CentraleSupélec

BLOOM

DEBRAY Clarisse
GLERANT Pierre
FERNANDES DE ALMEIDA Maxsuel
LOFTI Aymane
VOGELS Arthur

Introduction

Introduction

Premiers modèles NLP

Prédit des n-grammes à partir de leur fréquence dans un corpus. Cependant croissance exponentielle des ressources et difficultés sur les séquences rares.

Neural Language Models

Utilisation d'un NN pour estimer la proba du token suivant.

Transfer learning

Entraîne les NN sur des tâches plus riches et réutilise des paramètres pré-appris.

Few and Zero Shot learning

On prend le modèle tel quel en ne lui faisant rien apprendre (ou peu)

Ces modèles sont conçus par et pour des industriels à cause de leurs besoins de ressources. Ils peuvent donc avoir des biais.

Introduction

176 Milliards de paramètres

(dans la moyenne des LLM, dans les mêmes chiffres que GPT)

Open Source

Accent sur la collaboration et l'inclusivité

- **46 langues naturelles dont des langues sous représentées** *(ex : Yoruba)*
- **13 langages de programmation**
- **Initiative de la communauté NLP française**
- **plus de 1 200 chercheurs venant de 38 pays** *(et donc pas une seule entreprise privée)*
- **Pas que des métiers scientifiques proches du machine learning** *ex: droit, anthropologie sociale, philosophie.*

BigScience n'est pas une entreprise mais une initiative de recherche collaborative

Introduction

Les participants sont encouragés à rejoindre plusieurs groupes de travail

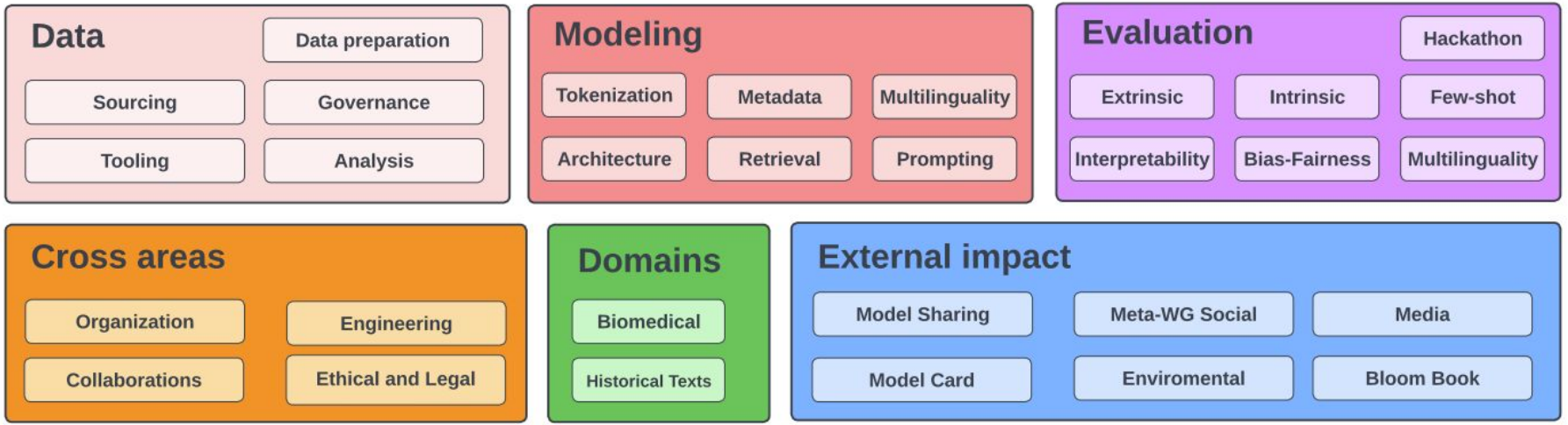
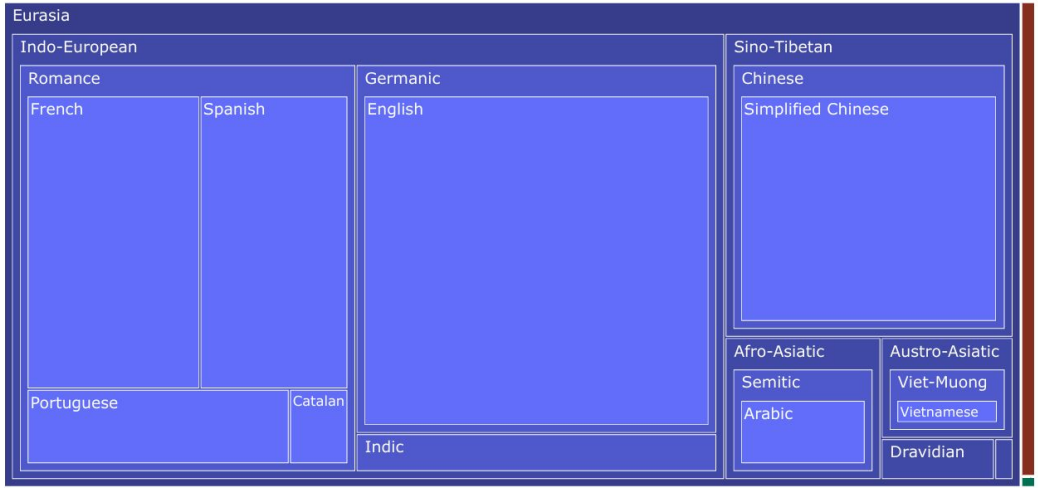


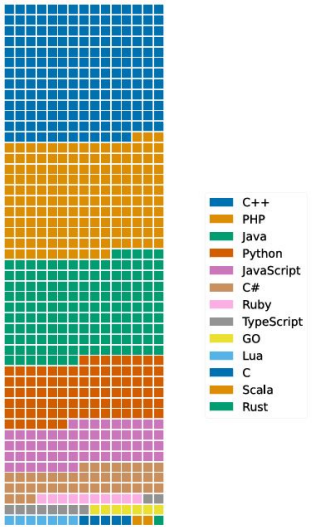
Figure 1: Organization of BigScience working groups.

TRAINING SET

Base principale du corpus d'entraînement : ROOTS

- 1,61 To de texte provenant de 498 ensembles
- 46 langages naturels
- 13 langages de programmation





Aperçu graphique du corpus ROOTS.

PREPROCESSING

3 étapes principales :

1. Acquisition des données
   - Téléchargement de jeux de données NLP variés
   - Extraction depuis des PDF (ex. archives scientifiques françaises)
   - Récupération depuis 192 sites du catalogue et 456 autres sites

2. Filtrage “qualitatif”
   - Critère principal: Texte "écrit par des humains pour des humains"

3. Déduplication et anonymisation
   - Suppression des doublons (deux étapes).
   - Retrait des informations personnelles (regex).

Architecture du réseau

Décodeur Causal:

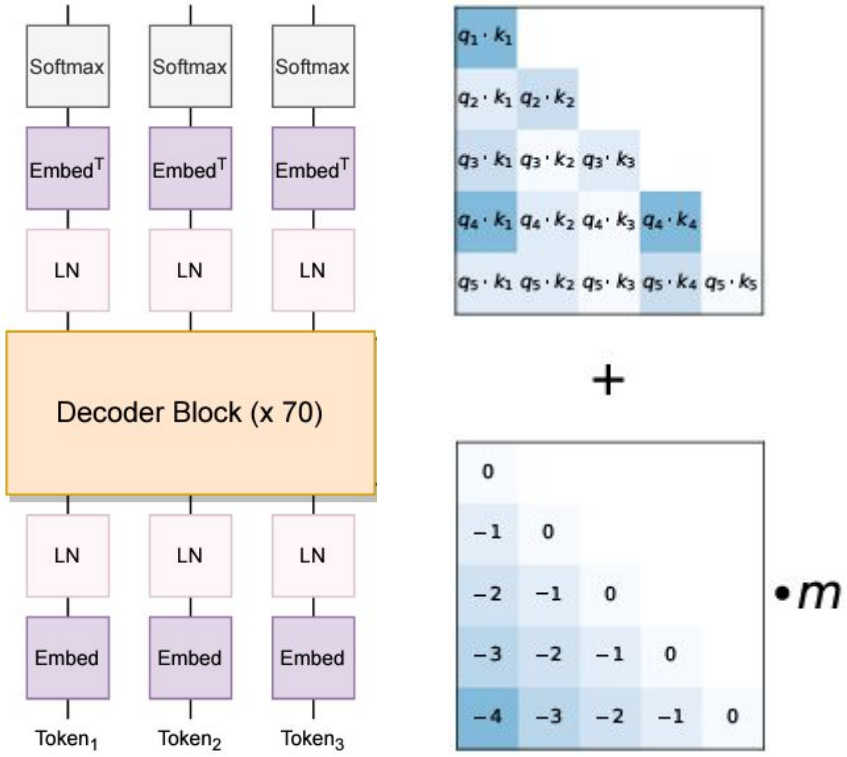
- Architecture privilégiée pour les gros modèles
- Meilleure généralisation après entraînement

Etude d'ablation:

- Sur la fonction objectif et l'architecture
- Effectuée sur des modèles plus petits
- Évaluée sur les capacités zéro-shot

Deux différences avec le Transformer:

- Ajout d'une LN pour plus de stabilité
- ALiBi Positional Embeddings



Tokenizer

Entraînement d'un tokenizer:

- Lié à l'aspect multilingue
- Utilise l'algorithme Byte-Level BPE
- Entraînement sur une partie du dataset ROOTS

Evaluation:

- Comparer la fertilité avec des tokenizer multilingue

Tokenizer	fr	en	es	zh	hi	ar
Monolingual	1.30	1.15	1.12	1.50	1.07	1.16
BLOOM	1.17 (-11%)	1.15 (+0%)	1.16 (+3%)	1.58 (+5%)	1.18 (+9%)	1.34 (+13%)

Training

Aperçu des hyperparamètres :

- Tailles des modèles : De 560M à 176B paramètres.
- Taille de lot : De 256 à 2048.
- Taux d'apprentissage : Ajusté selon la taille du modèle (ex. : 6e-5 pour 176B).
- Nombre de tokens : Entraîné sur 341 milliards de tokens avec un plan de décroissance en cosinus.

Points clés du fine-tuning :

- Maintien de l'architecture des modèles pré-entraînés.
- Taux d'apprentissage doublé par rapport au minimum utilisé pour le pré-entraînement.
- Performances stabilisées après 1 à 6 milliards de tokens.

Empreinte carbone :

- 81 t CO_2eq : 14 % fabrication, 30 % entraînement, 55 % veille.
- BLOOM : 25 t CO_2eq, bien moins que GPT-3 (502 t) et OPT (70 t), grâce à l'énergie nucléaire française.

Publication

Accessibilité : Accompagné d'une fiche technique détaillant usages, limitations, et spécifications.

Licence RAIL : Responsable et gratuite, avec restrictions pour prévenir les usages nuisibles.

Collaboratif : Documentation co-écrite par plusieurs groupes.

Déploiement : Instance GCP à 16 GPUs émettant 20 kg CO_2eq/jour.

Évaluation

Axé sur les contextes *zero-shot* et *few-shot* :

- Mise en avant d'une utilisation réaliste et pratique des modèles

Plusieurs modèles de référence :

- Ex : mGPT, T0, OPT, XGLM, M2M

Plusieurs tâches prises en considération :

- SuperGLUE (ensemble de tâches pour évaluer la compréhension avancée du langage naturel)
- Machine translation
- Summarization
- Code generation

Performances sur SuperGLUE

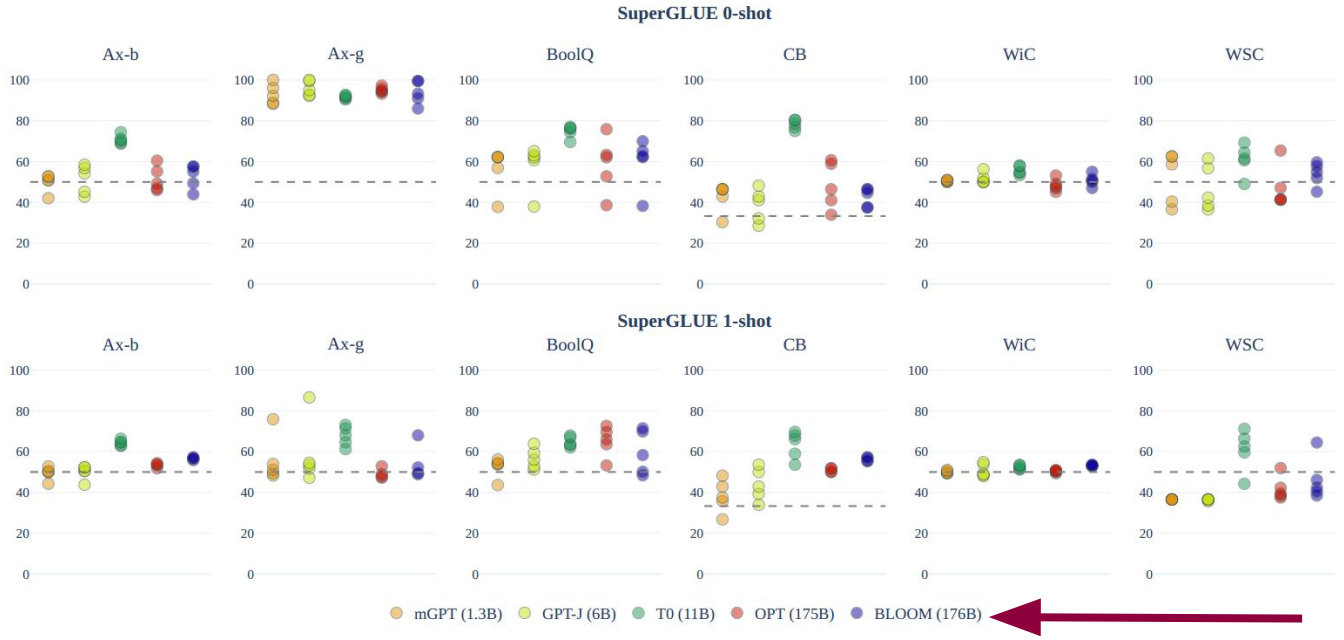


Figure 7: Performance of various LLMs on subset of tasks from SuperGLUE benchmark in zero- and one-shot prompt-based setting.

Performances en machine translation

Performance mesurée en utilisant la métrique ***spBLEU*** et le jeu de données ***Flores-101***

Langues à faibles ressources

Src↓	Trg→	en	bn	hi	sw	yo
en	BLOOM	–	24.6	27.2	20.5	2.6
	M2M	–	23.0	28.1	26.9	2.2
bn	BLOOM	29.9	–	16.3	–	–
	M2M	22.9	–	21.8	–	–
hi	BLOOM	35.1	23.8	–	–	–
	M2M	27.9	21.8	–	–	–
sw	BLOOM	37.4	–	–	–	1.3
	M2M	30.4	–	–	–	1.3
yo	BLOOM	4.1	–	–	0.9	–
	M2M	4.2	–	–	1.9	–

Langues à hautes ressources

Src ↓	Trg →	ar	en	es	fr	zh
ar	BLOOM	–	40.3	23.3	33.1	17.7
	M2M	–	25.5	16.7	25.7	13.1
	AlexaTM	–	41.8	23.2	35.5	–
en	BLOOM	28.2	–	29.4	45.0	26.7
	M2M	17.9	–	25.6	42.0	19.3
	AlexaTM	32.0	–	31.0	50.7	–
es	BLOOM	18.8	32.7	–	24.8	20.9
	M2M	12.1	25.1	–	29.3	14.9
	AlexaTM	20.8	34.6	–	33.4	–
fr	BLOOM	23.4	45.6	27.5	–	23.2
	M2M	15.4	37.2	25.6	–	17.6
	AlexaTM	24.7	47.1	26.3	–	–
zh	BLOOM	15.0	30.5	20.5	26.0	–
	M2M	11.55	20.9	16.9	24.3	–
	AlexaTM	–	–	–	–	–

Performances en summarization

Performance mesurée en utilisant la métrique *ROUGE-2* et le jeu de données *WikiLingua*

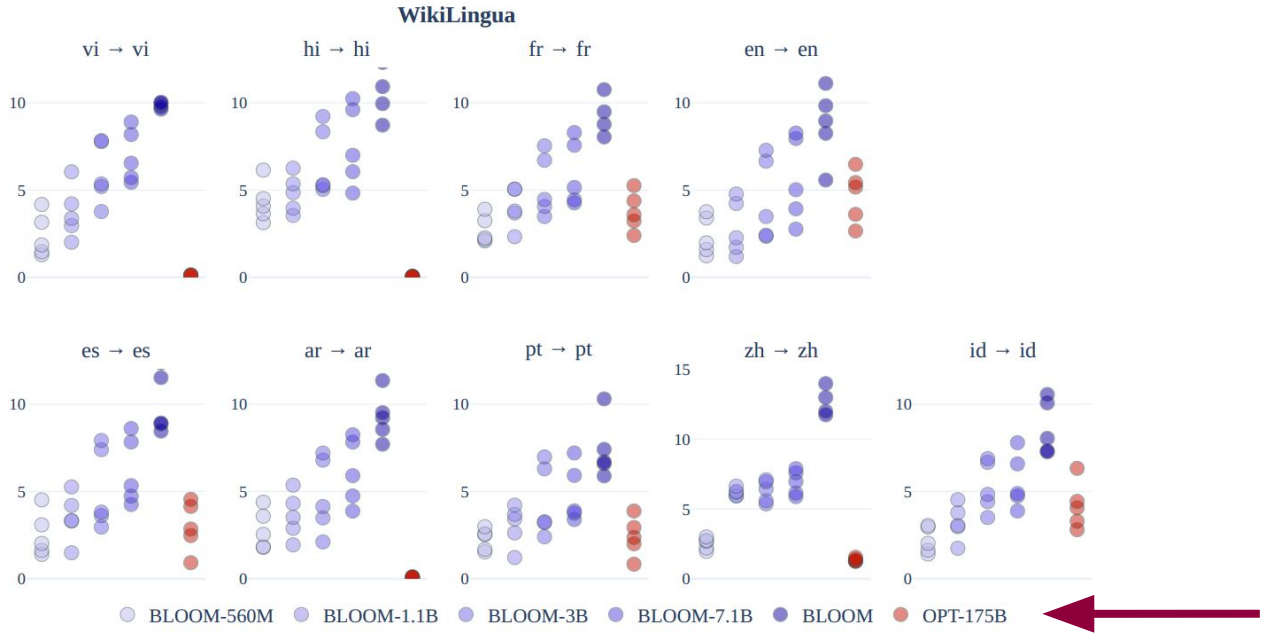


Figure 9: WikiLingua One-shot Results. Each plot represents a different language with per-prompt ROUGE-2 F-measure scores.

Conclusion