

# Les modèles Bloom pour le traitement automatique de la langue française



INSTITUT DU DÉVELOPPEMENT ET DES RESSOURCES EN INFORMATIQUE SCIENTIFIQUE

Rachel Bawden<sup>1</sup> Hatim Bourfoune<sup>2</sup> Bertrand Cabot<sup>2</sup> Nathan Cassereau<sup>2</sup>

Pierre Cornette<sup>2</sup> Marco Naguib<sup>3</sup> Aurélie Névéol<sup>3</sup> François Yvon<sup>4</sup>

(1) Inria Paris, France

(2) IDRIS, CNRS, Orsay, France

(3) LISN, Université Paris-Saclay et CNRS, Orsay, France

(4) ISIR, CNRS et Sorbonne-Université, Paris, France



## BigScience

### Résultats de BigScience

- Bloom, une famille de modèles multilingues.
- Évaluation multitâches (prompts anglais, tâches multilingues)
- Multi-évaluations ultérieures

Que valent les modèles Bloom pour des évaluations en français ?

## La famille Bloom

6 modèles de tailles croissantes

	Hyperparamètres					
Dimension	1024	1536	2048	2560	4096	14336
# Couches	24	24	24	30	30	70
# Têtes	16	16	16	32	32	112
# Param. (m)	559	1065	1722	3003	7069	176247

Entraînés sur ROOTS (46 langues, 350m tokens)

Bloomz dérive de Bloom (176g) par affinage multitâche

## Tâches considérées

Modélisation

Flores  
wikiFR

Classification

Analyse d'opinions  
Implication textuelle

Étiquetage séquences

WikiNER

QuaeroFrenchMed

Question-réponse  
PIAF

Génération

Traduction automatique  
Résumé de textes

Mesure des biais

CrowS-pairs

Shades

## Contributions

### Code

- Configs standard pour Bloom sur JZ
- Améliorations de LMeval (tâches et métriques)

### Ressources

- Invites en français pour 9 tâches
- Corpus pour l'évaluation de GML
- Jeux de données pour NER / BioMed

### Résultats d'évaluation

- Étalonnage de Bloom
- Comparaison de modèles
- Comparaison d'invites
- Comparaisons entre langues

## Outillage: Promptsources + LMeval

- PromptSource: spécification des invites
- LMharness: génération des invites, génération des réponses, évaluation

## Modélisation des langues [flores]

Langue (longueur)	Perplexité (Flores-101)				Bits/Octet (Flores-101)			
	deu (45129)	eng (26404)	fra (32008)	spa (32072)	deu (156358)	eng (132096)	fra (163927)	spa (159899)
Bloom-560m	144,30	44,48	25,68	31,54	2,07	1,09	0,91	1,00
Bloom-1b1	93,11	36,68	22,01	27,22	1,89	1,04	0,87	0,96
Bloom-1b7	65,94	32,49	19,66	24,42	1,74	1,00	0,84	0,92
Bloom-3b	51,80	29,58	18,08	22,75	1,64	0,98	0,82	0,90
Bloom-7b1	35,99	25,94	16,44	20,74	1,49	0,94	0,79	0,88
Bloom	18,33	20,17	13,55	17,26	1,21	0,87	0,73	0,82

- Anglais et français sont comparables, meilleurs que l'espagnol et à fortiori que l'allemand
- De Bloom-560 à Bloom division par 2 de la perplexité (confirmé sur un corpus français)

## Implication Textuelle [xnli]

Données	Bloom				Bloomz			
	en	en	fr	fr	en	en	fr	fr
Invites								
based_on_the_previous_passage	0,43	0,40	0,42	0,42	0,53	0,51	0,55	0,55
can_we_infer	0,40	0,37	0,42	0,42	0,45	0,45	0,50	0,50
does_it_follow_that	0,37	0,36	0,38	0,38	0,47	0,48	0,54	0,54
take_the_following_as_truth	0,40	0,36	0,40	0,40	0,47	0,44	0,38	0,38
moyenne	0,40	0,37	0,41	0,41	0,48	0,47	0,49	0,49

- La cohérence entre langues (invites donnée) améliore les résultats
- Résultats comparables entre en/en et fr/fr
- Bloomz >> Bloom, mais plus de variation entre invites

## Réponses aux Questions [PIAF]

invite	Langue	F1 agrégés sur tous les modèles Bloom + Bloomz					
		zéro-exemple			mono-exemple		
		moy.	min	max	moy.	min	max
after_reading	en	0,19	0,04	0,73	0,38	0,14	0,71
	fr	0,18	0,04	0,72	0,37	0,15	0,71
given_above_context	en	0,16	0,02	0,72	0,34	0,09	0,71
	fr	0,16	0,02	0,71	0,33	0,08	0,70
given_passage_answer	en	0,17	0,02	0,74	0,37	0,15	0,71
	fr	0,17	0,02	0,73	0,36	0,15	0,70
moyenne	en	0,18	0,02	0,74	0,36	0,06	0,71
	fra	0,18	0,02	0,73	0,36	0,06	0,71

- Bloom zéro-exemple, mono-exemple très loin de l'état de l'art
- Peu de différences en-fr dans la formulation des invites
- Bloomz (zéro-exemple !) >> Bloom, très proche de (Cattan et al, 2021 - 0,73 F1)

## Traduction automatique [WMT 14]

Invite	BLEU (zéro-exemple)			BLEU (mono-exemple)		
	moy.	min	max	moy.	min	max
en	5,3 → 9,2	0,4 → 0,6	15,4 → 32,2	19,1 → 21,0	3,2 → 3,4	36,4 → 37,1
fr	11,8 → 19,5	0,3 → 0,3	27,3 → 37,8	21,0 → 22,2	1,4 → 1,4	37,4 → 37,8
en-fr (avant / après post-traitement)						
en	10,4 → 18,8	2,1 → 2,3	16,8 → 37,2	23,8 → 25,2	7,6 → 8,2	36,6 → 38,2
fr	9,1 → 15,2	0,1 → 0,1	22,2 → 36,6	20,3 → 21,8	1,2 → 1,3	37,6 → 38,0
fr-en (avant / après post-traitement)						

- Variance entre invites, problème de génération zéro-exemple confirmés pour les invites en français
- Bénéfice de la cohérence invite- langue cible
- Aussi: Bloomz ≈ Bloom pour la TA

## Remerciements

Ce projet a reçu le soutien du CNRS et a utilisé les ressources de calcul de l'IDRIS (allocation 2023-AD010614012) à travers GENCI.