction Fine tuning**



Optimisation des LLMs

- **Instruction Fine tuning single task**



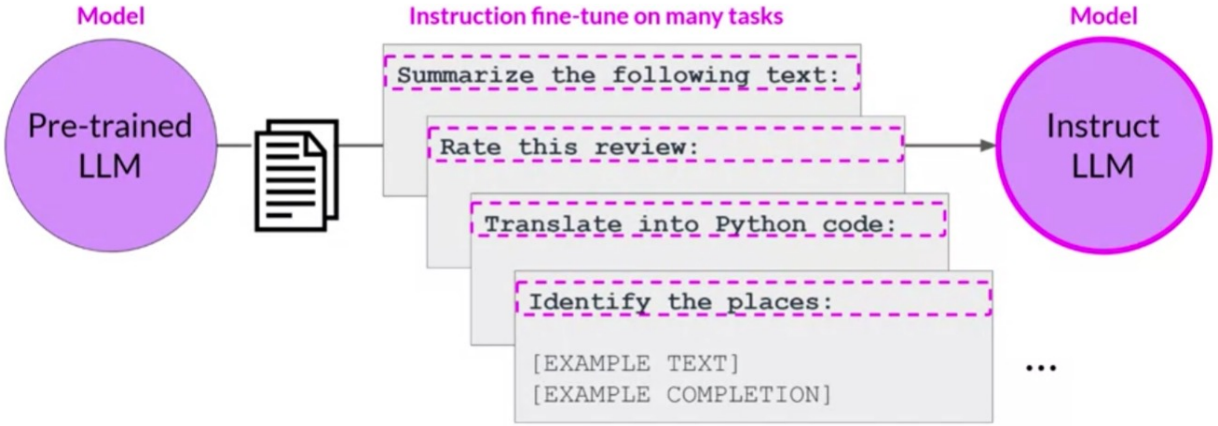
Seulement besoin de 500 – 1000 exemples pour des bons résultats!

Mais réduit les perfs des autres tasks

Solution :
- multiple task fine tuning
- Parameter-Efficient Fine-Tuning

Optimisation des LLMs

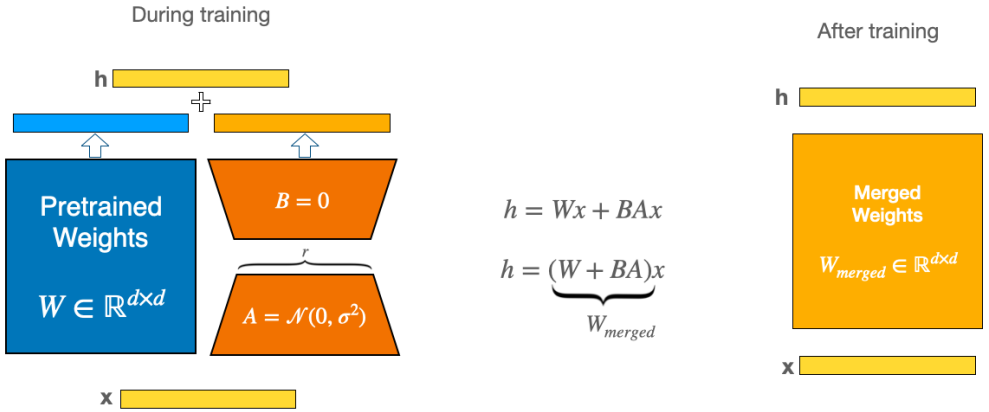
- **Multi task instruction fine tuning**



Inconvénient : 50K – 100K exemples

Optimisation des LLMs

- **LoRA**



https://huggingface.co/docs/peft/conceptual_guides/lora

https://arxiv.org/pdf/2106.09685.pdf