

Large Language Models

Clément Romac (Hugging Face & Inria)

clement.romac@gmail.com

https://github.com/ClementRomac/Teaching/tree/main/ENSC3A_LLMs_2025-2026

- Les RNNs:
 - souffrent de gradient vanishing
 - sont très bien adaptés au très longs exemples
 - ne permettent pas de générer du texte
-

Petit Quizz

Petit Quizz



- Les RNNs:
 - souffrent de gradient vanishing
 - sont très bien adaptés au très longs exemples
 - ne permettent pas de générer du texte
- Le mécanisme d'attention:
 - se base sur une moyenne pondérée d'entrées
 - a été très utilisé pour des problèmes de traduction
 - est apparu avec le papier introduisant l'architecture Transformer

Petit Quizz



- Les RNNs:
 - souffrent de gradient vanishing
 - sont très bien adaptés au très longs exemples
 - ne permettent pas de générer du texte
- Le mécanisme d'attention:
 - se base sur une moyenne pondérée d'entrées
 - a été très utilisé pour des problèmes de traduction
 - est apparu avec le papier introduisant l'architecture Transformer
- Un Transformer:
 - est constitué d'un Encoder et un Decoder
 - utilise des LSTMs
 - n'est constitué que de couches de Self-Attention (aucune autre architecture de réseau de neurones)

Petit Quizz



- Les RNNs:
 - souffrent de gradient vanishing
 - sont très bien adaptés au très longs exemples
 - ne permettent pas de générer du texte
- Le mécanisme d'attention:
 - se base sur une moyenne pondérée d'entrées
 - a été très utilisé pour des problèmes de traduction
 - est apparu avec le papier introduisant l'architecture Transformer
- Un Transformer:
 - est constitué d'un Encoder et un Decoder
 - utilise des LSTMs
 - n'est constitué que de couches de Self-Attention (aucune autre architecture de réseau de neurones)
- Un Transformer:
 - contient pour chaque bloc plusieurs têtes donc les sorties sont concaténées (appelé multi-head)
 - contient un Decoder qui utilise les sorties de l'encoder comme Queries
 - contient plusieurs blocs "empilés" (appelé multi-hop)

Petit Quizz



- Les RNNs:
 - souffrent de gradient vanishing
 - sont très bien adaptés au très longs exemples
 - ne permettent pas de générer du texte
- Le mécanisme d'attention:
 - se base sur une moyenne pondérée d'entrées
 - a été très utilisé pour des problèmes de traduction
 - est apparu avec le papier introduisant l'architecture Transformer
- Un Transformer:
 - est constitué d'un Encoder et un Decoder
 - utilise des LSTMs
 - n'est constitué que de couches de Self-Attention (aucune autre architecture de réseau de neurones)
- Un Transformer:
 - contient pour chaque bloc plusieurs têtes donc les sorties sont concaténées (appelé multi-head)
 - contient un Decoder qui utilise les sorties de l'encoder comme Queries
 - contient plusieurs blocs "empilés" (appelé multi-hop)
- Dans un Transformer:
 - l'Encoder est "bi-directionnel" (toute la séquence est utilisée, les opérations sur chaque token utilisent les tokens d'avant mais aussi ceux d'après)
 - Le Decoder est "bi-directionnel" (toute la séquence est utilisée, les opérations sur chaque token utilisent les tokens d'avant mais aussi ceux d'après)

Petit Quizz



- Les RNNs:
 - souffrent de gradient vanishing
 - sont très bien adaptés au très longs exemples
 - ne permettent pas de générer du texte
- Le mécanisme d'attention:
 - se base sur une moyenne pondérée d'entrées
 - a été très utilisé pour des problèmes de traduction
 - est apparu avec le papier introduisant l'architecture Transformer
- Un Transformer:
 - est constitué d'un Encoder et un Decoder
 - utilise des LSTMs
 - n'est constitué que de couches de Self-Attention (aucune autre architecture de réseau de neurones)
- Un Transformer:
 - contient pour chaque bloc plusieurs têtes donc les sorties sont concaténées (appelé multi-head)
 - contient un Decoder qui utilise les sorties de l'encoder comme Queries
 - contient plusieurs blocs "empilés" (appelé multi-hop)
- Dans un Transformer:
 - l'Encoder est "bi-directionnel" (toute la séquence est utilisée, les opérations sur chaque token utilisent les tokens d'avant mais aussi ceux d'après)
 - Le Decoder est "bi-directionnel" (toute la séquence est utilisée, les opérations sur chaque token utilisent les tokens d'avant mais aussi ceux d'après)

Contenu

- Language Modeling objectives
- Encoder-only (e.g. BERT)
- Decoder-only (e.g. GPT)
- Prompting
- Chat models

A retenir

- Masked Language Modeling
- Causal Language Modeling
- Transfer Learning
- Alignement

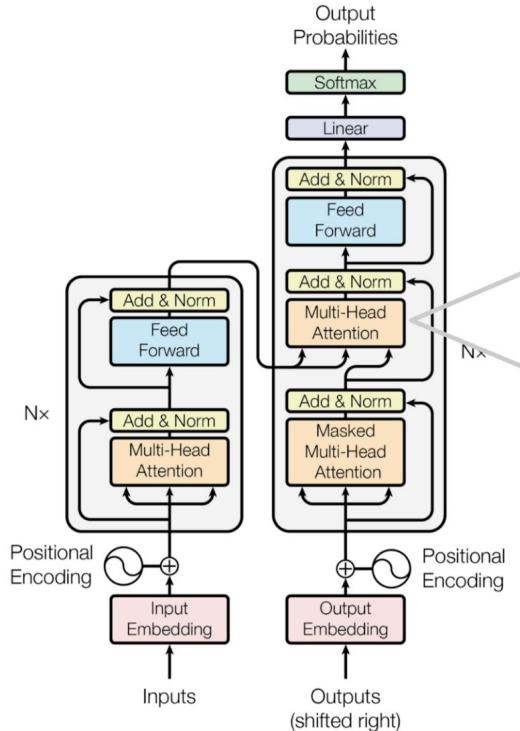
Ressources

Lectures:

- <https://huggingface.co/learn/nlp-course/>
- <https://transformersbook.com/>
- <https://arxiv.org/abs/1910.10683> (T5)

Aux origines: le Transformer

Original Transformer



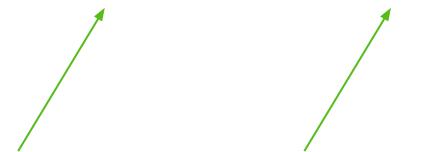
Trained on WMT EN-GER or EN-FR

Table 2: The Transformer achieves better BLEU scores than previous state-of-the-art models on the English-to-German and English-to-French newstest2014 tests at a fraction of the training cost.

Model	BLEU		Training Cost (FLOPs)	
	EN-DE	EN-FR	EN-DE	EN-FR
ByteNet [18]	23.75			
Deep-Att + PosUnk [39]		39.2		$1.0 \cdot 10^{20}$
GNMT + RL [38]	24.6	39.92	$2.3 \cdot 10^{19}$	$1.4 \cdot 10^{20}$
ConvS2S [9]	25.16	40.46	$9.6 \cdot 10^{18}$	$1.5 \cdot 10^{20}$
MoE [32]	26.03	40.56	$2.0 \cdot 10^{19}$	$1.2 \cdot 10^{20}$
Deep-Att + PosUnk Ensemble [39]		40.4		$8.0 \cdot 10^{20}$
GNMT + RL Ensemble [38]	26.30	41.16	$1.8 \cdot 10^{20}$	$1.1 \cdot 10^{21}$
ConvS2S Ensemble [9]	26.36	41.29	$7.7 \cdot 10^{19}$	$1.2 \cdot 10^{21}$
Transformer (base model)	27.3	38.1		$3.3 \cdot 10^{18}$
Transformer (big)	28.4	41.8		$2.3 \cdot 10^{19}$

(Causal) Language Modeling

Given a corpus of tokens: $U = \{u_1, \dots, u_N\}$

$$\max_{\theta} L(U) = \sum_i \log P_{\theta}(u_i | u_{i-k}, \dots, u_{i-1})$$


model context window

(Causal) Language Modeling

En pratique:

- Etant donné un **corpus de texte tokenisé**
- On casse le corpus en **blocs de taille k**
- On souhaite apprendre un **modèle** qui maximise la probabilité de chaque **token** d'apparaître dans sa séquence

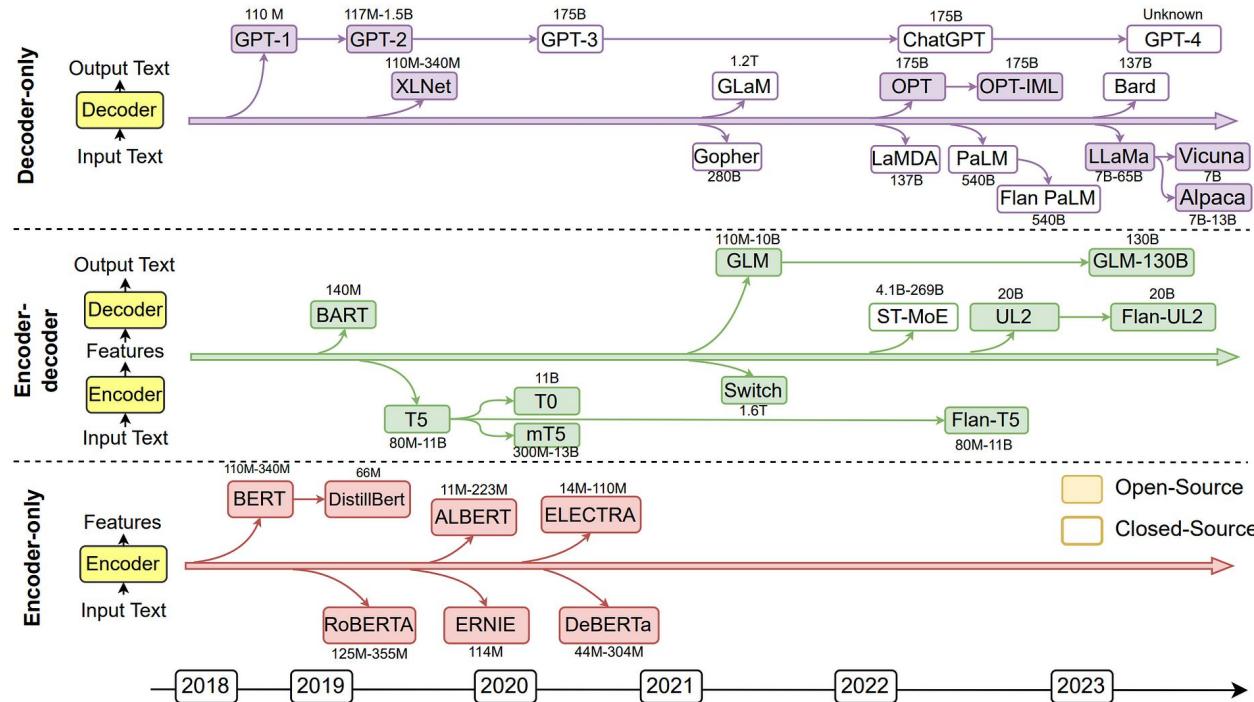
Du Transformer aux LLMs

Scaling up & Transfer Learning

A partir de 2018/2019:

- De nombreux travaux se mettent à entraîner des Transformers sur des **gros corpus généraux**

Panel des LLMs

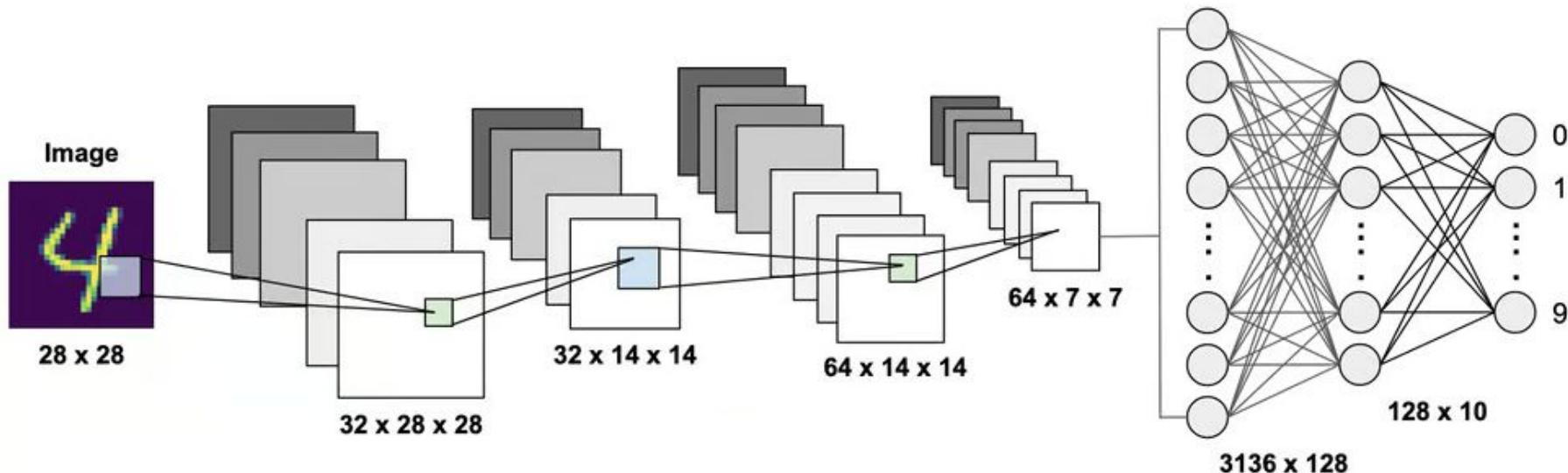


Scaling up & Transfer Learning

A partir de 2018/2019:

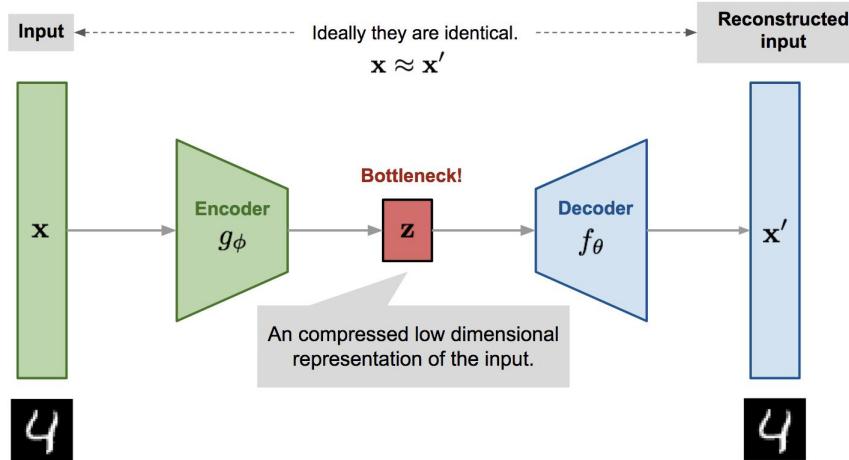
- De nombreux travaux se mettent à entraîner des Transformers sur des **gros corpus généraux**
- Le modèle obtenu est ensuite utilisé comme base **pré-entraînée**

Une inspiration: Computer Vision



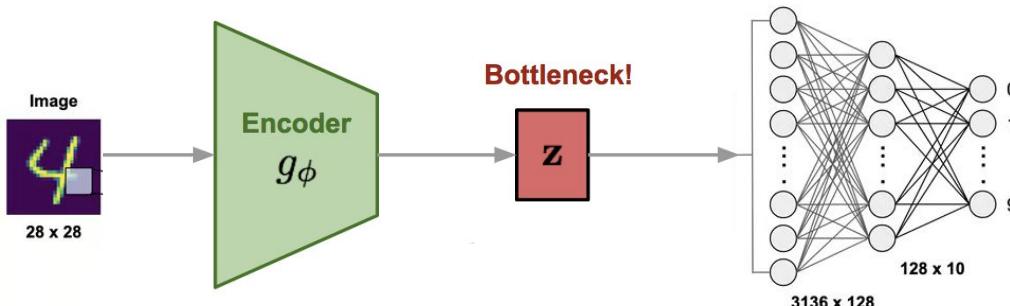
Comment avoir suffisamment de données pour apprendre une bonne représentation ?

Une inspiration: Computer Vision



1. On pré-entraîne notre réseau de neurones avec des données diverses

Une inspiration: Computer Vision



1. On pré-entraîne notre réseau de neurones avec des données diverses
1. On affine notre réseau avec des données liées à notre problème

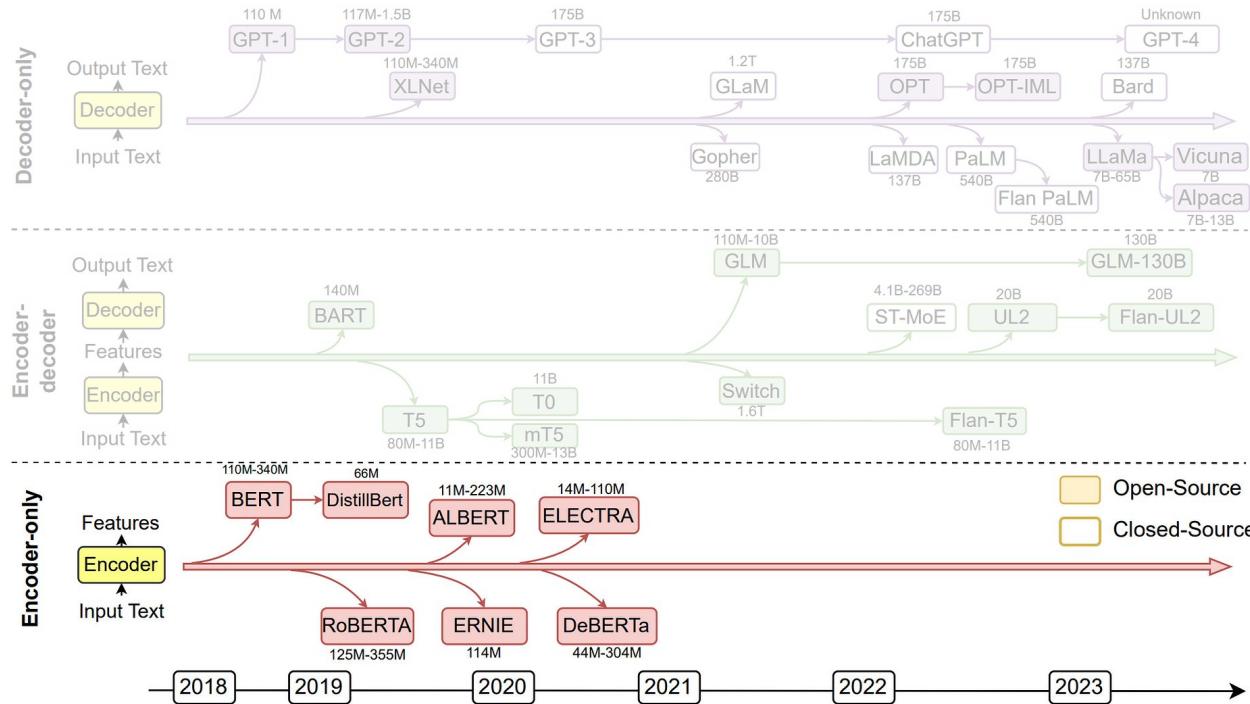
Scaling up & Transfer Learning

A partir de 2018/2019:

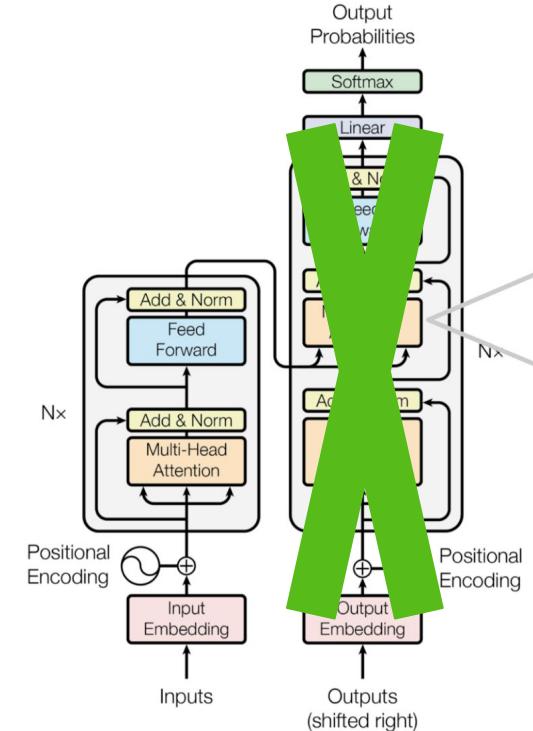
- De nombreux travaux se mettent à entraîner des Transformers sur des **gros corpus généraux**
- Le modèle obtenu est ensuite utilisé comme base **pré-entraînée**
- Un premier courant apparaît: **Encoder-Only**

Encoder-Only LLMs

Panel des LLMs



Encoder-Only as Foundation Models

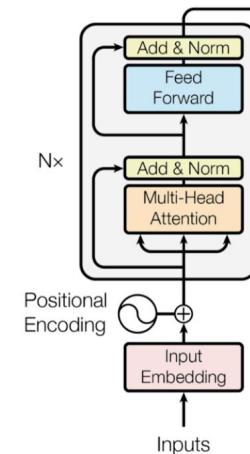


Encoder-Only as Foundation Models

Objectif

- Encoder un texte et obtenir une représentation exploitable pour différentes tâches
- Apprendre ensuite une “tête” spécifique pour un problème

=> Transfer Learning



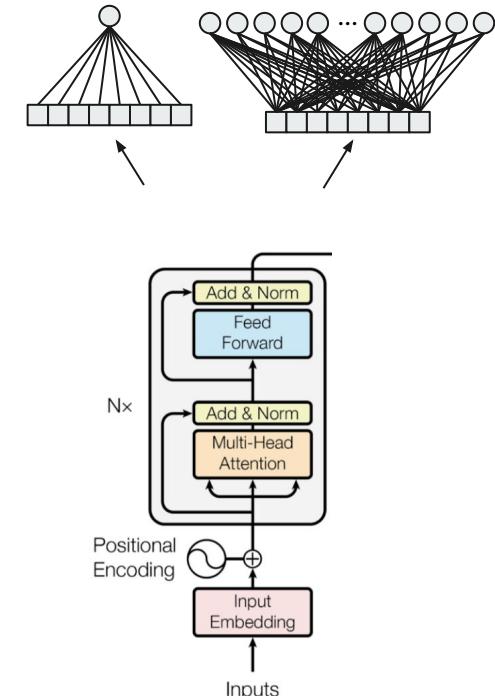
Encoder-Only as Foundation Models

Objectif

- Encoder un texte et obtenir une représentation exploitable pour différentes tâches
- Apprendre ensuite une "tête" spécifique pour un problème

=> Transfer Learning

1. (Pré-) Entrainer le LM
1. Entrainer des têtes
 - classification (topics, sentiment analysis...)



BERT (Devlin et al., 2018)

Masked Language Modeling objective

MLM:

- On remplace aléatoirement certains tokens par un token de mask
- Le modèle doit reconstruire la séquence
- Self-Supervised Learning (SSL)

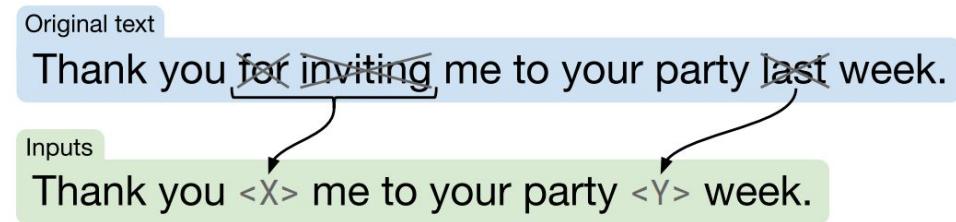
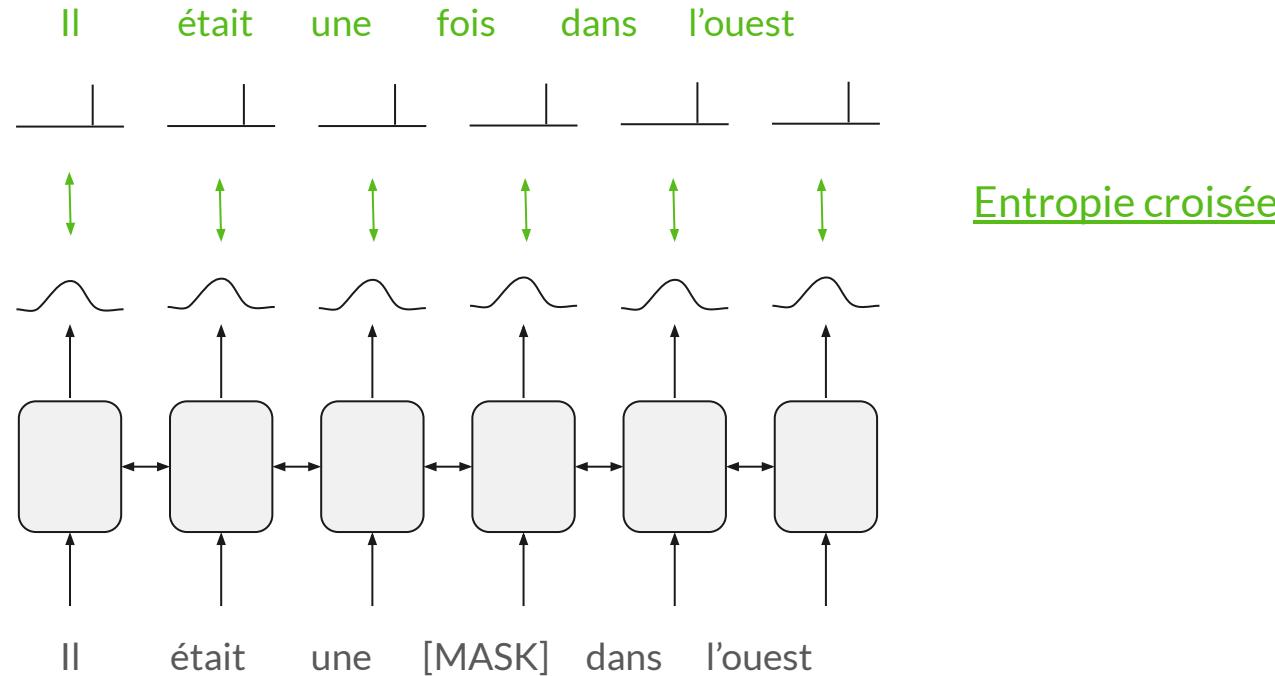


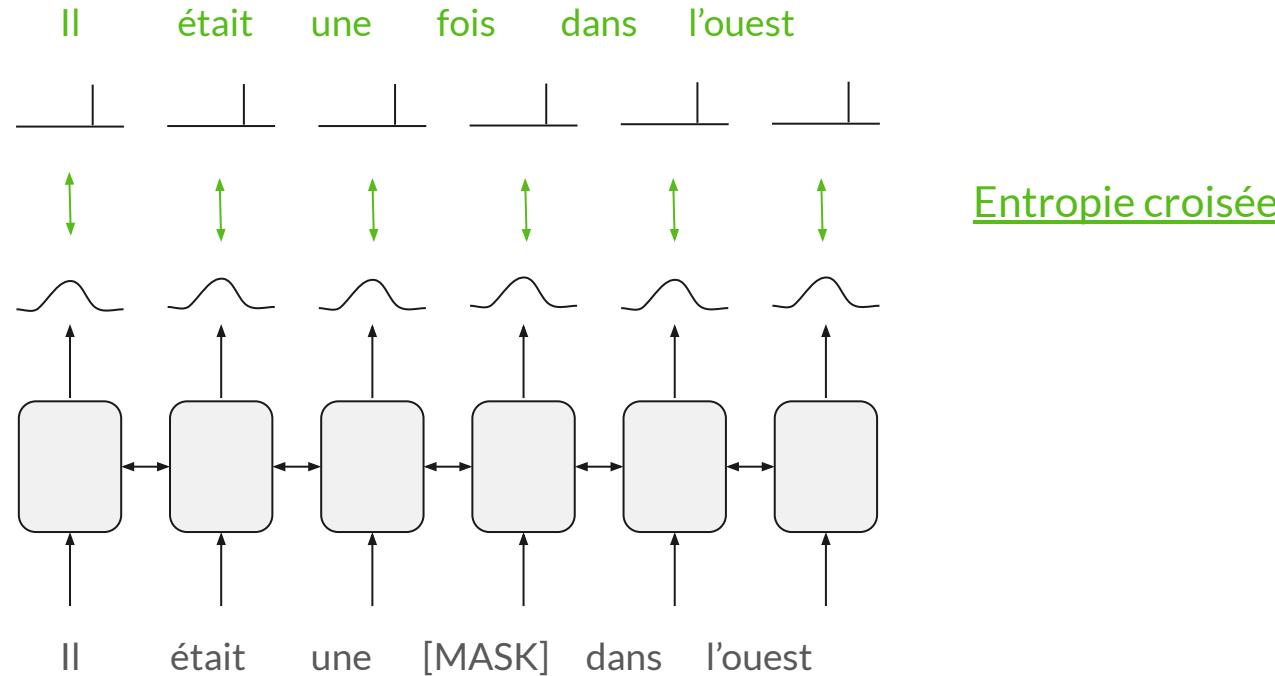
Image for T5 (Raffel et al., 2019)

Masked Language Modeling objective



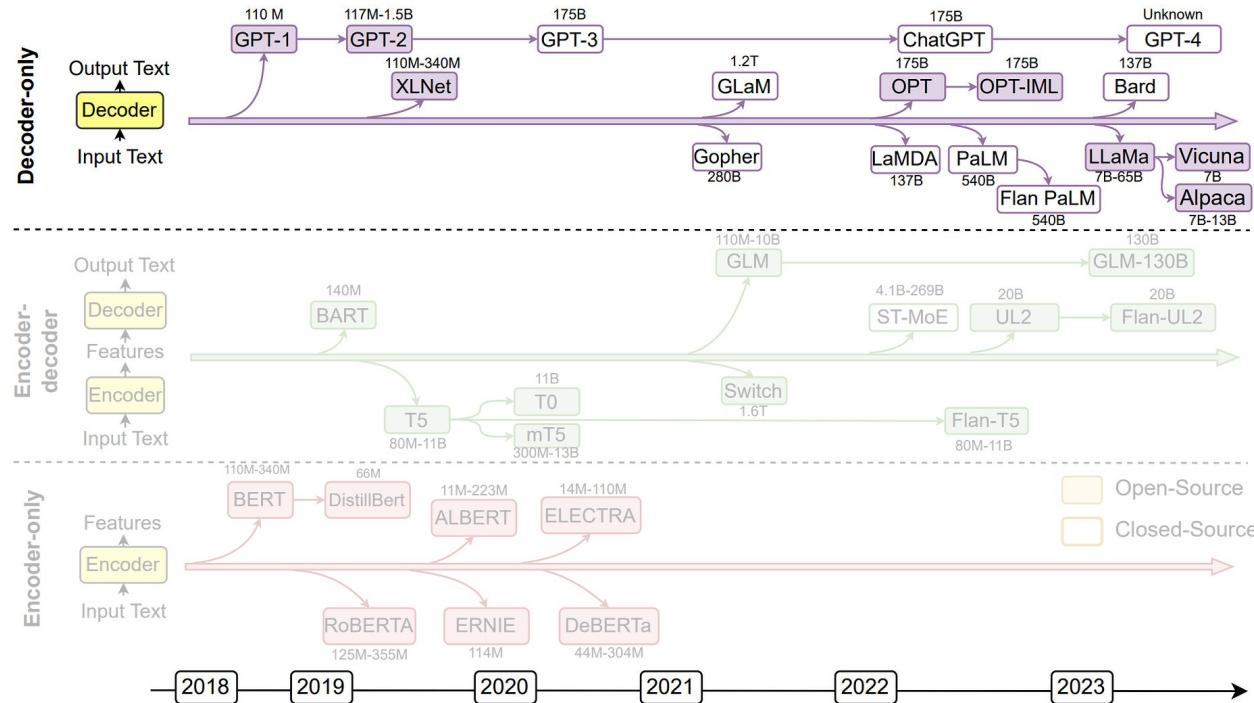


Masked Language Modeling objective



Decoder-Only LLMs

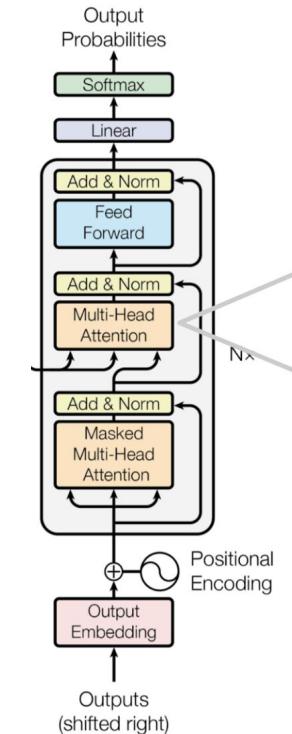
Panel des LLMs



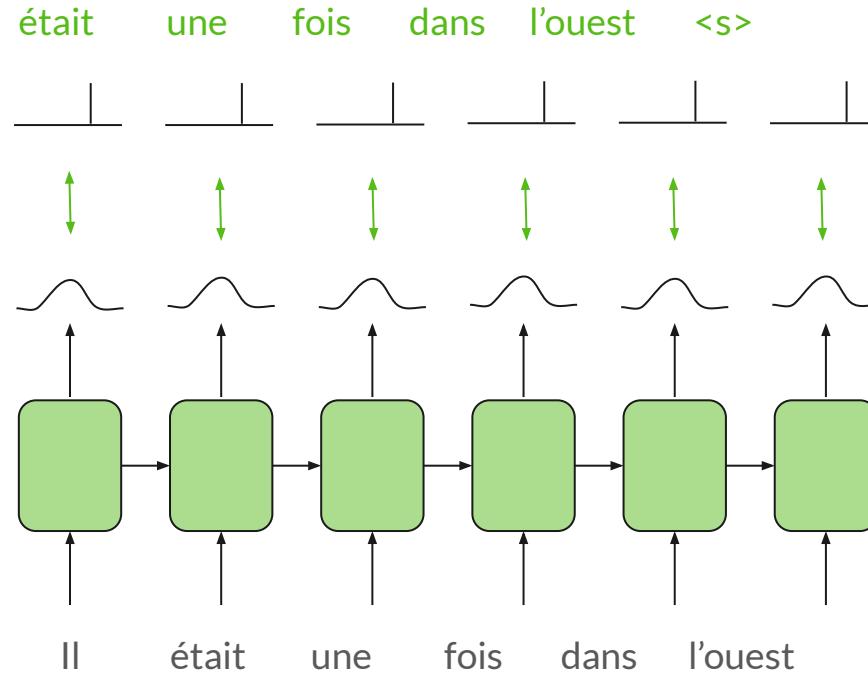
Decoder-Only Models

Objectif

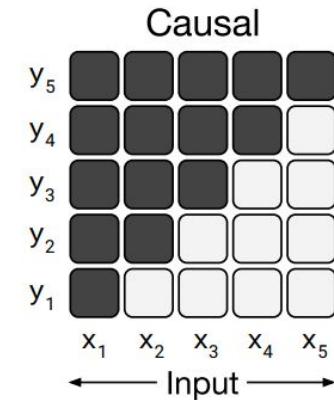
- N'utiliser **QUE** le Language Modeling objective
- **Génération** de la suite de chaque bloc de texte



Causal Language Modeling objective



Entropie croisée



Generative Pre-Training (GPT) *(Radford et al., 2018)*

- GPT-1 partage le principe de **Foundation Model**
(Pre-entraînement + Finetuning)
- Introduit un **formatage de l'entrée** qui est spécifique à chaque tâche (e.g. ajouter la réponse après la question)

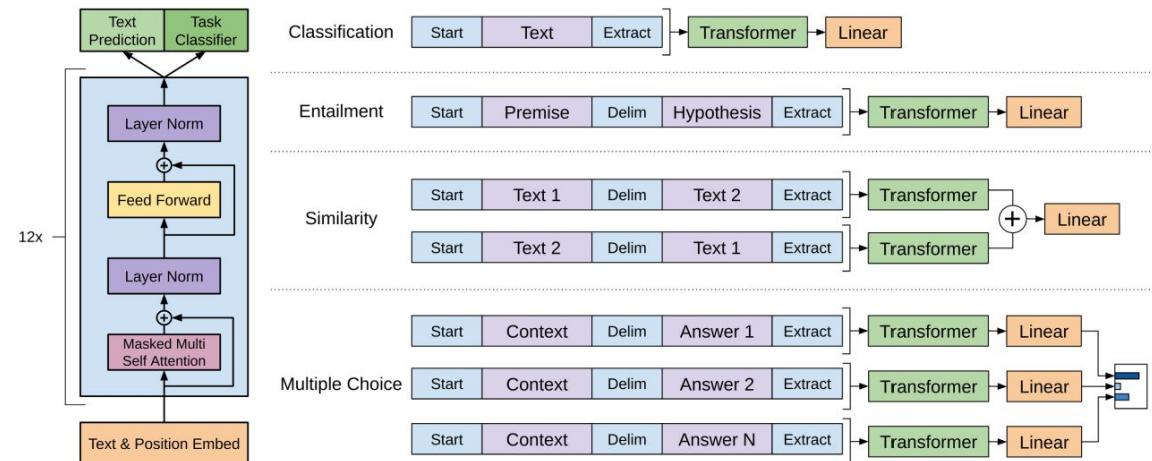
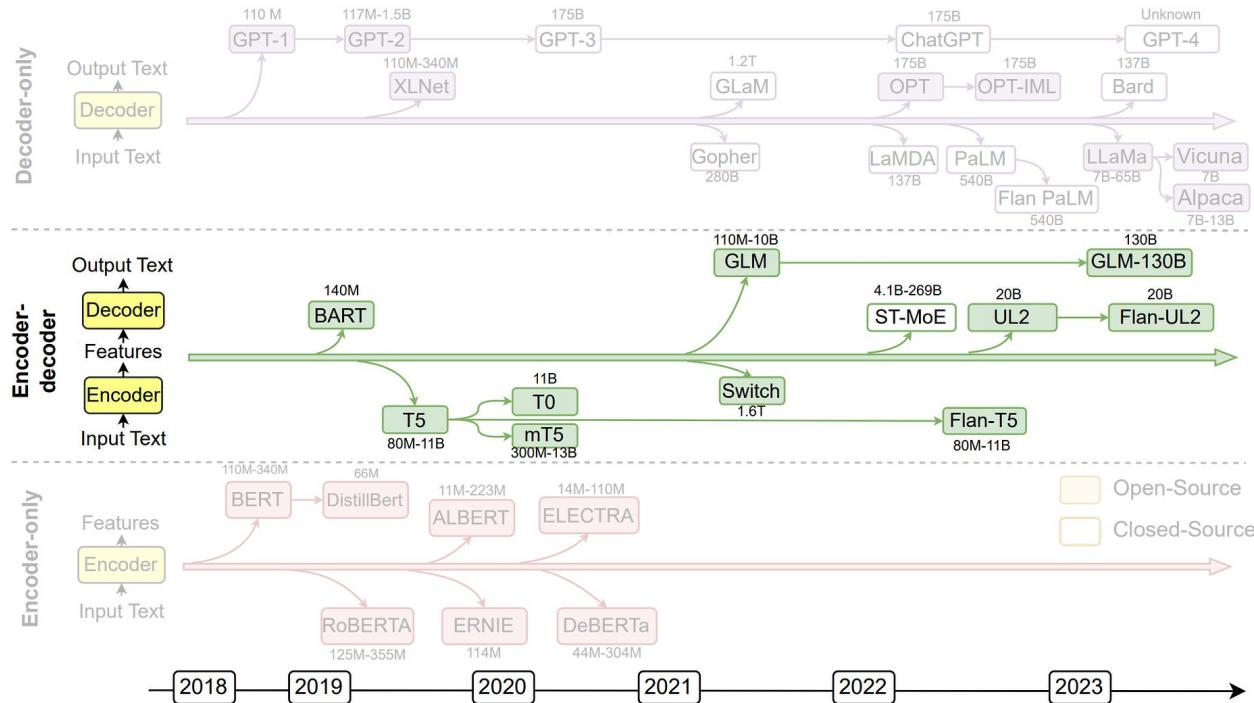


Figure 1: **(left)** Transformer architecture and training objectives used in this work. **(right)** Input transformations for fine-tuning on different tasks. We convert all structured inputs into token sequences to be processed by our pre-trained model, followed by a linear+softmax layer.

Et l'architecture complète dans
tout ça ?

Panel des LLMs



T5

(Raffel et al., 2019)

Architecture	Objective	Params	Cost	GLUE	CNNDM	SQuAD	SGLUE	EnDe	EnFr	EnRo
★ Encoder-decoder	Denoising	$2P$	M	83.28	19.24	80.88	71.36	26.98	39.82	27.65
Enc-dec, shared	Denoising	P	M	82.81	18.78	80.63	70.73	26.72	39.03	27.46
Enc-dec, 6 layers	Denoising	P	$M/2$	80.88	18.97	77.59	68.42	26.38	38.40	26.95
Language model	Denoising	P	M	74.70	17.93	61.14	55.02	25.09	35.28	25.86
Prefix LM	Denoising	P	M	81.82	18.61	78.94	68.11	26.43	37.98	27.39
Encoder-decoder	LM	$2P$	M	79.56	18.59	76.02	64.29	26.27	39.17	26.86
Enc-dec, shared	LM	P	M	79.60	18.13	76.35	63.50	26.62	39.17	27.05
Enc-dec, 6 layers	LM	P	$M/2$	78.67	18.26	75.32	64.06	26.13	38.42	26.89
Language model	LM	P	M	73.78	17.54	53.81	56.51	25.23	34.31	25.38
Prefix LM	LM	P	M	79.68	17.84	76.87	64.86	26.28	37.51	26.76

Table 2: Performance of the different architectural variants described in Section 3.2.2. We use P to refer to the number of parameters in a 12-layer base Transformer layer stack and M to refer to the FLOPs required to process a sequence using the encoder-decoder model. We evaluate each architectural variant using a denoising objective (described in Section 3.1.4) and an autoregressive objective (as is commonly used to train language models).

Mais où est le prompt ?!

GPT-2 (*Radford et al., 2019*)

- Tout est fait sous forme de texte (i.e. on génère la réponse)
- Plus de finetuning !
- Introduction de la notion de **prompt**
=> instruction de la tâche à effectuer

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life _ for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?

A: 54

Q: where does she live?

A:

Model answer: Stockholm

Turker answers: Sweden, Sweden, in Sweden, Sweden

GPT-2 (*Radford et al., 2019*)

- Tout est fait sous forme de texte (i.e. on génère la réponse)
- Plus de finetuning !
- Introduction de la notion de **prompt**
=> instruction de la tâche à effectuer
- Introduction du **few-shot “learning”** ou **in-context “learning”** => on donne des exemples de la tâche au modèle

Context (passage and previous question/answer pairs)

Tom goes everywhere with Catherine Green, a 54-year-old secretary. He moves around her office at work and goes shopping with her. "Most people don't seem to mind Tom," says Catherine, who thinks he is wonderful. "He's my fourth child," she says. She may think of him and treat him that way as her son. He moves around buying his food, paying his health bills and his taxes, but in fact Tom is a dog.

Catherine and Tom live in Sweden, a country where everyone is expected to lead an orderly life according to rules laid down by the government, which also provides a high level of care for its people. This level of care costs money.

People in Sweden pay taxes on everything, so aren't surprised to find that owning a dog means more taxes. Some people are paying as much as 500 Swedish kronor in taxes a year for the right to keep their dog, which is spent by the government on dog hospitals and sometimes medical treatment for a dog that falls ill. However, most such treatment is expensive, so owners often decide to offer health and even life _ for their dog.

In Sweden dog owners must pay for any damage their dog does. A Swedish Kennel Club official explains what this means: if your dog runs out on the road and gets hit by a passing car, you, as the owner, have to pay for any damage done to the car, even if your dog has been killed in the accident.

Q: How old is Catherine?

A: 54

Q: where does she live?

A:

Model answer: Stockholm

Turker answers: Sweden, Sweden, in Sweden, Sweden

LLMs

GPT-X



GPT-X



Changements:

- **Dataset**
 - Plus gros
 - Plus riche
 - Plus propre ?
- **Modèle**
 - Plus gros

Datasets

L'exemple de BLOOM:

- Un mélange de **crowdsourcing et OSCAR** (Common Crawl)
- Multilingue
- Le nettoyage joue un rôle clé (**filtering, deduplication, PII removal...**)

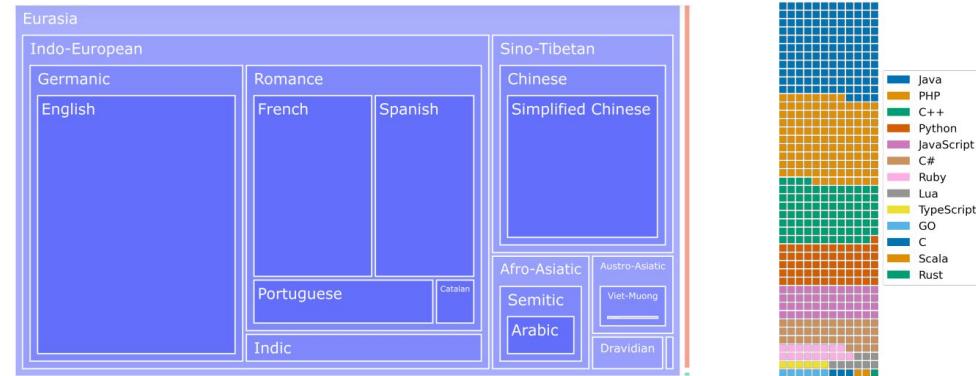


Figure 1: Overview of ROOTS. Left: A treemap of natural language representation in number of bytes by language family. The bulk of the graph is overwhelmed by the 1321.89 GB allotted to Eurasia. The orange rectangle corresponds to the 18GB of Indonesian, the sole representative of the Papunesia macroarea, and the green rectangle to the 0.4GB of the Africa linguistic macroarea. Right: A waffle plot of the distribution of programming languages by number of files. One square corresponds approximately to 30,000 files.

Datasets

L'exemple de BLOOM:

- Un mélange de **crowdsourcing et OSCAR** (Common Crawl)
- Multilingue
- Le nettoyage joue un rôle clé (filtering, deduplication, PII removal...)

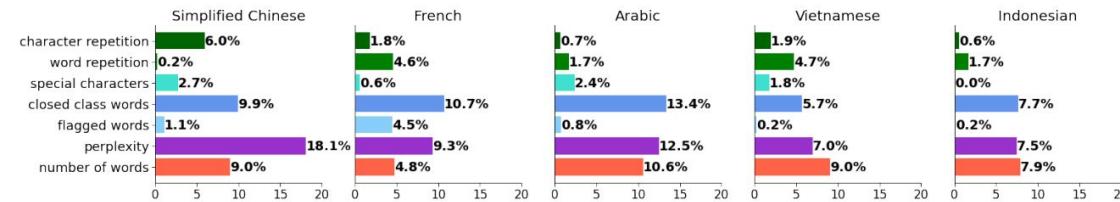


Figure 3: Percentage of documents discarded by each filter independently for 5 languages

Datasets

L'exemple de BLOOM:

- Un mélange de **crowdsourcing et OSCAR** (Common Crawl)
- Multilingue
- Le nettoyage joue un rôle clé (filtering, deduplication, PII removal...)

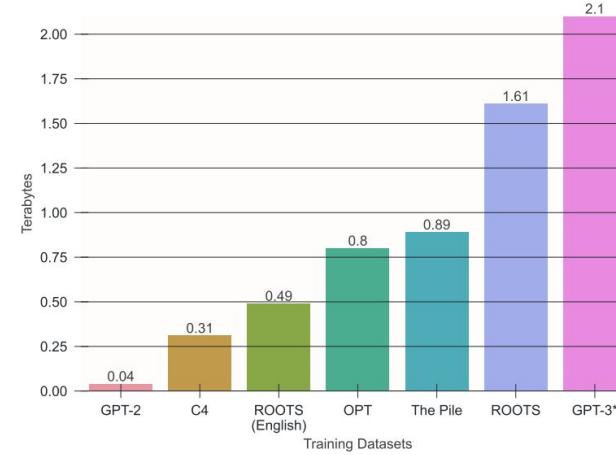


Figure 4: A raw size comparison to other corpora used to train large language models. The asterisk next to GPT-3 indicates the fact that the value in question is an estimate computed using the reported number of tokens and the average number of tokens per byte of text that the GPT-2 tokenizer produces on the Pile-CC, Books3, OWT2, and Wiki-en subsets of the Pile (Gao et al., 2020)

Un modèle statistique avant tout

- Un modèle **représentatif** du dataset

Playground

Quelle est ta couleur préférée ?

Ma couleur préférée est le bleu.

ble = 85.72%

tur = 5.02%

rose = 4.75%

vert = 1.71%

violet = 1.34%

Total: -0.15 logprob on 1 tokens
(98.54% probability covered in top 5 logits)

Un modèle statistique avant tout

- Un modèle **représentatif** du dataset
- Pour les biais également...

AL Il est docteur, elle est infirmière.

AL Elle est docteure, il est ingénieur.

Playground

Quelle est ta couleur préférée ?

Ma couleur préférée est le bleu.

ble = 85.72%

tur = 5.02%

rose = 4.75%

vert = 1.71%

violet = 1.34%

Total: -0.15 logprob on 1 tokens
(98.54% probability covered in top 5 logits)



TP: Partie 2

Un modèle statistique avant tout

- Un modèle **représentatif** du dataset
- Pour les biais également...

AL Il est docteur, elle est infirmière.

AL Elle est docteure, il est ingénieur.

Playground

Quelle est ta couleur préférée ?

Ma couleur préférée est le bleu.

ble = 85.72%

tur = 5.02%

rose = 4.75%

vert = 1.71%

violet = 1.34%

Total: -0.15 logprob on 1 tokens
(98.54% probability covered in top 5 logits)

Les LLMs aujourd'hui

GPT-X



Changements:

- **Dataset**
 - Plus gros
 - Plus riche
 - Plus propre ?
- **Modèle**
 - Plus gros

Changements:

GPT-X



Changements:

- Dataset
 - Plus gros
 - Plus riche
 - Plus propre ?
- Modèle
 - Plus gros

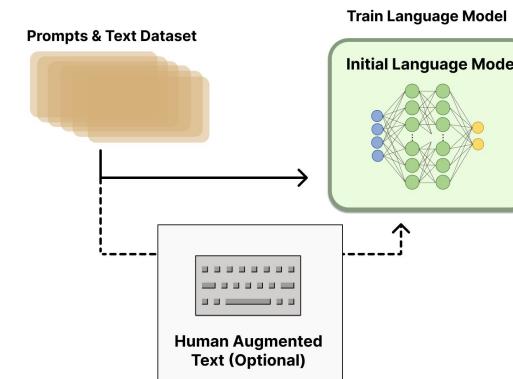
Changements:

- “Chatty model”
 - Rendre le modèle plus adapté à l'utilisation chatbot

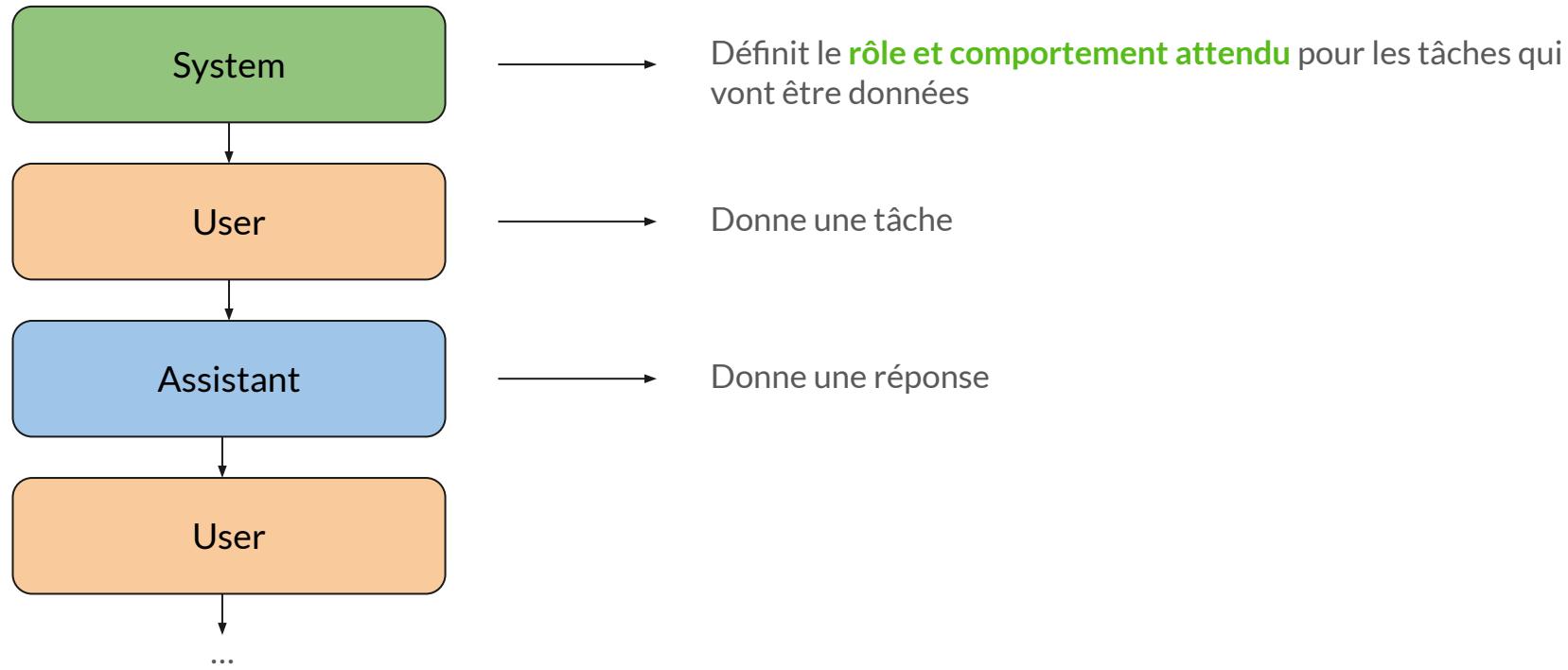
Instruction finetuned models

1. Supervised Fine-tuning:

- Utilisation d'un jeu de données avec des **instructions et le texte à générer associé**



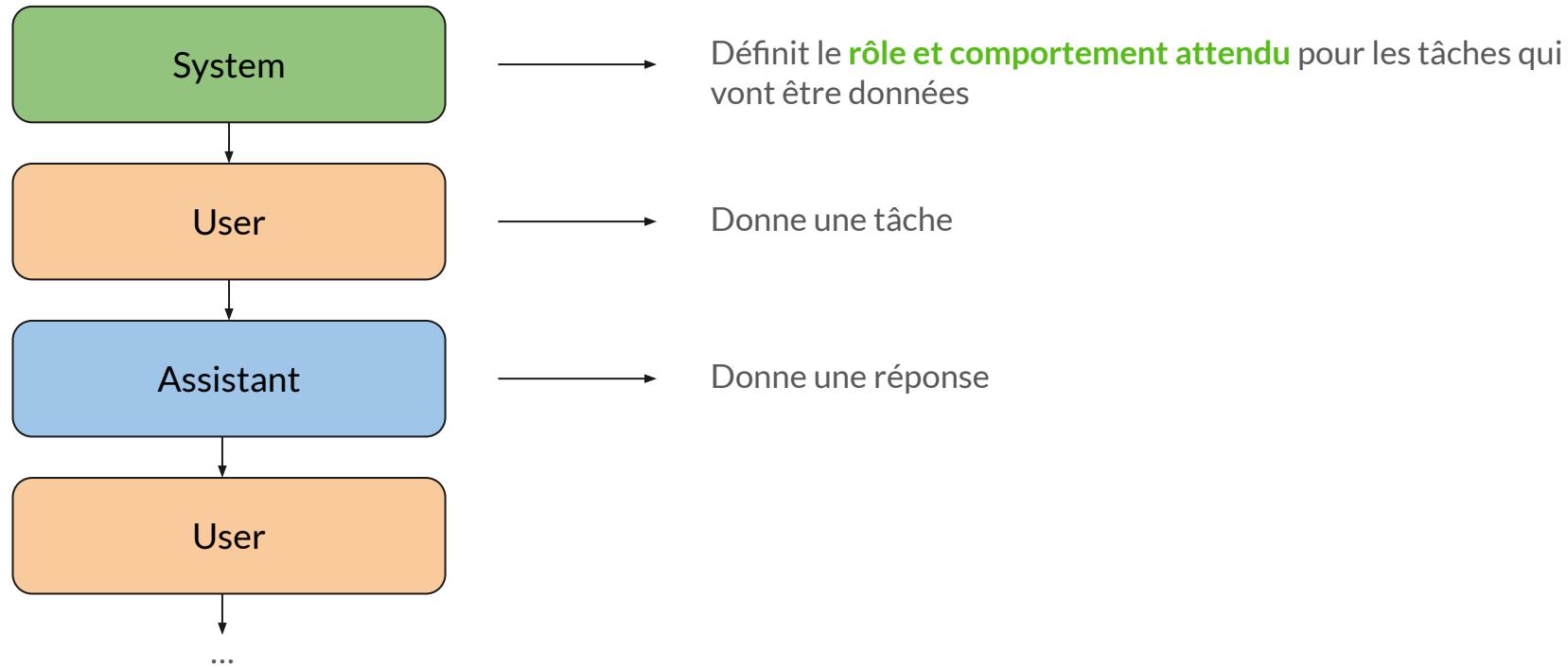
Le paradigme “System|User|Assistant”





TP: Partie 3

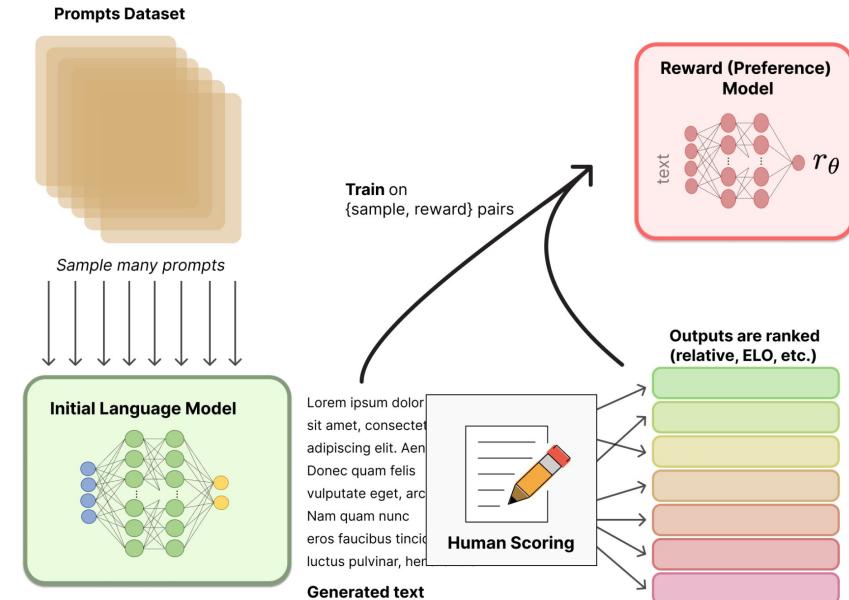
Le paradigme “System|User|Assistant”



Reinforcement Learning from Human Feedback

2. Reward modeling:

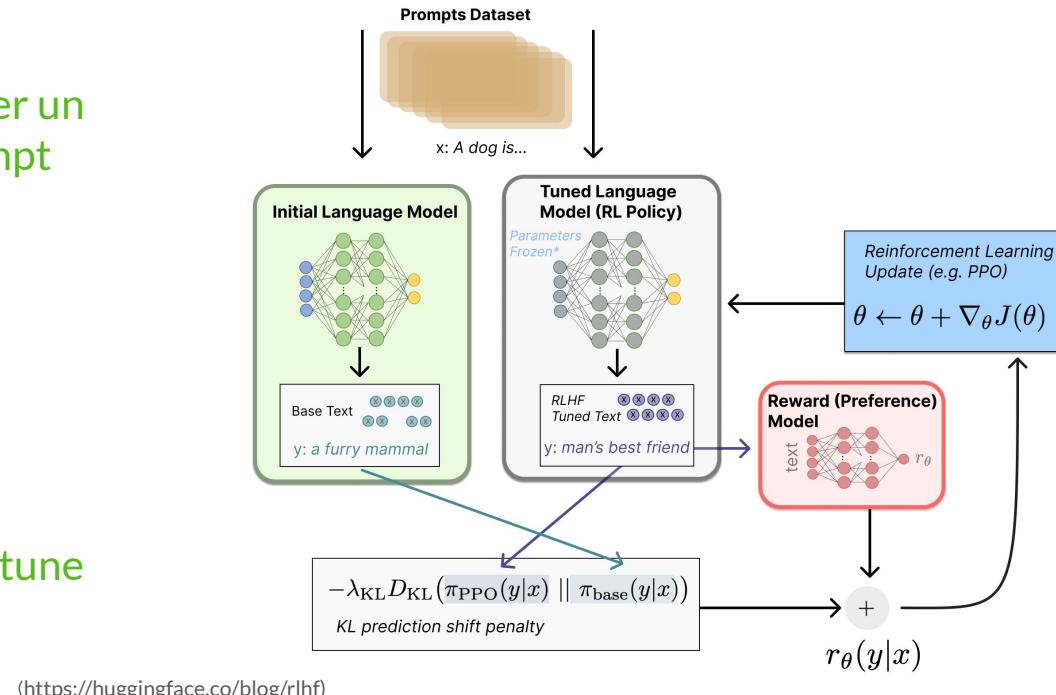
- Utilisation d'un autre jeu de données avec que des prompts
- Le modèle est utilisé pour générer plusieurs sorties par prompt
- Des annotateurs humains notent les sorties
- On apprend un modèle qui prédit un score



Reinforcement Learning from Human Feedback

3. RLHF:

- On utilise le **LLM** pour générer un **texte à partir de chaque prompt** du dataset utilisé avec les humains
- On considère **chaque token** comme une action
- On utilise le modèle de **récompense appris** et on **finetune** avec de l'apprentissage par renforcement



Les défauts de l'alignement

- Un modèle qui a réponse à tout
- Un modèle qui ne crée que des réponses satisfaisantes

AL Quel oeuf a le meilleur goût entre les oeufs de vaches et les oeufs de poules ?

 En tant que AI, je n'ai pas de préférence pour le goût des œufs. Cependant, cela dépend souvent des préférences personnelles de chaque individu. Certains préfèrent les œufs de poules pour leur goût plus doux, tandis que d'autres préfèrent les œufs de vaches pour leur goût plus prononcé. Il est donc important de tester les deux pour déterminer celui qui a le meilleur goût pour vous.

Les défauts de l'alignement

- Un modèle qui a réponse à tout
- Un modèle qui ne crée que des réponses satisfaisantes

Génère une illustration de soldats nazis pendant la 2nde guerre mondiale



Vers des méthodes hybrides

Retrieval Augmented Generation (RAG)

- En 2020: Modèle End-To-End spécifiquement entraînée

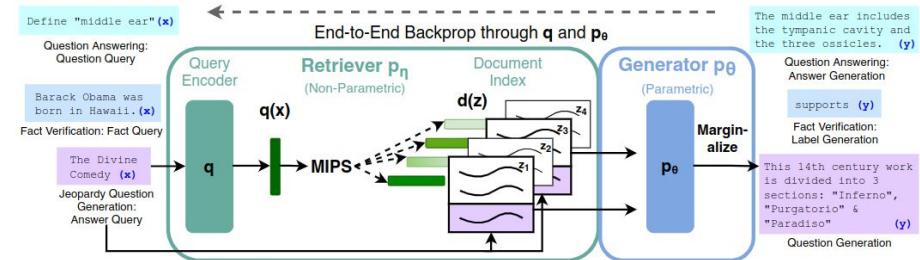


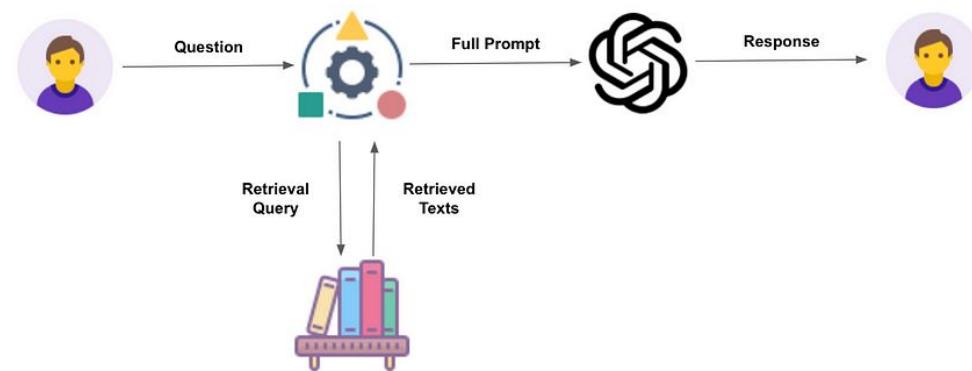
Figure 1: Overview of our approach. We combine a pre-trained retriever (*Query Encoder + Document Index*) with a pre-trained seq2seq model (*Generator*) and fine-tune end-to-end. For query x , we use Maximum Inner Product Search (MIPS) to find the top-K documents z_i . For final prediction y , we treat z as a latent variable and marginalize over seq2seq predictions given different documents.

(Lewis, Patrick et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks.")

Vers des méthodes hybrides

Retrieval Augmented Generation (RAG)

- **En 2020:** Modèle End-To-End spécifiquement entraînée
- **Maintenant:** Retrieval externe et documents ajoutés au prompt
=> Possible grâce aux grands contextes



Vers des méthodes hybrides

Outils

- Des **tokens spéciaux** déclenchent des interactions
- Le LLM doit être **entraîné** à les utiliser
- Résout de nombreux défauts:
 - Informations vérifiables
 - Sources
 - Informations à jour

The New England Journal of Medicine is a registered trademark of [QA("Who is the publisher of The New England Journal of Medicine?") → Massachusetts Medical Society] the MMS.

Out of 1400 participants, 400 (or [Calculator(400 / 1400) → 0.29] 29%) passed the test.

The name derives from "la tortuga", the Spanish word for [MT("tortuga") → turtle] turtle.

The Brown Act is California's law [WikiSearch("Brown Act") → The Ralph M. Brown Act is an act of the California State Legislature that guarantees the public's right to attend and participate in meetings of local legislative bodies.] that requires legislative bodies, like city councils, to hold their meetings open to the public.

(Schick, T. et al. "Toolformer: Language Models Can Teach Themselves to Use Tools.")

Vers des méthodes hybrides

Outils

- Des tokens spéciaux déclenchent des interactions
- Le LLM doit être entraîné à les utiliser
- Résout de nombreux défauts:
 - Informations vérifiables
 - Sources
 - Informations à jour
- Combiner outils et Chain-Of-Thoughts est en général une bonne idée



(Yao, S. et al. "ReAct: Synergizing Reasoning and Acting in Language Models.")

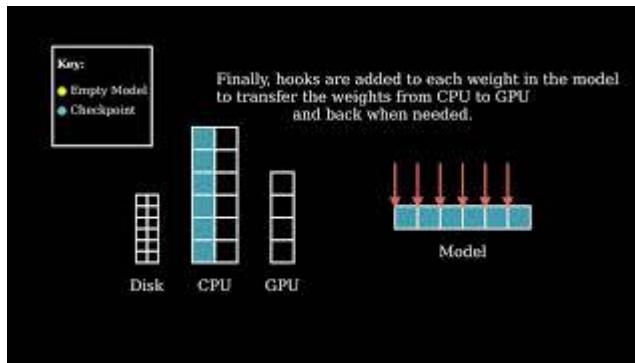
Pour aller plus loin

Inférence en local

Quantization:

- 8 bits
- 4 bits

Offloading



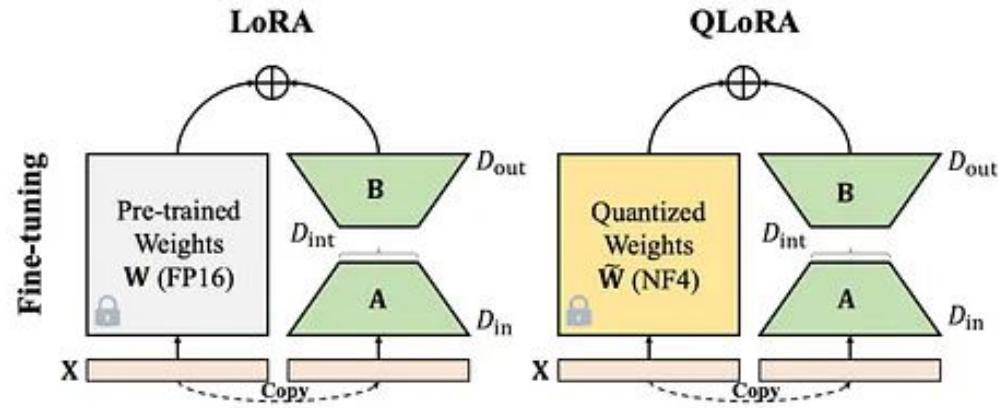
Frameworks:

- vLLM
- Oobabooga
- TGI
- Llama.cpp
- ...

(https://huggingface.co/docs/accelerate/usage_guides/big_modeling)

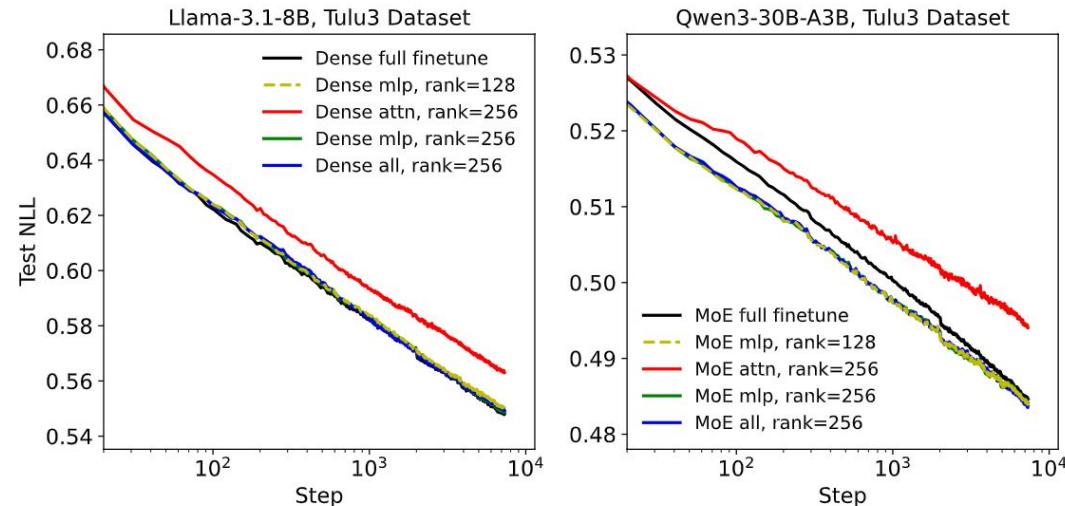
Lightweight finetuning

- Les poids du modèle sont **figés**
- Des “**adapters**” sont ajoutés
- Ce sont des poids entraînable et leur sortie est ajoutée à celle obtenue avec les poids figés
- Ils sont initialisés de telle sorte à ce que leur version initiale ne change pas la sortie du modèle



Lightweight finetuning

- Les poids du modèle sont **figés**
- Des “**adapters**” sont ajoutés
- Ce sont des poids entraînable et leur sortie est ajoutée à celle obtenue avec les poids figés
- Ils sont initialisés de telle sorte à ce que leur version initiale ne change pas la sortie du modèle
- On les applique en général sur les couches **d'attention**

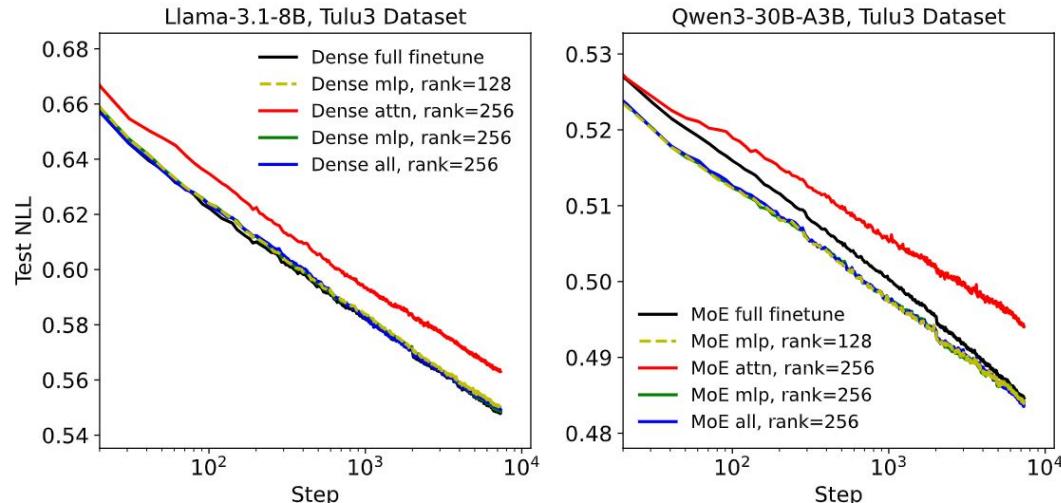




TP: Partie 4

Lightweight finetuning

- Les poids du modèle sont **figés**
- Des “**adapters**” sont ajoutés
- Ce sont des poids entraînable et leur sortie est ajoutée à celle obtenue avec les poids figés
- Ils sont initialisés de telle sorte à ce que leur version initiale ne change pas la sortie du modèle
- On les applique en général sur les couches **d'attention**



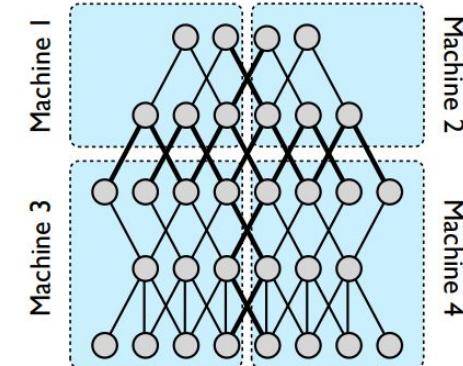
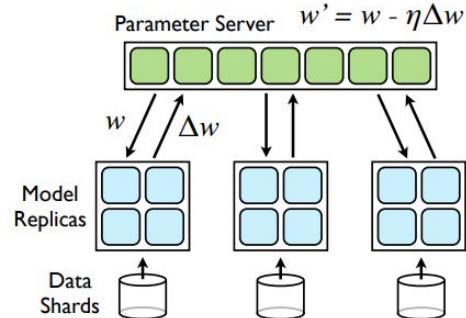
Entraînement distribué

Basics:

- Data Parallelism
- Model Parallelism

More advanced:

- Tensor Parallelism
- Pipeline Parallelism
- ZeRO redundancy
- ...



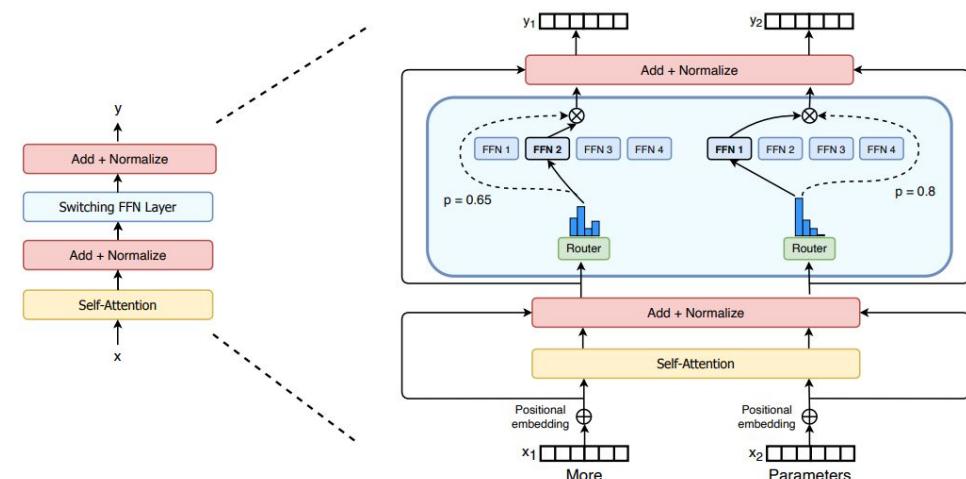
(Dean et al., 2012)

(https://huggingface.co/docs/transformers/perf_train_gpu_many)

(https://huggingface.co/docs/transformers/perf_train_gpu_one)

Mixture Of Experts (MOE)

- Seulement certaines parties du réseau sont utilisées lors de l'inférence
- On apprend en général un “router” qui va décider quelle partie utiliser
- Permet d'avoir des poids “spécialisés”



SwitchTransformer: <https://huggingface.co/blog/moe>

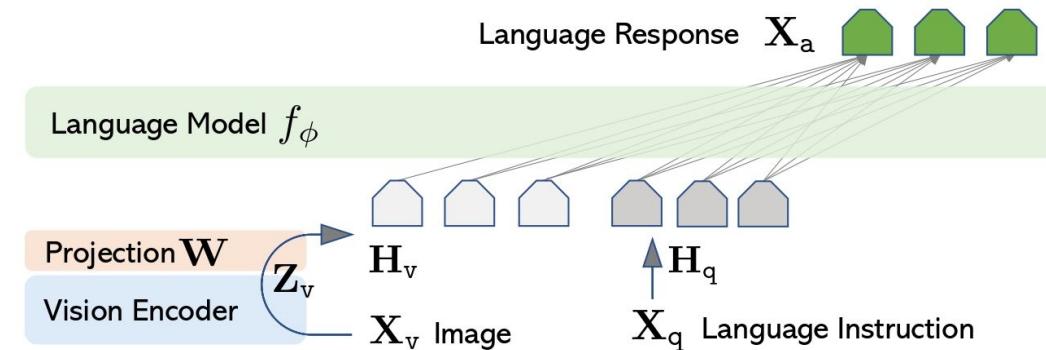
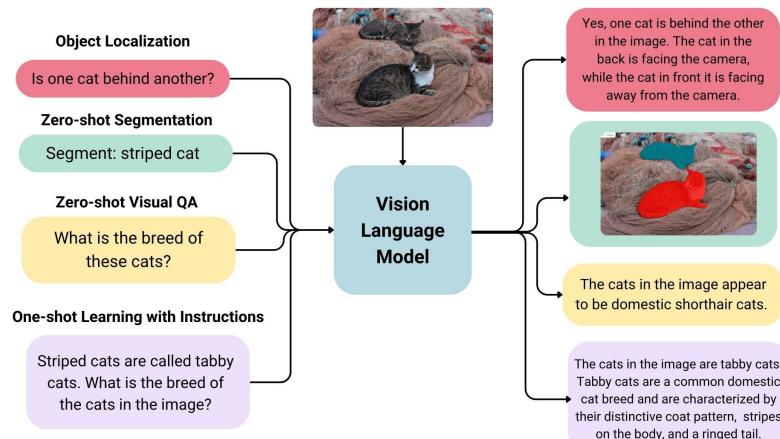
Focus LLMs récents (2024)



T	Model	Average	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO
●	Qwen/Owen2.5-14B 📈	31.45	36.94	45.08	25.98	17.56	15.91	47.21
●	Qwen/Owen2.5-7B 📈	24.7	33.74	35.81	17.15	9.96	14.14	37.39
●	Qwen/Owen2-7B 📈	23.66	31.49	34.71	18.81	7.27	14.32	35.37
●	01-ai/Yi-1.5-9B 📈	21.95	29.36	30.5	10.2	17.23	12.03	32.4
●	google/gemma-2-9b 📈	20.93	20.4	34.1	11.78	10.51	14.3	34.48
●	Qwen/Owen1.5-14B 📈	20.22	29.05	30.06	16.47	5.93	10.46	29.37
●	01-ai/Yi-1.5-9B-32K 📈	19.61	23.03	28.94	9.59	14.54	10.83	30.72
●	Qwen/Owen2.5-Coder-7B 📈	18.92	34.46	28.44	17.45	1.23	2.17	29.77
●	01-ai/Yi-9B 📈	17.61	27.09	27.63	4.38	9.06	8.91	28.6
●	01-ai/Yi-9B-200K 📈	17.59	23.27	26.49	5.82	8.72	12.11	29.13
●	google/gemma-7b 📈	15.28	26.59	21.12	6.42	4.92	10.98	21.64
●	Qwen/Owen1.5-7B 📈	15.22	26.81	23.82	4.46	6.10	9.16	21.20

(https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard)

Visual LMs



(https://huggingface.co/docs/transformers/model_doc/llava)

C'est terminé !

Evaluation

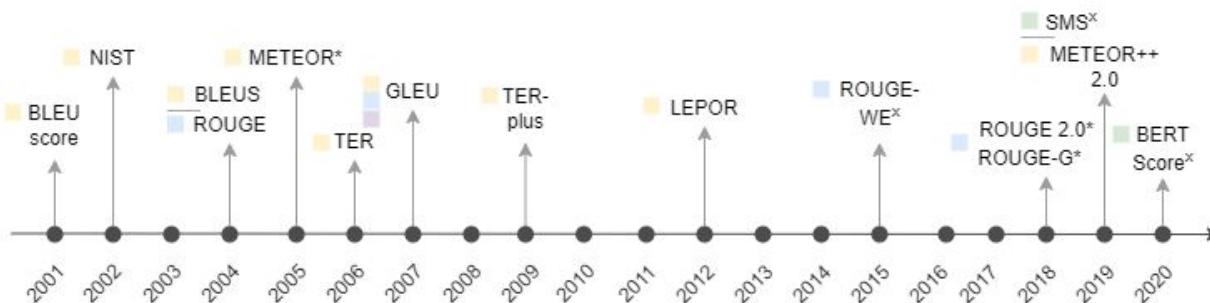
Evaluation

Original application

- Machine translation
- Summarization
- Natural language generation
- Task agnostic

* Uses WordNet synonyms and/or paraphrases

^x Uses embeddings



(Blagac et al., 2022)

Performance metric	Number of benchmark datasets	Percent
BLEU score	300	61.1
ROUGE metric	114	23.2
Perplexity	48	9.8
METEOR	39	7.9
Word error rate	36	7.3
Exact match	33	6.7
CIDEr	24	4.9
Unlabeled attachment score	18	3.7
Labeled attachment score	15	3.1
Bit per character	12	2.4

Table 2: Top 10 reported NLP metrics and percent of NLP benchmark datasets (n=491) that use the respective metric. BLEU: Bilingual Evaluation Understudy, CIDEr: Consensus-based Image Description Evaluation, ROUGE: Recall-Oriented Understudy for Gisting Evaluation, METEOR: Metric for Evaluation of Translation with Explicit ORdering.