Evaluer BLOOM en français

Rachel Bawden, Hatim Bourfoune, Bertrand Cabot, Nathan Cassereau, Pierre Cornette, Marco Naguib, François Yvon

Évaluer `Bloom` en français

Rachel Bawden [2] Hatim Bourfoune [1] Bertand Cabot [1] Nathan Cassereau [1]
Pierre Cornette [1] Marco Naguib [3] François Yvon [4]
[1] CNRS, IDRIS, 91 403, Orsay, France
[2] Inria, 75 013, Paris, France
[3] Université Paris-Saclay & CNRS, LISN, 91 403, Orsay, France
[4] Sorbonne-Université & CNRS, ISIR, 75 005, Paris, France

RÉSUMÉ

Le développement de très grands modèles de langue, capables de traiter de multiples tâches, implique de développer les infrastructures requises pour évaluer ces modèles sous toutes leurs facettes. De nombreux ensembles de données de référence ont ainsi été rassemblés pour l'anglais, permettant d'apprécier en détail leur capacité à traiter cette langue. Dans cet article, nous présentons nos efforts pour assembler un ensemble d'évaluation multi-tâche pour le français, qui est utilisé pour évaluer le modèle `Bloom`. Nos résultats complètent les évaluations de `Bloom` en anglais ; ils suggèrent que les performances pour le français et l'anglais sont très voisines, et encore meilleures lorsque les amorces utilisées pour l'inférence en contexte sont dans la même langue que les textes soumis à l'analyse.

ABSTRACT

BLOOM Models for French Natural Language Processing

The development of very large language models, capable of performing multipes tasks, implies to develop the necessary infrastructures to evaluate these models, ideally covering as many facets as possible. Numerous benchmarks have already been compiled for English, making it possible to precisely gauge their ability to process this language. In this paper, we present our own efforts to assemble a multi-task evaluation set for French, which is then used to evaluate models from the `Bloom` family. Our results complement the main evaluation results for `Bloom` in English ; they suggest that the performance obtained in French and English are very similar, and even better when the amorces used for contextual inference are in the same language as the texts to analyze.

MOTS-CLÉS : giga modèles de langues ; évaluations automatiques.

KEYWORDS: Large Language Models ; automatic evaluations.

1 Introduction

Le développement de giga modèles de langue (*Large Language Models*, ou GML) constitue une rupture dans l'évolution des méthodes de traitement automatique des langues. Pré-entraînés sur de gigantesques corpus de textes en s'appuyant sur des tâches simples (prédiction de mots masqués (Devlin *et al.*, 2019), prédiction du mot suivant (Radford *et al.*, 2019; Brown *et al.*, 2020), débruitage (Lewis *et al.*, 2020)) pour lesquelles des annotations « naturelles » sont disponibles, ces modèles

permettent de calculer, pour chaque unité [1] d'un texte en entrée, une représentation contextuelle numérique dense dans un espace de grande dimension (Raffel *et al.*, 2020). Sur la base de ces représentations, il est ensuite possible de construire des systèmes performants soit en affinant le modèle à l'aide de données de supervision dédiées à une tâche, soit (pour les modèles causaux), en amorçant la génération par une amorce (*prompt*) en langue naturelle. L'amorce peut contenir une *instruction* décrivant la tâche à accomplir, ainsi qu'un ou plusieurs exemples (des *démonstrations*, ou en anglais *few-shot examples*), qui constituent un contexte enrichi pour la génération. L'approche à base d'amorces présente l'avantage d'utiliser un même modèle pour de multiples tâches (McCann *et al.*, 2018; Radford *et al.*, 2019). Ces deux approches ne sont pas exclusives entre elles, puisqu'on peut affiner le modèle sur des séquences complétées par les amorces idoines [2] ; ni non plus exclusives d'autres méthodes d'optimisation qui pourraient encore améliorer les performances de ces modèles (apprentissage et optimisation des amorces, ajout de couches ou de modules d'adaptation (Houlsby *et al.*, 2019), etc.).

Le projet BigScience a produit un GML *multilingue* selon les principes de la science ouverte. Ce modèle, `Bloom`, existe en plusieurs tailles, allant de 560 millions de paramètres jusqu'à 176 milliards (BigScience *et al.*, 2022). Une dérivation du modèle par affinage multitâche et multilingue donnant lieu aux familles `mT0` et `Bloomz` est décrite dans (Muennighoff *et al.*, 2022). Ces modèles sont présentés dans l'annexe A. `Bloom` est doublement intéressant : (a) il est totalement ouvert et facilement disponible ; de surcroît, tous les détails concernant son entraînement (y compris les corpus) sont publics, ce qui permet d'étudier son fonctionnement en profondeur, ou de l'inclure comme modèle de base dans des évaluations comparatives ; (b) par rapport à d'autres modèles comparables, il a été entraîné avec une une large portion de documents français (environ 15% des données d'apprentissage) et aux langues romanes (35% des données), et dans un autre genre, aux langages de programmation.

Dans cet article, nous présentons une évaluation des modèles `Bloom` et `Bloomz` sur une sélection de jeux de données en langue française afin (a) d'établir des performances de référence sur un large éventail de tâches standard du traitement des langues, incluant des jeux de données « généraux » ou plus spécifiques à un domaine de spécialité ; (b) de comparer, lorsque cela était possible et pertinent, les résultats obtenus en faisant varier la langue de l'amorce (français / anglais), voire en faisant varier la langue du jeu de test pour les tests multilingues. L'ensemble des codes et des résultats d'évaluation est accessible à l'adresse `https://github.com/fyvo/EvaluerBloom`.

2 Travaux connexes

Évaluations multitâches et multilingues Un des traits les plus saillants des grands modèles de langue est leur capacité à apprendre des représentations utiles à plusieurs tâches. Depuis (Collobert *et al.*, 2011) (4 tâches d'étiquetage de séquences), les évaluations de ces modèles considèrent un nombre croissant de tâches. Ainsi `SentEval` (Conneau & Kiela, 2018) introduit 7 tâches de classification, quand `GLUE` (Wang *et al.*, 2019b) en considère 9. Avec DecaNLP (McCann *et al.*, 2018) la variété augmente, des tâches plus complexes (traduction, étiquetage de séquence) étant reformulées comme des tâches de réponses à des questions. `SuperGLUE` (Wang *et al.*, 2019a) étend `GLUE` avec des tâches de question-réponses et de coréférence. Les évaluations massivement multi-tâches se sont depuis développées jusqu'à inclure des centaines de tâches (Srivastava *et al.*, 2023). Seuls les

1. Identifiée par des algorithmes de segmentation sous-lexicale (Gage, 1994; Kudo & Richardson, 2018).
2. On parle alors d'affinage par instruction (*instruction fine-tuning*).

modèles autorégressifs, capables de produire des énoncés complets, peuvent aborder toutes ces tâches sans aucune supervision. Pour les modèles *multilingues*, une autre dimension de l'évaluation vise à mesurer la généralisation entre langues par des mécanismes de transfert. Ces capacités sont par exemple mises à l'épreuve par les jeux de test XGLUE (Liang *et al.*, 2020), XTREME (Hu *et al.*, 2020), ou encoreMEGA (Ahuja *et al.*, 2023) et BUFFET (**?**). La construction et l'exploitation de ces benchmarks multilingues peuvent s'appuyer sur des traductions, qui concernent soit les données (ainsi PAWS-X (Yang *et al.*, 2019), X-COPA (Ponti *et al.*, 2020), etc.), soit les amorces (Lin *et al.*, 2022). Ils ont été utilisés de manière répétée pour évaluer des GLM monolingues et multilingues.

En dépit de leur diversité, ces jeux de test ne s'intéressent souvent qu'au niveau d'accomplissement des tâches, mesuré par des métriques idoines. Liang *et al.* (2022) introduit d'autres mesures, comme celles du biais ou des incertitudes, qui ont également leur importance pour l'acceptation de ces modèles. Une autre limitation des grands benchmarks est le caractère parfois artificiel ou académique des traitements considérés ; en réaction, XTREME-UP (Ruder *et al.*, 2023) s'intéresse à un petit nombre de tâches finalisées jugées essentielles (transcription, reconnaissance de caractères, etc).

Pour le français, CamemBERT (Martin *et al.*, 2020) est évalué sur 4 tâches, quand FlauBERT (Le *et al.*, 2020a) est évalué sur l'ensemble des tâches du benchmark FLUE (Le *et al.*, 2020b). Pour évaluer GPT-FR, FLUE est aussi utilisé par Simoulin & Crabbé (2021), qui s'attaquent, de surcroît, au résumé de textes (*zéro-exemple*). Comme expliqué supra, de nombreux jeux de tests multilingues incluent des données en français, et sont donc disponibles pour évaluer des modèles français. C'est le cas, par exemple, de XNLI (Conneau *et al.*, 2018) qui fait partie de nos évaluations. Un travail contemporain (Faysse *et al.*, 2024), introduit enfin *FrenchBench*, qui inclut de multiples tâches pour le traitement du français, certaines originales et portant sur des données nouvelles, par exemple pour l'évaluation des connaissances linguistiques du modèle.

Les évaluations de Bloom La présentation des modèles Bloom et Bloomz s'appuie sur des tâches variées, en particulier celles de SuperGLUE, complétées par des tâches de génération de textes (résumé, traduction) ainsi que des mesures des biais. (BigScience *et al.*, 2022) utilise des amorces formulées en anglais, alors que (Muennighoff *et al.*, 2022) considère de surcroît des démonstrations multilingues à l'apprentissage, associant des exemples multilingues et des amorces en anglais. Les modèles Bloom sont ouverts, et s'appuient sur un corpus d'apprentissage bien documenté – ils constituent donc des modèles de base avec lesquels il est facile de se comparer sur de nombreuses tâches. Notre travail complète donc un certain nombre de travaux présentant des évaluations comparatives de Bloom et Bloomz, voir en particulier (Liang *et al.*, 2022) pour une évaluation « holistique », ainsi que (Wu *et al.*, 2023), qui traite des tâches variées, (Bawden & Yvon, 2023) pour la traduction automatique, (Gallienne & Poibeau, 2023) pour une évaluation des biais, (**?**), qui s'intéresse à l'évaluation du transfert crosslingue, ou encore le travail déjà cité de Faysse *et al.* (2024).

3 Évaluer Bloom avec des données françaises

Pour les évaluations, nous réexploitons le cadre méthodologique utilisé du projet BigScience : les performances sont mesurées en interrogeant directement à l'aide d'amorces pouvant inclure des exemples de la tâche à accomplir. Par comparaison avec l'affinage de modèles, cette méthode est plus légère car elle ne demande pas de réapprentissage ; pour de nombreuses tâches elle conduit

probablement à des performances un peu moindres. Deux composants logiciels sont nécessaires pour la mettre en œuvre : `promptsource` [3], pour construire des amorces [4] et des instructions dédiées au traitement d'un jeu de données et d'une tâche particulière, et `lmharness` [5] (Gao *et al.*, 2021; Biderman *et al.*, 2024), pour spécifier des tâches (définies par des jeux de données, et des métriques) et la manière dont elles s'exécutent (le modèle utilisé, l'amorce, le nombre de démonstrations, les paramètres du processus de génération, etc.) [6].

3.1 Modélisation de la langue

`Bloom` est un modèle causal entraîné sur la tâche de prédiction du prochain mot. Une première évaluation porte donc sur sa capacité à accomplir cette tâche. Les principales métriques sont la *perplexité*, utilisée classiquement en théorie de l'information et le *nombre de bits par octet* (*bits per byte*), introduite pour comparer des modèles utilisant des segmentations différentes :

$$CE(L) = -\frac{1}{L}\log P(w_1 \dots w_T) \text{ [7]} \tag{1}$$

$$\text{bits-par-octet} = CE(|B|), \text{avec |B| la longueur en bytes de } w_1 \dots w_T \text{ [8]} \tag{2}$$

$$\text{perplexité} = 2^{CE(T)} \tag{3}$$

Ces mesures impliquent de choisir un corpus de textes approprié. Simoulin & Crabbé (2021) utilisent un ensemble de 60 articles tirés de la Wikipédia française. Les perplexités obtenues sont respectivement de 109,2 et 12,9 pour le petit (resp. grand) modèle `GPT-FR` (voir (Simoulin & Crabbé, 2021), tableau 4) [9]. Ces articles faisant partie du corpus d'apprentissage de `Bloom`, nous utilisons à la place deux corpus similaires en contenu. Le premier est la partie française de `Flores-101` (Goyal *et al.*, 2022), qui comprend 1 002 phrases traduites de la Wikipédia anglaise : ce corpus multiparallèle permet des comparaisons entre langues et n'a pas été utilisé pour entraîner `Bloom`. En complément, nous reproduisons la méthodologie de (Simoulin & Crabbé, 2021) pour construire le corpus `wikitext-fr-2022` [10]. Ces deux corpus sont décrits dans le tableau 1.

	# articles	# lignes	# tokens
`Flores`-101 (fr) (Goyal *et al.*, 2022)	-	1012	26k
`wikitext-fr` (Simoulin & Crabbé, 2021)	60		897k
`wikitext-fr-2022` (ce travail)	60	4594	443k

TABLE 1 – Corpus pour l'évaluation de la modélisation de la langue.

Le tableau 2 présente les perplexités et bits/octets pour 5 variantes de `Bloom` sur le corpus `Flores-101` et compare les prédictions en français avec trois langues, deux bien représentées dans le corpus `ROOTS` (l'anglais et l'espagnol), et l'allemand qui n'est présent qu'à l'état de traces [11].

3. https://github.com/bigscience-workshop/promptsource.
4. Voir l'annexe B pour une discussion terminologique.
5. https://github.com/bigscience-workshop/lm-evaluation-harness/
6. Nous avons réalisé de diverses contributions à `lmharness`, en particulier pour intégrer des tâches de reconnaissance d'entités nommées et les métriques associées.
9. Le facteur de normalisation pour le calcul de la perplexité est le nombre total de tokens lors de la segmentation du texte.
10. Le code utilisé est celui des auteurs https://github.com/AntoineSimoulin/gpt-fr. Chaque ligne de ce corpus correspond à un court paragraphe.
11. Muennighoff *et al.* (2022) estiment que les textes allemands comptent pour 0,21% des données d'apprentissage.

	Perplexité				Bits/octet			
Langue (Longueur)	de (45 129)	en (26 404)	fr (32 008)	es (32 072)	de (156 358)	en (132 096)	fr (163 927)	es (159 899)
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

TABLE 2 – Performances de Bloom en modélisation des langues (corpus flores-101). Les longueurs indiquées correspondent à un nombre de tokens (pour la perplexité) et à un nombre d'octets.

Comme attendu, pour toutes les langues, les deux métriques s'améliorent avec la taille du modèle. Les comparaisons entre langues sont plus délicates : l'utilisation d'un corpus parallèle rend comparable la complexité du contenu des corpus d'évaluation, mais les facteurs de normalisation impliqués dans le calcul des métriques varient fortement selon les langues (voir la ligne « longueur »). Comme attendu, l'allemand obtient les moins bons scores, malgré des facteurs de normalisation favorables (par rapport à l'anglais) ; les scores pour le français sont également artificiellement favorables par rapport à l'anglais, du fait d'une segmentation en un plus grand nombre de tokens, et d'une plus grande longueur en octets. La comparaison la plus claire est entre français et l'espagnol, qui sont proches en termes de longueur : ici les résultats pour le français semblent indiscutablement meilleurs.

Modèle	Tokens	Perplexité	Bits/Octet	Modèle	Tokens	Perplexité	Bits/Octet
Bloom-560m	698 385	23,88	1,15	Bloom-1b1	698 385	18,69	1,06
Bloom-1b7	698 385	15,81	1,00	Bloom-3b	698 385	13,78	0,95
Bloom-7b1	698 385	11,60	0,89	Bloom	698 385	8,33	0,77
gpt-fr-cased-base	821 598	15,76	1,17				

TABLE 3 – Performance de Bloom en modélisation des langues (corpus Wikipédia-fr).

Les résultats du tableau 3 confirment sur un corpus plus volumineux l'analyse faite sur Flores-101, avec des valeurs de la métrique bits/octets s'étalant de 1,15 pour le plus petit modèle à 0,77 pour le plus grand. L'utilisation du modèle gpt-fr-cased-base de (Simoulin & Crabbé, 2021) (1,017b paramètres) implique un nombre plus élevé d'unités sous-lexicales, ce qui rend sa perplexité incomparable avec celles des modèles Bloom ; en bits/octets il est très proche de Bloom-560m.

3.2 Classification de textes

Pour la classification de textes, nous considérons la tâche Multilingual Amazon Reviews [12], qui propose plusieurs problèmes de prédiction du niveau d'appréciation d'un produit à partir d'un commentaire. La version de base consiste à assigner un score entre 1 et 5 à partir du commentaire ; des variantes n'utilisent que le titre du commentaire, ou bien à la fois le titre et le corps du texte. Cette tâche fait partie de divers benchmarks comme FLUE (Le *et al.*, 2020a) et BUFFET (Asai *et al.*, 2024).

12. https://huggingface.co/datasets/amazon_reviews_multi/viewer/all_languages/test

Langue utilisée (commentaires–amorces)	en–en		fr–en		fr–fr	
Modèle / # exemples	0	1	0	1	0	1
Bloom-560m	0,22	0,22	0,21	0,21	0,21	0,21
Bloom-1b1	0,28	0,24	0,25	0,23	0,27	0,22
Bloom-3b	0,29	0,25	0,25	0,22	0,27	0,22
Bloom-7b1	0,30	0,26	0,27	0,23	0,26	0,23
Bloom	0,34	0,35	0,31	0,29	0,33	0,31
Bloomz	0,48	0,50	0,41	0,45	0,44	0,45

TABLE 4 – Résultats de Bloom et Bloomz pour la classification de textes (Multilingual Amazon Reviews) en variant la langue du commentaire et la langue de l'amorce.

La version de FLUE est toutefois plus simple et distingue deux classes : positifs (score au moins 4) et négatifs (scores au plus 2), les commentaires intermédiaires étant supprimés. Avec cette restriction, FlauBERT (Large) après affinage atteint environ 95% de correction (Le *et al.*, 2020a, tableau 3).

Nos évaluations respectent le cadre standard de tâches de classification : la réponse du système est correcte lorsque la classe de référence est la plus probable parmi les alternatives possibles *listées dans l'amorce* (ici les chiffres de 1 à 5) [13]. Les amorces utilisées s'apparentent toutes au modèle suivant :

(1) Sur la base du titre du commentaire, attribuez un nombre d'étoiles au produit : (1 étoile est la pire note, 5 la meilleure).

Trois amorces sont utilisées : soit le titre seul, soit le corps seul du commentaire, soit la conjonction du titre et du corps du commentaire [14]. Ces formulations reprennent les choix faits en anglais par Sanh *et al.* (2022) et permettent de réaliser des comparaison entre langues. Nous n'utilisons que les données de test, soit 5 000 commentaires pour le français et autant pour l'anglais.

Une première observation est que les scores de classification sont toujours meilleurs quand on exploite simultanément le commentaire et son titre, et ce pour tous les modèles et langues : la correction moyenne est d'environ 6 points meilleure pour cette condition que pour le titre seul. Les résultats détaillés sont dans le tableau 4, qui donne la moyenne (sur les trois amorces) des résultats obtenus en considérant trois associations possibles entre la langue des commentaires et des amorces. Les résultats d'ensemble sont assez médiocres, le meilleur système obtenant environ 50% de correction (en moyenne) sur la prédiction des 5 classes. Bloomz obtient de loin les meilleurs résultats, les autres modèles s'ordonnant en fonction de leur taille. Enfin, les résultats obtenus lorsque commentaires et amorces sont en anglais sont toujours meilleurs que lorsque l'on utilise les mêmes amorces avec des textes français ; traiter la tâche intégralement en français améliore les résultats, sans toutefois complètement combler l'écart avec l'anglais, qui est d'environ 5 points de correction pour Bloomz.

Notons enfin que pour la tâche plus simple de discrimination des commentaires positifs (note > 3) et négatifs (note < 3) le meilleur système français (Bloomz mono-exemple) atteint une correction de 85,6%, près de dix points inférieure au modèle FlauBERT avec affinage.

13. Une discussion des subtilités de l'évaluation des GML est dans (Fournier *et al.*, 2023).
14. Les formulations correspondantes sont « Sur la base du commentaire... », « Sur la base du commentaire et de son titre ».

Amorce / Langues (données–amorces)	Bloom en–en	Bloom fr–en	Bloom fr–fr	Bloomz en–en	Bloomz fr–en	Bloomz fr–fr
based_on_the_previous_passage	0,43	0,40	0,42	0,53	0,51	0,55
can_we_infer	0,40	0,37	0,42	0,45	0,45	0,50
does_it_follow_that	0,37	0,36	0,38	0,47	0,48	0,54
take_the_following_as_truth	0,40	0,36	0,40	0,47	0,44	0,38
Moyenne	0,40	0,37	0,41	0,48	0,47	0,49

TABLE 5 – Résultats de Bloom et Bloomz pour la tâche de calcul d'équivalences sémantiques (corpus de test de XNLI) en *mono-exemple* en faisant varier les langues des textes et des amorces.

3.3 Équivalences sémantiques

Le calcul d'équivalences sémantiques (Monz & de Rijke, 2001; Dagan & Glickman, 2004) prend des formes variées. Elle consiste dans sa version la plus simple à exprimer un jugement binaire ou ternaire sur la relation qui existe entre deux énoncés A et B : A implique-t-il logiquement B ? Ou bien au contraire implique-t-il le contraire de B ? Les deux énoncés sont-ils des paraphrases mutuelles ? Ou bien encore n'ont-ils aucune relation ? Grau & Gleize (2018) discutent les variantes les plus communes, ainsi que des difficultés qu'elles posent aux machines. Cette tâche est intégrée dans les benchmarks GLUE (Wang *et al.*, 2019b) et SuperGLUE (Wang *et al.*, 2019a).

Nous avons ici utilisé XNLI (Conneau *et al.*, 2018), construit par traduction depuis l'anglais en 14 langues d'un ensemble de phrases extraites du corpus MNLI (Bowman *et al.*, 2015; Williams *et al.*, 2018). XNLI fait également partie des tâches considérées pour évaluer Bloom et Bloomz, de nombreux résultats pour d'autres langues sont disponibles dans les présentations de ces modèles. Les données de test comprennent 5 010 paires d'énoncés, chacune associée à une des trois étiquettes possibles pour décrire la relation (implication, contradiction, neutre). Ces données de test (A et B en français) font partie de FLUE (Le *et al.*, 2020a), et sont utilisés dans plusieurs travaux en français : (Le *et al.*, 2020a) rapporte (tableau 5) des scores de correction [15] compris entre 76,9 et 85,2, les modèles FlauBERT atteignant respectivement 80,6 (base) et 83,4 (Large) et CamemBERT 81,2.

Pour réaliser ces évaluations, nous avons traduit en français les amorces de (Bach *et al.*, 2022) (voir l'annexe C, tableau 9). Comme précédemment, la réponse du système est jugée correcte lorsqu'elle est la plus probable *des réponses associées à l'amorce*. Chaque amorce du tableau 9 correspond à un ensemble de trois réponses possibles, formulées à chaque fois dans la même langue que l'amorce.

Nous nous limitons ici à comparer Bloom et Bloomz pour des données anglaises et françaises. Les résultats du tableau 5 confirment les résultats déjà publiés, qui sont assez mauvais quand le nombre d'exemples est faible [16] : devant choisir une des trois continuations possibles de l'amorce, Bloom et Bloomz font à peine mieux que le hasard. Utiliser des amorces en anglais pour les données françaises est moins bon que d'utiliser les amorces en français, avec lesquelles on retrouve des scores proches de ceux que l'on observe avec des amorces et des textes en anglais.

15. Égale au pourcentage de bonnes réponses parmi les sorties du modèle.

16. Pour Bloom et Bloomz (*zéro-exemple*), les chiffres de (BigScience *et al.*, 2022) pour XNLI-FR sont respectivement inférieurs à 0,4 et 0,6 ; ceux de (Ahuja *et al.*, 2023) sont entre 0,6 et 0,7, dans des conditions expérimentales un peu différentes.

3.4 Extraction d'information (reconnaissance d'entités nommées)

La reconnaissance des entités nommées (REN) consiste à identifier et classifier des *entités nommées*, qui sont des mentions qui font référence à des entités du monde réel. On peut s'intéresser à des entités du domaine général (personnes, lieux, organisations), ou à des entités spécifiques à un domaine comme les domaines clinique (maladie, symptôme, partie du corps) ou juridique (court, loi, délit).

Reconnaissance d'entités nommées générales Nous utilisons `WikiNER_fr` [17] (Nothman *et al.*, 2013) qui comporte 13 410 exemples, extraits de Wikipédia et dont les annotations sont construites automatiquement en projetant les annotations en langue anglaise. Les résultats doivent être analysés en gardant à l'esprit le fait que ces annotations sont des pseudo-références (*silver annotations*) [18].

Nous évaluons en utilisant deux stratégies : (i) cibler chaque type d'entité (personne : PER, lieu : LOC et organisation : ORG) et demander aux modèles de lister pour chaque exemple l'ensemble des entités d'un certain type (`LIST_{PER,LOC,ORG}`), (ii) sélectionner aléatoirement une entité de chaque exemple et demander aux modèles de prédire la classe de l'entité parmi les trois types possibles (`choose_entity`), ce qui réduit la tâche à une classification multiclasse. Les amorces utilisées pour les deux types d'évaluation sont présentées dans le tableau 10 de l'annexe C. Une instance du jeu de données initiales (annotations au niveau de chaque token ; 1=lieu, 2=personne, 4=organisation) est reproduit dans l'exemple (2) et sa transformation en amorce de type `LIST_PER` est dans l'exemple (3). Nous fixons la longueur maximum des prédictions à 64 tokens.

(2) Les poètes Joachim du Bellay et Pierre Ronsard sympathisent à l' Université de Poitiers , avant de monter à Paris .
0 0 2 2 2 0 2 2 0 0 0 4 4 4 0 0 0 0 0 1 0

(3) Contexte : Lister les entités de type "personne" dans le texte suivant : Les poètes Joachim du Bellay et Pierre Ronsard sympathisent à l'Université de Poitiers, avant de monter à Paris. Cible : Joachim du Bellay\nPierre Ronsard

En termes de métriques d'évaluation, nous utilisons la correction pour la deuxième stratégie, car il s'agit d'une tâche de classification simple. Pour la première, nous concevons des mesures adaptées au fait que les prédictions, comme les annotations de références, sont des listes. Ces mesures, que l'on désigne comme des approximations *fuzzy* de la précision, du rappel et de la F-mesure, se fondent sur (i) une stratégie heuristique pour diviser les prédictions et les annotations de référence en listes [19], (ii) la suppression d'apostrophes, guillemets et espaces dans chaque élément (pour éviter que les variantes de ponctuation soient comptées en erreur) ; (iii) le calcul du nombre de prédictions correctes sur la base des deux listes résultantes, sans tenir compte de l'ordre de leurs éléments, et enfin (iv) le calcul de la précision, du rappel et de la F-mesure sur toutes les prédictions, tous exemples confondus.

Reconnaissance d'entités nommées cliniques Pour le domaine clinique, nous utilisons le test de `QuaeroFrenchMed` [20] (Névéol *et al.*, 2014), composé de deux parties. La première, `EMEA` est une collection de notices concernant des médicaments commercialisés en Europe, fournis par l'Agence Européenne des Médicaments. La seconde, `MEDLINE`, consiste en 2 500 titres d'articles

17. https://huggingface.co/datasets/Jean-Baptiste/wikiner_fr
18. Nous avons préféré `WikiNER` à `WikiAnn` (Pan *et al.*, 2017), en raison de sa nature plus réaliste ; il contient des phrases plus longues et plus complètes (une moyenne de 26 mots par exemple contre 7 mots en moyenne pour `WikiAnn`).
19. Le délimiteur utilisé est par ordre de priorité le retour chariot, le point-virgule puis la virgule en fonction de la présence ou non du délimiteur dans le texte. Si aucun des délimiteurs est présent, nous considérons que le texte contient un seul item.
20. https://huggingface.co/datasets/mnaguib/QuaeroFrenchMed

Amorce	Modèle	Précision	
		zéro-exemple	*mono-exemple*
choose_entity	Bloom_560m	0,39	0,40
	Bloom_1b1	0,43	0,34
	Bloom_3b	0,42	0,33
	Bloom_7b1	0,41	0,30
	Bloom	0,47	0,53
	Bloomz	0,54	0,55

TABLE 6 – Précision de Bloom et Bloomz pour la REN dans le domaine général (WikiNER-fr) en *zéro-exemple* et *mono-exemple* pour l'amorce choose_entity.

scientifiques indexés dans MEDLINE[21]. Ces deux parties sont annotées en 10 types d'EN correspondant à des groupes sémantiques de l'UMLS (*Unified Medical Language System*) (Bodenreider & McCray, 2003) : symptômes et maladies, parties du corps, composants chimiques, êtres vivants, procédures médicales, physiologie humaine, phénomènes physiologiques, dispositifs médicaux, zones géographiques et objets.

(4) Tysabri est utilisé dans le traitement des adultes atteints de sclérose en plaques
2 0 0 0 0 10 0 6 0 0 4 4 4

La figure 4 présente un extrait d'un document EMEA annoté au niveau de chaque token (2=composants chimiques, 4=symptômes et maladies, 6=êtres humains, 10=procédures médicales). Nous utilisons les deux stratégies d'évaluation décrites ci-dessus, en utilisant des amorces explicitant la nature du document, selon le sous-corpus considéré (annexe C, tableau 11). Nous utilisons les mêmes métriques que pour la REN en domaine général et fixons la longueur maximum des prédictions à 32 tokens.

Résultats Les résultats sur la tâche de REN montrent qu'elle est loin d'être résolue par Bloom, qu'il s'agisse du domaine général ou clinique. Pour WikiNER_fr, les scores restent très bas pour tous les modèles, en dessous de 0,2 (pour la précision comme pour le rappel) pour les trois types d'entités (LOC, PER, ORG, par ordre décroissant de scores) - le tableau complet est en annexe D.1. Le modèle Bloomz en *mono-exemple* atteint des scores nettement supérieurs aux variantes de Bloom, avec des F-mesures de 0,23, de 0,27 et de 0,05 pour PER, LOC, et ORG respectivement. Les scores faibles pour le type ORG, même en *mono-exemple*, illustrent la difficulté posée par ce type, qui inclut des mentions qui ne correspondent pas à des organisations, mais à des groupes de musique, des livres et des produits. Une analyse détaillée est en annexe D.2, qui donne des indications sur les différents types d'erreurs. Les performances sur l'amorce *choose_entity* (tableau 6) sont meilleures, la tâche étant plus contrainte ; les scores vont jusqu'à 0,6 de précision pour Bloomz en *zéro-exemple*. Pour cette tâche, les résultats en *zéro-exemple* sont parfois meilleurs qu'en *mono-exemple*.

Des résultats similaires sont obtenus pour les entités nommées cliniques. Nous remarquons que dans la partie EMEA (notices patients) le modèle réussit à bien identifier la classe prépondérante CHEM (composant chimique). Globalement, les performances augmentent avec la taille des modèles et le nombre d'exemples fournis dans l'amorce. Tous les résultats restent néanmoins inférieurs à ceux des approches supervisées ou symboliques (voir aussi les tableaux 14 et 15 de l'annexe D.3).

21. http://pubmed.ncbi.nlm.nih.gov/

3.5 Réponses aux questions

Amorce		*zéro-exemple* moy.	min.	max.	*mono-exemple* 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
	fr	0,18	0,02	0,73	0,36	0,06	0,71

TABLE 7 – Résultats de Bloom et Bloomz pour la réponse aux questions avec différentes amorces (scores F1, agrégés sur tous les modèles). La colonne 'max' correspond aux résultats de Bloomz.

Pour cette tâche, nous utilisons PIAF (Keraron *et al.*, 2020), un ensemble de paires (question, réponse) extrait de la Wikipédia française au terme d'un développement participatif. La conception de PIAF reproduit celle de sQuad (Rajpurkar *et al.*, 2016, 2018) ; en particulier elle assure que la réponse à la question posée figure toujours dans le passage présenté [22]. Pour cette tâche nous considérons quatre amorces utilisées dans des études similaires, chacune existant en français et en anglais. Un trait important de ces amorces est la position relative de l'instruction, du passage et de la question.

L'évaluation repose sur deux métriques classiques : EM (pour *Exact Match*) correspond au pourcentage de questions pour lesquelles la réponse fournie est exactement la réponse de référence ; F1 combine précision et rappel au niveau des mots de la réponse et fournit une évaluation moins stricte des performances. L'évaluation porte sur l'intégralité des 3 835 questions de PIAF 1.0. L'article qui présente PIAF donne également des scores de performance obtenus par affinage du modèle CamemBERT (Martin *et al.*, 2020) ; ces scores F1 sont tous voisins de 70, avec un maximum de 71,1. Ce jeu de données est également utilisé par Cattan *et al.* (2021) : les meilleurs résultats (pour un ensemble restreint de questions) sont obtenus en affinant le modèle XLM-R$_{large}$ (F1=73,2, EM=45,8).

Le tableau 7 : en premier lieu, la très nette supériorité de Bloomz qui obtient systématiquement les meilleurs résultats, avec un F1 supérieur à 0,73 en *zéro-exemple*. Rappelons que l'ensemble de datasets (xP3) utilisé pour affiner Bloomz comprend plusieurs tâches de réponses aux questions (c.f. §A.1). Par comparaison, le meilleur score de Bloom (*mono-exemple*) est F1=0,53. Pour cette tâche, on observe toujours une variabilité des résultats en fonction de l'amorce, avec une très faible variabilité liée à la langue.

4 Conclusions

Dans cet article, nous avons présenté une évaluation de la famille Bloom sur diverses tâches de traitement automatique de la langue française. Ces expériences complètent les évaluations déjà disponibles pour ces modèles, qu'elles aient été réalisées dans un cadre monolingue ou multilingue : (a) les performances des différents modèles s'ordonnent selon leur taille, avec une nette différence

22. PIAF ne représente que très imparfaitement la variété des tâches de réponses aux questions, voir (Rogers *et al.*, 2023).

entre (a) le modèle le plus gros (176B) et les autres ; (b) le scénario *zéro-exemple* et le scénario *mono-exemple*. L'affinage multitâche améliore les performances *zéro-exemple* pour les tâches de classification ; pour les autres tâches, le bénéfice est moins net, et disparaît presque dès que l'on présente un exemple à l'inférence. Enfin, pour nos tâches, utiliser des amorces en français est toujours meilleur qu'utiliser des entrées multilingues, mélangeant amorces (anglais) et texte en français.

Remerciements

Ce projet a été réalisé grâce au soutien du projet ANR GEM, financé par l'Agence Nationale de la Recherche sous le numéro ANR-19-CE38-0012, à la chaire de R. Bawden à l'institut PRAIRIE, financée par l'Agence Nationale de la Recherche (ANR) dans le cadre du programme « Investissements d'avenir » sous la référence ANR-19-P3IA-0001 et au projet « Émergence », DadaNMT, financé par Sorbonne-Université. Il a également reçu le soutien du CNRS dans le cadre du réseau des ingénieurs du Programme national de recherche en intelligence artificielle (PNRIA) et a utilisé les ressources de calcul de l'IDRIS (allocations 2023-AD010614012 et AD011012254R2 et AD011012254R3 et AD011014533) à travers GENCI.

Références

AHUJA K., HADA R., OCHIENG M., JAIN P., DIDDEE H., MAINA S., GANU T., SEGAL S., AXMED M., BALI K. & SITARAM S. (2023). MEGA : multilingual evaluation of generative AI. *CoRR*, **abs/2303.12528**. DOI : 10.48550/ARXIV.2303.12528.

ASAI A., KUDUGUNTA S., YU X., BLEVINS T., GONEN H., REID M., TSVETKOV Y., RUDER S. & HAJISHIRZI H. (2024). BUFFET : Benchmarking large language models for few-shot cross-lingual transfer. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 1771–1800, Mexico City, Mexico : Association for Computational Linguistics.

BACH S., SANH V., YONG Z. X., WEBSON A., RAFFEL C., NAYAK N. V., SHARMA A., KIM T., BARI M. S., FEVRY T., ALYAFEAI Z., DEY M., SANTILLI A., SUN Z., BEN-DAVID S., XU C., CHHABLANI G., WANG H., FRIES J., AL-SHAIBANI M., SHARMA S., THAKKER U., ALMUBARAK K., TANG X., RADEV D., JIANG M. T.-J. & RUSH A. (2022). PromptSource : An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, p. 93–104, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-demo.9.

BAWDEN R. & YVON F. (2023). Investigating the Translation Performance of a Large Multilingual Language Model : the Case of BLOOM. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, Tampere, Finland. DOI : 10.48550/ARXIV.2303.01911, HAL : hal-04015863.

BIDERMAN S., SCHOELKOPF H., SUTAWIKA L., GAO L., TOW J., ABBASI B., AJI A. F., AMMANAMANCHI P. S., BLACK S., CLIVE J., DIPOFI A., ETXANIZ J., FATTORI B., FORDE J. Z., FOSTER C., HSU J., JAISWAL M., LEE W. Y., LI H., LOVERING C., MUENNIGHOFF N.,

Pavlick E., Phang J., Skowron A., Tan S., Tang X., Wang K. A., Winata G. I., Yvon F. & Zou A. (2024). Lessons from the trenches on reproducible evaluation of language models.

BigScience W., Scao T. L., Fan A., Akiki C., Pavlick E., Ilić S., Hesslow D., Castagné R., Luccioni A. S., Yvon F., Gallé M., Tow J., Rush A. M., Biderman S., Webson A., Ammanamanchi P. S., Wang T., Sagot B., Muennighoff N., del Moral A. V., Ruwase O., Bawden R., Bekman S., McMillan-Major A., Beltagy I., Nguyen H., Saulnier L., Tan S., Suarez P. O., Sanh V., Laurençon H., Jernite Y., Launay J., Mitchell M., Raffel C., Gokaslan A., Simhi A., Soroa A., Aji A. F., Alfassy A., Rogers A., Nitzav A. K., Xu C., Mou C., Emezue C., Klamm C., Leong C., van Strien D., Adelani D. I., Radev D., Ponferrada E. G., Levkovizh E., Kim E., Natan E. B., De Toni F., Dupont G., Kruszewski G., Pistilli G., Elsahar H., Benyamina H., Tran H., Yu I., Abdulmumin I., Johnson I., Gonzalez-Dios I., de la Rosa J., Chim J., Dodge J., Zhu J., Chang J., Frohberg J., Tobing J., Bhattacharjee J., Almubarak K., Chen K., Lo K., Von Werra L., Weber L., Phan L., allal L. B., Tanguy L., Dey M., Muñoz M. R., Masoud M., Grandury M., Šaško M., Huang M., Coavoux M., Singh M., Jiang M. T.-J., Vu M. C., Jauhar M. A., Ghaleb M., Subramani N., Kassner N., Khamis N., Nguyen O., Espejel O., de Gibert O., Villegas P., Henderson P., Colombo P., Amuok P., Lhoest Q., Harliman R., Bommasani R., López R. L., Ribeiro R., Osei S., Pyysalo S., Nagel S., Bose S., Muhammad S. H., Sharma S., Longpre S., Nikpoor S., Silberberg S., Pai S., Zink S., Torrent T. T., Schick T., Thrush T., Danchev V., Nikoulina V., Laippala V., Lepercq V., Prabhu V., Alyafeai Z., Talat Z., Raja A., Heinzerling B., Si C., Taşar D. E., Salesky E., Mielke S. J., Lee W. Y., Sharma A., Santilli A., Chaffin A., Stiegler A., Datta D., Szczechla E., Chhablani G., Wang H., Pandey H., Strobelt H., Fries J. A., Rozen J., Gao L., Sutawika L., Bari M. S., Al-shaibani M. S., Manica M., Nayak N., Teehan R., Albanie S., Shen S., Ben-David S., Bach S. H., Kim T., Bers T., Fevry T., Neeraj T., Thakker U., Raunak V., Tang X., Yong Z.-X., Sun Z., Brody S., Uri Y., Tojarieh H., Roberts A., Chung H. W., Tae J., Phang J., Press O., Li C., Narayanan D., Bourfoune H., Casper J., Rasley J., Ryabinin M., Mishra M., Zhang M., Shoeybi M., Peyrounette M., Patry N., Tazi N., Sanseviero O., von Platen P., Cornette P., Lavallée P. F., Lacroix R., Rajbhandari S., Gandhi S., Smith S., Requena S., Patil S., Dettmers T., Baruwa A., Singh A., Cheveleva A., Ligozat A.-L., Subramonian A., Névéol A., Lovering C., Garrette D., Tunuguntla D., Reiter E., Taktasheva E., Voloshina E., Bogdanov E., Winata G. I., Schoelkopf H., Kalo J.-C., Novikova J., Forde J. Z., Clive J., Kasai J., Kawamura K., Hazan L., Carpuat M., Clinciu M., Kim N., Cheng N., Serikov O., Antverg O., van der Wal O., Zhang R., Zhang R., Gehrmann S., Mirkin S., Pais S., Shavrina T., Scialom T., Yun T., Limisiewicz T., Rieser V., Protasov V., Mikhailov V., Pruksachatkun Y., Belinkov Y., Bamberger Z., Kasner Z., Rueda A., Pestana A., Feizpour A., Khan A., Faranak A., Santos A., Hevia A., Unldreaj A., Aghagol A., Abdollahi A., Tammour A., HajiHosseini A., Behroozi B., Ajibade B., Saxena B., Ferrandis C. M., Contractor D., Lansky D., David D., Kiela D., Nguyen D. A., Tan E., Baylor E., Ozoani E., Mirza F., Ononiwu F., Rezanejad H., Jones H., Bhattacharya I., Solaiman I., Sedenko I., Nejadgholi I., Passmore J., Seltzer J., Sanz J. B., Dutra L., Samagaio M., Elbadri M., Mieskes M., Gerchick M., Akinlolu M., McKenna M., Qiu M., Ghauri M., Burynok M., Abrar N., Rajani N., Elkott N., Fahmy N., Samuel O., An R., Kromann R., Hao R., Alizadeh S., Shubber S., Wang S., Roy S., Viguier S., Le T., Oyebade T., Le T., Yang Y., Nguyen Z., Kashyap A. R.,

PALASCIANO A., CALLAHAN A., SHUKLA A., MIRANDA-ESCALADA A., SINGH A., BEILHARZ B., WANG B., BRITO C., ZHOU C., JAIN C., XU C., FOURRIER C., PERIÑÁN D. L., MOLANO D., YU D., MANJAVACAS E., BARTH F., FUHRIMANN F., ALTAY G., BAYRAK G., BURNS G., VRABEC H. U., BELLO I., DASH I., KANG J., GIORGI J., GOLDE J., POSADA J. D., SIVARAMAN K. R., BULCHANDANI L., LIU L., SHINZATO L., DE BYKHOVETZ M. H., TAKEUCHI M., PÀMIES M., CASTILLO M. A., NEZHURINA M., SÄNGER M., SAMWALD M., CULLAN M., WEINBERG M., DE WOLF M., MIHALJCIC M., LIU M., FREIDANK M., KANG M., SEELAM N., DAHLBERG N., BROAD N. M., MUELLNER N., FUNG P., HALLER P., CHANDRASEKHAR R., EISENBERG R., MARTIN R., CANALLI R., SU R., SU R., CAHYAWIJAYA S., GARDA S., DESHMUKH S. S., MISHRA S., KIBLAWI S., OTT S., SANG-AROONSIRI S., KUMAR S., SCHWETER S., BHARATI S., LAUD T., GIGANT T., KAINUMA T., KUSA W., LABRAK Y., BAJAJ Y. S., VENKATRAMAN Y., XU Y., XU Y., XU Y., TAN Z., XIE Z., YE Z., BRAS M., BELKADA Y. & WOLF T. (2022). BLOOM : A 176b-parameter open-access multilingual language model. *CoRR*, **abs/2211.05100**. DOI : 10.48550/arXiv.2211.05100.

BODENREIDER O. & MCCRAY A. T. (2003). Exploring semantic groups through visual approaches. *Journal of biomedical informatics*, **36**(6), 414–432.

BOWMAN S. R., ANGELI G., POTTS C. & MANNING C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 632–642, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1075.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CATTAN O., SERVAN C. & ROSSET S. (2021). On the usability of transformers-based models for a French question-answering task. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, p. 244–255, Held Online : INCOMA Ltd.

COLLOBERT R., WESTON J., BOTTOU L., KARLEN M., KAVUKCUOGLU K. & KUKSA P. (2011). Natural language processing (almost) from scratch. *Journal Machine Learning Research*, **12**, 2493–2537.

CONNEAU A. & KIELA D. (2018). SentEval : An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

CONNEAU A., RINOTT R., LAMPLE G., WILLIAMS A., BOWMAN S., SCHWENK H. & STOYANOV V. (2018). XNLI : Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, p. 2475–2485, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-1269.

DAGAN I. & GLICKMAN O. (2004). Probabilistic textual entailment : Generic applied modeling of language variability. *Learning Methods for Text Understanding and Mining*, **2004**(26-29), 2–5.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the*

North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), p. 4171–4186, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1423.

FAYSSE M., FERNANDES P., GUERREIRO N., LOISON A., ALVES D., CORRO C., BOIZARD N., ALVES J., REI R., MARTINS P., CASADEMUNT A. B., YVON F., MARTINS A., VIAUD G., HUDELOT C. & COLOMBO P. (2024). CroissantLLM : A Truly Bilingual French-English Language Model. *CoRR*, **abs/2402.00786**.

FOURNIER C., HABIB N., LAUNAY J. & WOLF T. (2023). What's going on with the Open LLM leaderboard ? Blog post, last visited on december 6th, 2023.

GAGE P. (1994). A new algorithm for data compression. *Computer Users Journal*, **12**(2), 23–38.

GALLIENNE R. & POIBEAU T. (2023). Quelques observations sur la notion de biais dans les modèles de langue. In C. SERVAN & A. VILNAT, Éds., *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 3 : prises de position en TAL*, p. 1–13, Paris, France : ATALA.

GAO L., TOW J., BIDERMAN S., BLACK S., DIPOFI A., FOSTER C., GOLDING L., HSU J., MCDONELL K., MUENNIGHOFF N., PHANG J., REYNOLDS L., TANG E., THITE A., WANG B., WANG K. & ZOU A. (2021). A framework for few-shot language model evaluation. https ://doi.org/10.5281/zenodo.5371628. DOI : 10.5281/zenodo.5371628.

GOYAL N., GAO C., CHAUDHARY V., CHEN P.-J., WENZEK G., JU D., KRISHNAN S., RANZATO M., GUZMÁN F. & FAN A. (2022). The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, **10**, 522–538. DOI : 10.1162/tacl_a_00474.

GRAU B. & GLEIZE M. (2018). Implication textuelle : problèmes et méthodes pour le TAL. *Langages*, **212**(4), 105–122. DOI : 10.3917/lang.212.0105.

HOULSBY N., GIURGIU A., JASTRZEBSKI S., MORRONE B., DE LAROUSSILHE Q., GESMUNDO A., ATTARIYAN M. & GELLY S. (2019). Parameter-efficient transfer learning for NLP. In K. CHAUDHURI & R. SALAKHUTDINOV, Éds., *Proceedings of the 36th International Conference on Machine Learning*, volume 97 de *Proceedings of Machine Learning Research*, p. 2790–2799 : PMLR.

HU J., RUDER S., SIDDHANT A., NEUBIG G., FIRAT O. & JOHNSON M. (2020). XTREME : A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In H. D. III & A. SINGH, Éds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 de *Proceedings of Machine Learning Research*, p. 4411–4421 : PMLR.

KERARON R., LANCRENON G., BRAS M., ALLARY F., MOYSE G., SCIALOM T., SORIANO-MORALES E.-P. & STAIANO J. (2020). Project PIAF : Building a native French question-answering dataset. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5481–5490, Marseille, France : European Language Resources Association.

KUDO T. & RICHARDSON J. (2018). SentencePiece : A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing : System Demonstrations*, p. 66–71, Brussels, Belgium : Association for Computational Linguistics. DOI : 10.18653/v1/D18-2012.

LAURENÇON H., SAULNIER L., WANG T., AKIKI C., DEL MORAL A. V., SCAO T. L., WERRA L. V., MOU C., PONFERRADA E. G., NGUYEN H., FROHBERG J., ŠAŠKO M., LHOEST Q., MCMILLAN-MAJOR A., DUPONT G., BIDERMAN S., ROGERS A., ALLAL L. B., TONI F. D.,

PISTILLI G., NGUYEN O., NIKPOOR S., MASOUD M., COLOMBO P., DE LA ROSA J., VILLEGAS P., THRUSH T., LONGPRE S., NAGEL S., WEBER L., MUÑOZ M. R., ZHU J., STRIEN D. V., ALYAFEAI Z., ALMUBARAK K., CHIEN V. M., GONZALEZ-DIOS I., SOROA A., LO K., DEY M., SUAREZ P. O., GOKASLAN A., BOSE S., ADELANI D. I., PHAN L., TRAN H., YU I., PAI S., CHIM J., LEPERCQ V., ILIC S., MITCHELL M., LUCCIONI S. & JERNITE Y. (2022). The BigScience ROOTS corpus : A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020a). FlauBERT : des modèles de langue contextualisés pré-entraînés pour le français. In *Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 2 : Traitement Automatique des Langues Naturelles*, p. 268–278, Nancy, France : ATALA et AFCP.

LE H., VIAL L., FREJ J., SEGONNE V., COAVOUX M., LECOUTEUX B., ALLAUZEN A., CRABBÉ B., BESACIER L. & SCHWAB D. (2020b). FlauBERT : Unsupervised language model pre-training for French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France : European Language Resources Association.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIANG P., BOMMASANI R., LEE T., TSIPRAS D., SOYLU D., YASUNAGA M., ZHANG Y., NARAYANAN D., WU Y., KUMAR A., NEWMAN B., YUAN B., YAN B., ZHANG C., COSGROVE C., MANNING C. D., RÉ C., ACOSTA-NAVAS D., HUDSON D. A., ZELIKMAN E., DURMUS E., LADHAK F., RONG F., REN H., YAO H., WANG J., SANTHANAM K., ORR L., ZHENG L., YUKSEKGONUL M., SUZGUN M., KIM N., GUHA N., CHATTERJI N., KHATTAB O., HENDERSON P., HUANG Q., CHI R., XIE S. M., SANTURKAR S., GANGULI S., HASHIMOTO T., ICARD T., ZHANG T., CHAUDHARY V., WANG W., LI X., MAI Y., ZHANG Y. & KOREEDA Y. (2022). Holistic evaluation of language models. arXiv preprint. DOI : 10.48550/ARXIV.2211.09110.

LIANG Y., DUAN N., GONG Y., WU N., GUO F., QI W., GONG M., SHOU L., JIANG D., CAO G., FAN X., ZHANG R., AGRAWAL R., CUI E., WEI S., BHARTI T., QIAO Y., CHEN J.-H., WU W., LIU S., YANG F., CAMPOS D., MAJUMDER R. & ZHOU M. (2020). XGLUE : A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 6008–6018, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.484.

LIN X. V., MIHAYLOV T., ARTETXE M., WANG T., CHEN S., SIMIG D., OTT M., GOYAL N., BHOSALE S., DU J., PASUNURU R., SHLEIFER S., KOURA P. S., CHAUDHARY V., O'HORO B., WANG J., ZETTLEMOYER L., KOZAREVA Z., DIAB M., STOYANOV V. & LI X. (2022). Few-shot learning with multilingual generative language models. In Y. GOLDBERG, Z. KOZAREVA & Y. ZHANG, Éds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 9019–9052, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.616.

MARTIN L., MULLER B., ORTIZ SUÁREZ P. J., DUPONT Y., ROMARY L., DE LA CLERGERIE É., SEDDAH D. & SAGOT B. (2020). CamemBERT : a tasty French language model. In *Proceedings of*

the 58th Annual Meeting of the Association for Computational Linguistics, p. 7203–7219, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.645.

MCCANN B., KESKAR N. S., XIONG C. & SOCHER R. (2018). The natural language decathlon : Multitask learning as question answering. *CoRR*, **abs/1806.08730**.

MONZ C. & DE RIJKE M. (2001). Light-weight entailment checking for computational semantics. In *Proc. of the third workshop on inference in computational semantics (ICoS-3)*.

MUENNIGHOFF N., WANG T., SUTAWIKA L., ROBERTS A., BIDERMAN S., SCAO T. L., BARI M. S., SHEN S., YONG Z. X., SCHOELKOPF H., TANG X., RADEV D., AJI A. F., ALMUBARAK K., ALBANIE S., ALYAFEAI Z., WEBSON A., RAFF E. & RAFFEL C. (2022). Crosslingual generalization through multitask finetuning. *CoRR*, **abs/2211.01786**. DOI : 10.48550/arXiv.2211.01786.

NÉVÉOL A., GROUIN C., LEIXA J., ROSSET S. & ZWEIGENBAUM P. (2014). The quaero french medical corpus : A ressource for medical entity recognition and normalization. In *Proceedings of the Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing*, BioTxtM 2014, p. 24–30, Reykjavik, Iceland.

NOTHMAN J., RINGLAND N., RADFORD W., MURPHY T. & CURRAN J. R. (2013). Learning multilingual named entity recognition from wikipedia. *Artificial Intelligence*, **194**, 151–175. Artificial Intelligence, Wikipedia and Semi-Structured Resources, DOI : https ://doi.org/10.1016/j.artint.2012.03.006.

ORTIZ SUÁREZ P. J., SAGOT B. & ROMARY L. (2019). Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In P. BAŃSKI, A. BARBARESI, H. BIBER, E. BREITENEDER, S. CLEMATIDE, M. KUPIETZ, H. LÜNGEN & C. ILIADI, Éds., *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, p. 9–16, Mannheim : Leibniz-Institut für Deutsche Sprache. DOI : 10.14618/ids-pub-9021.

PAN X., ZHANG B., MAY J., NOTHMAN J., KNIGHT K. & JI H. (2017). Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 1946–1958, Vancouver, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/P17-1178.

PONTI E. M., GLAVAŠ G., MAJEWSKA O., LIU Q., VULIĆ I. & KORHONEN A. (2020). XCOPA : A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 2362–2376, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.185.

PRESS O., SMITH N. & LEWIS M. (2022). Train short, test long : Attention with linear biases enables input length extrapolation. In *International Conference on Learning Representations*.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D., SUTSKEVER I. *et al.* (2019). Language models are unsupervised multitask learners. OpenAI blog.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, **21**(140), 1–67.

RAJPURKAR P., JIA R. & LIANG P. (2018). Know what you don't know : Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2 : Short Papers)*, p. 784–789, Melbourne, Australia : Association for Computational Linguistics. DOI : 10.18653/v1/P18-2124.

RAJPURKAR P., ZHANG J., LOPYREV K. & LIANG P. (2016). SQuAD : 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods*

in Natural Language Processing, p. 2383–2392, Austin, Texas : Association for Computational Linguistics. DOI : 10.18653/v1/D16-1264.

ROGERS A., GARDNER M. & AUGENSTEIN I. (2023). Qa dataset explosion : A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, **55**(10). DOI : 10.1145/3560260.

RUDER S., CLARK J. H., GUTKIN A., KALE M., MA M., NICOSIA M., RIJHWANI S., RILEY P., SARR J. M. A., WANG X., WIETING J., GUPTA N., KATANOVA A., KIROV C., DICKINSON D. L., ROARK B., SAMANTA B., TAO C., ADELANI D. I., AXELROD V., CASWELL I., CHERRY C., GARRETTE D., INGLE R. R., JOHNSON M., PANTELEEV D. & TALUKDAR P. (2023). XTREME-UP : A user-centric scarce-data benchmark for under-represented languages. *CoRR*, **abs/2305.11938**. DOI : 10.48550/ARXIV.2305.11938.

SANH V., WEBSON A., RAFFEL C., BACH S., SUTAWIKA L., ALYAFEAI Z., CHAFFIN A., STIEGLER A., RAJA A., DEY M., BARI M. S., XU C., THAKKER U., SHARMA S. S., SZCZECHLA E., KIM T., CHHABLANI G., NAYAK N., DATTA D., CHANG J., JIANG M. T.-J., WANG H., MANICA M., SHEN S., YONG Z. X., PANDEY H., BAWDEN R., WANG T., NEERAJ T., ROZEN J., SHARMA A., SANTILLI A., FEVRY T., FRIES J. A., TEEHAN R., SCAO T. L., BIDERMAN S., GAO L., WOLF T. & RUSH A. M. (2022). Multitask prompted training enables zero-shot task generalization. In *Proceedings of the International Conference on Learning Representations*.

SIMOULIN A. & CRABBÉ B. (2021). Un modèle transformer génératif pré-entrainé pour le_____ français. In *Actes de la 28e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, p. 246–255, Lille, France : ATALA.

SRIVASTAVA A., RASTOGI A., RAO A., SHOEB A. A. M., ABID A., FISCH A., BROWN A. R., SANTORO A., GUPTA A., GARRIGA-ALONSO A., KLUSKA A., LEWKOWYCZ A., AGARWAL A., POWER A., RAY A., WARSTADT A., KOCUREK A. W., SAFAYA A., TAZARV A., XIANG A., PARRISH A., NIE A., HUSSAIN A., ASKELL A., DSOUZA A., SLONE A., RAHANE A., IYER A. S., ANDREASSEN A. J., MADOTTO A., SANTILLI A., STUHLMÜLLER A., DAI A. M., LA A., LAMPINEN A., ZOU A., JIANG A., CHEN A., VUONG A., GUPTA A., GOTTARDI A., NORELLI A., VENKATESH A., GHOLAMIDAVOODI A., TABASSUM A., MENEZES A., KIRUBARAJAN A., MULLOKANDOV A., SABHARWAL A., HERRICK A., EFRAT A., ERDEM A., KARAKAŞ A., ROBERTS B. R., LOE B. S., ZOPH B., BOJANOWSKI B., ÖZYURT B., HEDAYATNIA B., NEYSHABUR B., INDEN B., STEIN B., EKMEKCI B., LIN B. Y., HOWALD B., ORINION B., DIAO C., DOUR C., STINSON C., ARGUETA C., FERRI C., SINGH C., RATHKOPF C., MENG C., BARAL C., WU C., CALLISON-BURCH C., WAITES C., VOIGT C., MANNING C. D., POTTS C., RAMIREZ C., RIVERA C. E., SIRO C., RAFFEL C., ASHCRAFT C., GARBACEA C., SILEO D., GARRETTE D., HENDRYCKS D., KILMAN D., ROTH D., FREEMAN C. D., KHASHABI D., LEVY D., GONZÁLEZ D. M., PERSZYK D., HERNANDEZ D., CHEN D., IPPOLITO D., GILBOA D., DOHAN D., DRAKARD D., JURGENS D., DATTA D., GANGULI D., EMELIN D., KLEYKO D., YURET D., CHEN D., TAM D., HUPKES D., MISRA D., BUZAN D., MOLLO D. C., YANG D., LEE D.-H., SCHRADER D., SHUTOVA E., CUBUK E. D., SEGAL E., HAGERMAN E., BARNES E., DONOWAY E., PAVLICK E., RODOLÀ E., LAM E., CHU E., TANG E., ERDEM E., CHANG E., CHI E. A., DYER E., JERZAK E., KIM E., MANYASI E. E., ZHELTONOZHSKII E., XIA F., SIAR F., MARTÍNEZ-PLUMED F., HAPPÉ F., CHOLLET F., RONG F., MISHRA G., WINATA G. I., DE MELO G., KRUSZEWSKI G., PARASCANDOLO G., MARIANI G., WANG G. X., JAIMOVITCH-LOPEZ G., BETZ G., GUR-ARI G., GALIJASEVIC H., KIM H., RASHKIN H., HAJISHIRZI H., MEHTA H., BOGAR H., SHEVLIN H. F. A., SCHUETZE H., YAKURA H., ZHANG H., WONG H. M., NG I., NOBLE I., JUMELET J., GEISSINGER J., KERNION J., HILTON J., LEE J., FISAC J. F., SIMON

J. B., Koppel J., Zheng J., Zou J., Kocon J., Thompson J., Wingfield J., Kaplan J., Radom J., Sohl-Dickstein J., Phang J., Wei J., Yosinski J., Novikova J., Bosscher J., Marsh J., Kim J., Taal J., Engel J., Alabi J., Xu J., Song J., Tang J., Waweru J., Burden J., Miller J., Balis J. U., Batchelder J., Berant J., Frohberg J., Rozen J., Hernandez-Orallo J., Boudeman J., Guerr J., Jones J., Tenenbaum J. B., Rule J. S., Chua J., Kanclerz K., Livescu K., Krauth K., Gopalakrishnan K., Ignatyeva K., Markert K., Dhole K., Gimpel K., Omondi K., Mathewson K. W., Chiafullo K., Shkaruta K., Shridhar K., McDonell K., Richardson K., Reynolds L., Gao L., Zhang L., Dugan L., Qin L., Contreras-Ochando L., Morency L.-P., Moschella L., Lam L., Noble L., Schmidt L., He L., Oliveros-Colón L., Metz L., Senel L. K., Bosma M., Sap M., Hoeve M. T., Farooqi M., Faruqui M., Mazeika M., Baturan M., Marelli M., Maru M., Ramirez-Quintana M. J., Tolkiehn M., Giulianelli M., Lewis M., Potthast M., Leavitt M. L., Hagen M., Schubert M., Baitemirova M. O., Arnaud M., McElrath M., Yee M. A., Cohen M., Gu M., Ivanitskiy M., Starritt M., Strube M., Swędrowski M., Bevilacqua M., Yasunaga M., Kale M., Cain M., Xu M., Suzgun M., Walker M., Tiwari M., Bansal M., Aminnaseri M., Geva M., Gheini M., T M. V., Peng N., Chi N. A., Lee N., Krakover N. G.-A., Cameron N., Roberts N., Doiron N., Martinez N., Nangia N., Deckers N., Muennighoff N., Keskar N. S., Iyer N. S., Constant N., Fiedel N., Wen N., Zhang O., Agha O., Elbaghdadi O., Levy O., Evans O., Casares P. A. M., Doshi P., Fung P., Liang P. P., Vicol P., Alipoormolabashi P., Liao P., Liang P., Chang P. W., Eckersley P., Htut P. M., Hwang P., Miłkowski P., Patil P., Pezeshkpour P., Oli P., Mei Q., Lyu Q., Chen Q., Banjade R., Rudolph R. E., Gabriel R., Habacker R., Risco R., Millière R., Garg R., Barnes R., Saurous R. A., Arakawa R., Raymaekers R., Frank R., Sikand R., Novak R., Sitelew R., Bras R. L., Liu R., Jacobs R., Zhang R., Salakhutdinov R., Chi R. A., Lee S. R., Stovall R., Teehan R., Yang R., Singh S., Mohammad S. M., Anand S., Dillavou S., Shleifer S., Wiseman S., Gruetter S., Bowman S. R., Schoenholz S. S., Han S., Kwatra S., Rous S. A., Ghazarian S., Ghosh S., Casey S., Bischoff S., Gehrmann S., Schuster S., Sadeghi S., Hamdan S., Zhou S., Srivastava S., Shi S., Singh S., Asaadi S., Gu S. S., Pachchigar S., Toshniwal S., Upadhyay S., Debnath S. S., Shakeri S., Thormeyer S., Melzi S., Reddy S., Makini S. P., Lee S.-H., Torene S., Hatwar S., Dehaene S., Divic S., Ermon S., Biderman S., Lin S., Prasad S., Piantadosi S., Shieber S., Misherghi S., Kiritchenko S., Mishra S., Linzen T., Schuster T., Li T., Yu T., Ali T., Hashimoto T., Wu T.-L., Desbordes T., Rothschild T., Phan T., Wang T., Nkinyili T., Schick T., Kornev T., Tunduny T., Gerstenberg T., Chang T., Neeraj T., Khot T., Shultz T., Shaham U., Misra V., Demberg V., Nyamai V., Raunak V., Ramasesh V. V., vinay uday prabhu, Padmakumar V., Srikumar V., Fedus W., Saunders W., Zhang W., Vossen W., Ren X., Tong X., Zhao X., Wu X., Shen X., Yaghoobzadeh Y., Lakretz Y., Song Y., Bahri Y., Choi Y., Yang Y., Hao Y., Chen Y., Belinkov Y., Hou Y., Hou Y., Bai Y., Seid Z., Zhao Z., Wang Z., Wang Z. J., Wang Z. & Wu Z. (2023). Beyond the imitation game : Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. u. & Polosukhin I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan & R. Garnett, Éds., *Advances in Neural Information Processing Systems 30*, p. 5998–6008 : Curran Associates, Inc.

Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O. &

BOWMAN S. (2019a). SuperGLUE : A stickier benchmark for general-purpose language understanding systems. **32**.

WANG A., SINGH A., MICHAEL J., HILL F., LEVY O. & BOWMAN S. R. (2019b). GLUE : A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations*.

WILLIAMS A., NANGIA N. & BOWMAN S. (2018). A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long Papers)*, p. 1112–1122, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/N18-1101.

WU S., IRSOY O., LU S., DABRAVOLSKI V., DREDZE M., GEHRMANN S., KAMBADUR P., ROSENBERG D. S. & MANN G. (2023). Bloomberggpt : A large language model for finance. *CoRR*, **abs/2303.17564**. DOI : 10.48550/ARXIV.2303.17564.

YANG Y., ZHANG Y., TAR C. & BALDRIDGE J. (2019). PAWS-X : A cross-lingual adversarial dataset for paraphrase identification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, p. 3687–3692, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1382.

A Description de `Bloom`

Le projet BigScience [23] est une collaboration internationale assemblée durant les années 2021-2022 dans le but d'entraîner et d'évaluer un giga modèle de langue selon les principes de la science ouverte. Le résultat principal du projet est un ensemble de modèles de langue auto-régressifs (causaux) fondés sur des architectures *Transformer* (Vaswani *et al.*, 2017) dont la taille varie de 560 millions à 175 milliards de paramètres (voir le tableau 8, qui reprend les principales dimensions de ces modèles). Ces modèles sont documentés principalement dans (BigScience *et al.*, 2022), ainsi que sur le hub de HuggingFace [24] depuis lesquels ils peuvent être téléchargés. Tous ces modèles utilisent le même vocabulaire comprenant 250 680 unités sous-lexicales ou *tokens* déterminées par analyse d'un sous-ensemble du corpus d'apprentissage avec l'algorithme BPE (Gage, 1994). Cet algorithme est appliqué sur des textes pré-segmentés, traités comme des séquences d'octets. Tous les modèles de cette famille partagent également l'utilisation de la méthode ALIBI (Press *et al.*, 2022) en replacement des plongements positionnels du *Transformer* de base.

L'apprentissage de ces modèles exploite le corpus `ROOTS` [25] (Laurençon *et al.*, 2022), qui est un corpus multilingue comprenant des textes en 46 langues, ainsi que des programmes informatiques écrits en 14 langages de programmation (pour environ 11% du corpus). Les textes en langue française comptent sur environ 13% des données, à comparer avec 30% d'anglais, 16,2% de chinois et 11% d'espagnol [26]. La partie française du corpus d'apprentissage correspond donc à environ 210Gb, soit 45 milliards de tokens ; par comparaison les modèles `CamemBERT` (Martin *et al.*, 2020) et `FlauBERT`

23. https://bigscience.huggingface.co/
24. https://huggingface.co/bigscience/bloom
25. Consultable à https://huggingface.co/spaces/bigscience-data/roots-search.
26. Les langues officiellement couvertes par `Bloom` sont documentées dans la carte d'identité du modèle : https://huggingface.co/bigscience/bloom.

	Hyper-paramètres			
	Couches	Dimension interne	Têtes d'attention	Nombre paramètres
Bloom-560M	24	1 024	16	559M
Bloom-1.1B	24	1 536	16	1 065M
Bloom-1.7B	24	2 048	16	1 722M
Bloom-3B	30	2 560	32	3 003M
Bloom-7.1B	30	4 096	32	7 069M
Bloom	70	14 336	112	176 247M

TABLE 8 – Les modèles de la famille Bloom.

(Le *et al.*, 2020a) se fondent sur des corpus d'apprentissage de respectivement 138Gb et 71Gb. Ces textes proviennent pour partie de corpus déjà constitués, représentant un vaste ensemble de genres textuels et de thèmes (littérature, textes journalistiques, documents émanant d'institutions internationales, encyclopédie, matériel pédagogique, sous-titres, etc), soit à des documents collectés sur le web et rassemblés dans le corpus OSCAR (Ortiz Suárez *et al.*, 2019).

A.1 Description de Bloomz

Bloomz [27] (Muennighoff *et al.*, 2022) est un modèle dérivé de Bloom par affinage, en poursuivant le processus d'apprentissage sur des séquences textuelles reproduisant des amorces pour un grand nombre de tâches. Ce travail étend la méthodologie de (Sanh *et al.*, 2022) en considérant des jeux de tests multilingues [28]. Il existe également une version (Bloomz-mt) pour laquelle des amorces multilingues (obtenues par traduction automatique) sont utilisées durant l'affinage ; ce modèle n'a pas été considéré dans nos expériences. Les évaluations de Bloomz, publiées dans (Muennighoff *et al.*, 2022), montrent qu'il surpasse considérablement Bloom pour l'utilisation en *zéro-exemple*.

L'affinage de Bloomz prend en compte une grande variété de tâches : complétion de textes, classification, analyse de sentiments, réponse à des questions, identification de paraphrases, désambiguïsation sémantique, résumé, simplification et traduction automatique pour citer les principales. Il est important de noter que cet affinage utilise certains de jeux de tests fréquemment utilisés : c'est le cas de amazon pour l'analyse des sentiments (section 3.2) ; de wikilingua pour le résumé automatique et de flores pour la traduction. Pour ces jeux de test, les résultats de Bloomz surestiment les performances réelles de ce modèle.

B Note Terminologique

Nous avons recours dans ce rapport une terminologie délibérément francisée. En particulier, nous utilisons systématiquement le terme 'amorce' pour traduire l'anglais '*prompt*', sans nécessairement distinguer la partie qui désigne *l'instruction* (la spécification de la tâche à accomplir) de la

27. https://huggingface.co/bigscience/bloomz
28. Les données utilisées pour l'affinage constituent le corpus xP3 : voir https://huggingface.co/datasets/bigscience/xP3.

démonstration qui introduit dans le contexte gauche du décodeur des exemples de couples entrée-sortie. Concernant le nombre d'exemples, nous utilisons la série : '*zéro-exemple*, '*mono-exemple*', '*n-exemples*', '*oligo-exemples*' pour faire pendant aux termes anglais correspondants (*zero-shot*, *one-shot*, *n-shot*, *few-shot*, etc).

C Amorces

Dans cette section, nous donnons les amorces pour chacune des tâches évaluées dans l'article. Les amorces pour la tâche d'implication textuelle (corpus `XNLI`) se trouvent dans le tableau 9, pour la tâche de REN dans le domaine général (corpus `WikiNER_fr`) dans le tableau 10, pour la tâche de REN dans le domaine clinique (corpus `QuaeroFrenchMed`) dans le tableau 11 et pour la tâche de réponses aux questions (corpus `PIAF`) dans le tableau 12.

Lang.	Nom	Amorce	Continuation
en	`based_on_the_previous_passage`	[prémisse] Based on the previous passage, is it true that "[hypothèse]" ? ¶Yes, no, or maybe ?	[réponse]
fr	`based_on_the_previous_passage`	[prémisse] Etant donné le passage précédent, est-il vrai que : "[hypothèse]" ?¶Oui, Non ou Peut-être ?	[réponse]
en	`can_we_infer`	Suppose [prémisse] Can we infer that "[hypothèse]" ? Yes, no, or maybe ?	[réponse]
fr	`can_we_infer`	Supposons [prémisse] Peut-on en déduire que "[hypothèse]" ? Oui, Non ou Peut-être ?	[réponse]
en	`does_it_follow_that`	Given that [prémisse] Does it follow that "[hypothèse]" Yes, no, or maybe ?¶	[réponse]
fr	`does_it_follow_that`	Etant donné [prémisse] S'ensuit-il que : "[hypothèse]" Oui, Non ou Peut-être ? ¶	[réponse]
en	`take_the_following_as_truth`	Take the following as truth : [prémisse]¶Then the following statement : "[hypothèse]" is "true", "false", or "inconclusive" ?	[réponse]
fr	`take_the_following_as_truth`	Supposons que ce qui suit est vrai : [prémisse]¶Alors l'enoncé suivant : "[hypothèse]" est-il "vrai", "faux", ou "indécidable" ?	[réponse]

TABLE 9 – Amorces utilisées pour la tâche d'implication textuelle (corpus `XNLI`). Les amorces existent systématiquement en deux langues (en, fr).

Entité	Nom	Amorce	Continuation
Personne	`LIST_PER`	Lister les entités de type "personne" dans le texte suivant : [passage]	[liste d'entités]
Lieu	`LIST_PER`	Lister les entités de type "lieu" dans le texte suivant : [passage]	[liste d'entités]
Organisation	`LIST_ORG`	Lister les entités de type "organisation" dans le texte suivant : [passage]	[liste d'entités]
Mélangé	`choose_entity`	Dans le texte suivant, [mention] est de quel type entre "personne", "lieu" ou "organisation" ? [passage]	{personne,lieu,organisation}

TABLE 10 – Amorces pour la tâche de REN dans le domaine général (corpus `WikiNER_fr`).

Entité	Nom	Amorce	Continuation
Symptôme et maladie	LIST_DISO	Voici (le titre d'un article scientifique médical | une notice patient) : [passage] Lister tous les symptômes et maladies qui sont mentionnées dans (ce titre | cette notice).	[liste d'entités]
Partie du corps	LIST_ANAT	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister toutes les parties du corps qui sont mentionnées dans (ce titre | cette notice).	[liste d'entités]
Composant chimique	LIST_CHEM	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister tous les composants chimiques qui sont mentionnés dans (ce titre | cette notice).	[liste d'entités]
Être vivant	LIST_LIVB	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister tous les êtres vivants qui sont mentionnés dans (ce titre | cette notice).	[liste d'entités]
Procédure médicale	LIST_PROC	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister toutes les procédures médicales qui sont mentionnées dans (ce titre | cette notice).	[liste d'entités]
Physiologie humaine	LIST_PHYS	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister toutes les physiologies humaines qui sont mentionnées dans (ce titre | cette notice).	[liste d'entités]
Phénomène physiologique	LIST_PHEN	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister tous les phénomènes physiologiques qui sont mentionnés dans (ce titre | cette notice).	[liste d'entités]
Appareil	LIST_DEVI	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister tous les appareils qui sont mentionnés dans (ce titre | cette notice).	[liste d'entités]
Zone géographique	LIST_GEOG	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister toutes les zones géographiques qui sont mentionnées dans (ce titre | cette notice).	[liste d'entités]
Objet	LIST_OBJC	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Lister tous les objets qui sont mentionnés dans (ce titre | cette notice).	[liste d'entités]
Mélangé	choose_entity	Voici (le titre d'un article scientifique médical | une notice patient) : [passage]. Quel est le type de l'entité [mention] parmi "symptôme et maladie", "partie du corps", "composant chimique", "être vivant", "procédure médicale", "physiologie humaine", "phénomène physiologique", "appareil", "zone géographique" ?	{Type prédit}

TABLE 11 – Amorces utilisées pour la tâche de REN (corpus QuaeroFrenchMed). Les invites s'adaptent au type du document.

Langue	Nom	Amorce	Continuation
fr	after_reading	Après avoir lu le paragraphe, merci de répondre à la question qui le suit : ¶ [passage] ¶ [question] ¶	[réponse]
en	after_reading	After reading the following paragraph, please answer the question that follows : ¶ [passage] ¶ [question]	[réponse]
fr	given_above_context	[passage] ¶ Etant donné le contexte qui précède, [question]	[réponse]
en	given_above_context	[passage] ¶ Given the above context, [question]	[réponse]
fr	given_passage_answer	Étant donné le passage suivant, répondre à la question ci-dessous : ¶[passage] ¶[question]	[réponse]
en	given_passage_answer	Given the following passage answer the question that follows : ¶[passage] ¶[question]	[réponse]
-	context_follow_q	[passage] ¶ Q : [question] ¶A :	[réponse]

TABLE 12 – Amorces utilisées pour la tâche de réponses aux questions (corpus PIAF). Elles présentent diverses variations de l'ordre relatif de l'instruction, du passage, et de la question. Le symbole ¶ identifie les retours à la ligne.

D Reconnaissance d'entités nommées

D.1 Résultats pour le domaine générale

Les résultats pour la tâche de REN dans le domaine général pour les amorces de type list_{PER,LOC,ORG} se trouvent dans le tableau 13.

D.2 Analyse des erreurs (WikiNER_fr)

Pour mieux comprendre les résultats decevants de la section 3.4, nous avons étudié de façon automatique les erreurs produites pour l'amorce *LIST_PER* (*mono-exemple*) sur le jeu de test WikiNER_fr, en nous focalisant en particulier sur trois cas que nous avions identifiés comme étant sources potentielles d'erreurs :

Invite	Modèle	*zéro-exemple* P	R	F1	*mono-exemple* P	R	F1
list_PER	Bloom_560m	0,00	0,01	0,00	0,01	0,05	0,02
	Bloom_1b1	0,00	0,01	0,00	0,01	0,04	0,01
	Bloom_3b	0,00	0,01	0,00	0,02	0,09	0,03
	Bloom_7b1	0,00	0,01	0,00	0,03	0,13	0,05
	Bloom	0,00	0,01	0,00	0,03	0,15	0,05
	Bloomz	0,08	0,32	0,13	0,08	0,32	0,13
list_LOC	Bloom_560m	0,00	0,01	0,00	0,01	0,03	0,01
	Bloom_1b1	0,00	0,01	0,00	0,01	0,02	0,01
	Bloom_3b	0,00	0,01	0,00	0,03	0,08	0,04
	Bloom_7b1	0,00	0,01	0,01	0,05	0,12	0,07
	Bloom	0,00	0,01	0,00	0,06	0,19	0,10
	Bloomz	0,06	0,17	0,09	0,06	0,17	0,09
list_ORG	Bloom_560m	0,00	0,01	0,00	0,00	0,01	0,00
	Bloom_1b1	0,00	0,01	0,00	0,00	0,01	0,00
	Bloom_3b	0,00	0,01	0,00	0,00	0,02	0,00
	Bloom_7b1	0,00	0,01	0,00	0,00	0,03	0,01
	Bloom	0,00	0,01	0,00	0,01	0,05	0,01
	Bloomz	0,00	0,04	0,01	0,00	0,04	0,01

TABLE 13 – Résultats de Bloom et Bloomz pour la reconnaisance d'entités nommées dans le domaine général (corpus de test de WikiNER_fr) pour chaque type d'entité nommée (précision, rappel et F-mesure en utilisant la mesure *fuzzy_list*).

1. la prédiction partielle des entités par rapport aux annotations de référence (p. ex. *Clodion* au lieu de *Clodion le Chevelu*, *Noé* au lieu de *Gaspar Noé*, *Diana Spencer* au lieu de *Lady Diana Spencer* et *Fox* au lieu de *sœurs Fox*). Selon les métriques utilisées (précision, rappel et F-mesure), seules les prédictions complètes sont considérées comme correctes. Nous identifions le nombre de prédictions partielles en comptant le nombre d'entités prédites présentes dans le texte de référence, mais qui ne correspondent pas à une entité entière dans liste d'entités de référence.
2. le texte prédit correspond non pas à des entités mais à une continuation plausible du passage. Quand cela arrive, il a souvent pour résultat la génération d'un texte long. Nous comptons donc le nombre de prédictions qui contiennent plus de 10 unités, où une unités est une séquence de caractères séparés par des blancs.
3. la prédiction vide erronée, c'est-à-dire que la prédiction est la chaîne vide, tandis que la référence contenait au moins une entité. Ceci arrive souvent pour les trois amorces de type `LIST_` et représente un des défis majeurs de cette tâche. Nous contrastons ce nombre avec le nombre de prédictions vides *correctes*, où la prédiction est vide et il n'existe effectivement pas d'entité du type spécifié dans le texte.

La figure 1 donne une idée de l'ampleur de chacune des erreurs, pour quatre tailles du modèle `Bloom` et pour `Bloomz`. Comme mentionné ci-dessus, nous incluons aussi la catégorie *vide-correct* pour mieux analyser la catégorie *vide-erreur*. De loin, le plus grand problème correspond à la deuxième erreur (continuation du passage plutôt qu'une prédiction des entités). Le problème a tendance à diminuer lorsque la taille du modèle augmente et le nombre de ces erreurs. Le nombre de prédictions partielles reste faible pour tous les modèles, même s'il y a une légère augmentation lorsque la taille du modèle augmente. Enfin, le nombre de prédictions vide augmente avec la taille du modèle, que ce soit des prédictions vides correctes ou incorrectes, sauf pour le grand modèle `Bloom` qui montre une baisse dans le nombre de prédictions vides par rapport à `Bloom` _7b1 et `Bloom` _3b. Les meilleurs scores de `Bloomz` dans le tableau 13, sont peut-être expliqués par le nombre relativement faible de textes trop longs (qui indique une continuation du texte de l'amorce et aussi le plus grand taux de prédictions vides (ce qui augmente considérablement le nombre de prédictions vites correctes, même si cela mène à plus de prédictions vides de façon erronée).

D.3 Résultats pour le domaine clinique

Les résultats pour la tâche de REN dans le domaine clinique se trouvent dans les tableaux 14 et 15.

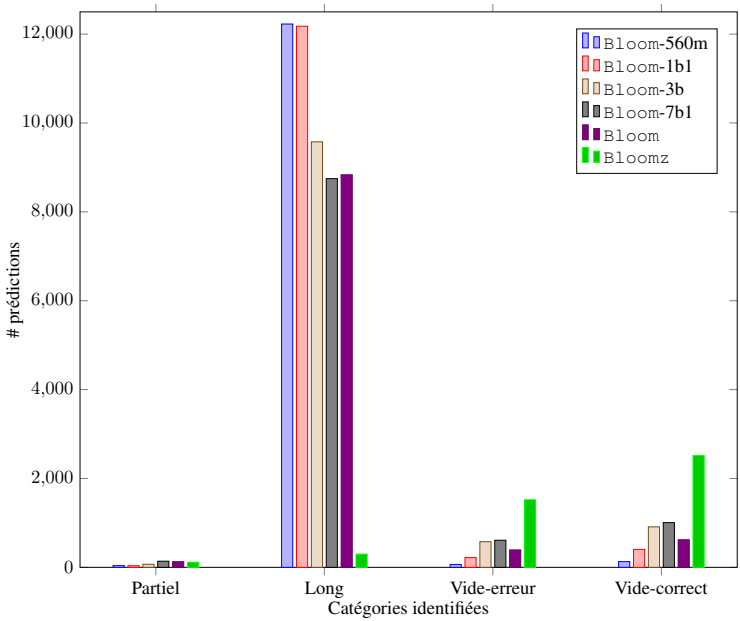

FIGURE 1 – Analyse d'erreurs des modèles `Bloom` et `Bloomz` pour la reconnaissance d'entités nommées (corpus de test de `WikiNER`) pour l'amorce LIST_PER en *mono-exemple*. Les trois catégories d'erreurs correspondent aux nombres de : (i) prédictions partielles (souvent correctes mais sans être une correspondence exacte) et (ii) prédictions trop longues (correspondant souvent à une continuation du texte) et (iii) prédictions vides de façon erronée. Nous incluons une quatrième analyse dans cette figure, correspondant au nombre de prédictions vides correctes pour permettre à donner un ordre de grandeur à la catégorie (iii).

Modèle	# exemples	ANAT			CHEM			DISO			Tous		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Bloom-560m	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.01	0.0
	1	0.08	0.07	0.07	0.05	0.04	0.05	0.1	0.07	0.08	0.06	0.05	0.05
	5	0.15	0.08	0.1	0.1	0.06	0.07	0.1	0.08	0.09	0.1	0.06	0.07
	10	0.19	0.12	0.15	0.14	0.1	0.11	0.14	0.12	0.13	0.13	0.09	0.11
Bloom-1b1	0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	0.09	0.09	0.09	0.05	0.05	0.05	0.07	0.07	0.07	0.06	0.06	0.06
	5	0.17	0.08	0.11	0.18	0.13	0.15	0.17	0.15	0.16	0.13	0.09	0.11
	10	0.18	0.14	0.16	0.18	0.19	0.19	0.16	0.18	0.17	0.12	0.13	0.13
Bloom-3b	0	0.0	0.0	0.0	0.0	0.02	0.01	0.0	0.01	0.01	0.0	0.01	0.0
	1	0.15	0.11	0.13	0.07	0.05	0.06	0.09	0.07	0.08	0.08	0.06	0.07
	5	0.2	0.16	0.18	0.15	0.11	0.13	0.18	0.16	0.17	0.14	0.12	0.13
	10	0.23	0.21	0.22	0.22	0.19	0.2	0.18	0.19	0.19	0.16	0.15	0.16
Bloom-7b1	0	0.0	0.04	0.01	0.01	0.07	0.02	0.01	0.05	0.02	0.0	0.04	0.01
	1	0.15	0.11	0.13	0.07	0.05	0.06	0.1	0.08	0.09	0.09	0.06	0.08
	5	0.16	0.15	0.16	0.15	0.07	0.1	0.17	0.19	0.18	0.13	0.12	0.13
	10	0.15	0.17	0.16	0.23	0.11	0.15	0.19	0.22	0.2	0.16	0.15	0.15
Bloom	0	0.0	0.01	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.01	0.0
	1	0.1	0.12	0.11	0.08	0.1	0.09	0.07	0.06	0.07	0.07	0.07	0.07
	5	0.15	0.2	0.17	0.16	0.07	0.1	0.15	0.1	0.12	0.12	0.09	0.1
	10	0.18	0.19	0.19	0.24	0.08	0.12	0.17	0.07	0.1	0.19	0.08	0.11
Bloomz-560m	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	0.03	0.03	0.03	0.02	0.01	0.02	0.06	0.04	0.05	0.03	0.03	0.03
	5	0.04	0.08	0.05	0.03	0.08	0.04	0.09	0.1	0.09	0.05	0.07	0.05
	10	0.07	0.12	0.09	0.06	0.14	0.08	0.13	0.15	0.14	0.07	0.11	0.09
Bloomz-1b1	0	0.0	0.0	0.0	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.02	0.01
	1	0.07	0.06	0.06	0.07	0.05	0.06	0.07	0.06	0.07	0.05	0.04	0.05
	5	0.12	0.11	0.11	0.1	0.11	0.11	0.12	0.12	0.12	0.09	0.09	0.09
	10	0.13	0.16	0.14	0.16	0.21	0.18	0.13	0.15	0.14	0.11	0.13	0.12
Bloomz-3b	0	0.02	0.06	0.03	0.08	0.24	0.12	0.05	0.0	0.01	0.04	0.08	0.04
	1	0.04	0.07	0.05	0.02	0.06	0.03	0.06	0.06	0.06	0.03	0.05	0.04
	5	0.07	0.15	0.09	0.06	0.17	0.08	0.11	0.13	0.12	0.06	0.11	0.08
	10	0.12	0.27	0.17	0.07	0.21	0.11	0.13	0.15	0.14	0.09	0.15	0.11
Bloomz-7b1	0	0.11	0.21	0.14	0.11	0.27	0.15	0.11	0.1	0.1	0.09	0.17	0.1
	1	0.05	0.11	0.07	0.02	0.07	0.04	0.07	0.08	0.07	0.04	0.07	0.05
	5	0.08	0.18	0.11	0.04	0.13	0.06	0.11	0.16	0.13	0.07	0.13	0.09
	10	0.14	0.18	0.16	0.07	0.2	0.1	0.07	0.15	0.09	0.07	0.15	0.1
Bloomz	0	0.07	0.26	0.11	0.11	0.42	0.17	0.05	0.25	0.08	0.05	0.27	0.08
	1	0.07	0.1	0.08	0.03	0.08	0.04	0.04	0.1	0.06	0.03	0.07	0.05
	5	0.12	0.15	0.13	0.06	0.17	0.08	0.08	0.19	0.12	0.07	0.14	0.09
	10	0.16	0.2	0.18	0.06	0.16	0.08	0.1	0.22	0.14	0.08	0.17	0.11

TABLE 14 – Résultats de Bloom et bloomz pour la reconnaissance d'entités nommées dans le domaine clinique (corpus MEDLINE) en faisant varier la taille du modèle et le nombre d'exemples pour chaque type d'entité nommée (précision, rappel et F-mesure en utilisant la mesure *fuzzy_list*). Nous rapportons les résultats pour les trois types d'entités nommées dans le corpus, ainsi que les résultats tout type confondu.

Modèle	# exemples	ANAT			CHEM			DISO			Tous		
		P	R	F1	P	R	F1	P	R	F1	P	R	F1
Bloom-560m	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0	0.0	0.0	0.0
	1	0.01	0.01	0.01	0.14	0.1	0.12	0.01	0.02	0.02	0.06	0.05	0.06
	5	0.0	0.0	0.0	0.3	0.14	0.19	0.03	0.01	0.01	0.13	0.06	0.08
	10	0.0	0.0	0.0	0.41	0.2	0.27	0.11	0.04	0.06	0.19	0.09	0.12
Bloom-1b1	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.02	0.01	0.0	0.0	0.0
	1	0.04	0.05	0.05	0.18	0.14	0.16	0.04	0.06	0.05	0.09	0.08	0.08
	5	0.07	0.03	0.04	0.39	0.22	0.28	0.1	0.12	0.11	0.19	0.13	0.15
	10	0.05	0.03	0.03	0.44	0.26	0.32	0.11	0.14	0.12	0.22	0.16	0.18
Bloom-3b	0	0.0	0.04	0.0	0.01	0.01	0.01	0.0	0.01	0.0	0.0	0.01	0.0
	1	0.07	0.06	0.06	0.18	0.14	0.16	0.05	0.07	0.06	0.1	0.09	0.09
	5	0.17	0.07	0.1	0.41	0.22	0.28	0.25	0.12	0.17	0.25	0.13	0.17
	10	0.17	0.06	0.09	0.5	0.29	0.37	0.23	0.16	0.19	0.29	0.17	0.22
Bloom-7b1	0	0.0	0.0	0.0	0.01	0.01	0.01	0.0	0.0	0.0	0.0	0.01	0.0
	1	0.06	0.04	0.05	0.23	0.15	0.18	0.07	0.06	0.06	0.12	0.09	0.1
	5	0.09	0.05	0.06	0.44	0.23	0.3	0.15	0.12	0.13	0.24	0.15	0.18
	10	0.09	0.06	0.07	0.54	0.34	0.42	0.15	0.15	0.15	0.28	0.21	0.23
Bloom	0	0.0	0.01	0.0	0.0	0.01	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	0.05	0.06	0.06	0.21	0.16	0.18	0.02	0.01	0.01	0.09	0.07	0.08
	5	0.08	0.09	0.08	0.45	0.32	0.37	0.07	0.02	0.03	0.21	0.14	0.16
	10	0.16	0.12	0.14	0.53	0.35	0.42	0.0	0.0	0.0	0.27	0.15	0.19
Bloomz-560m	0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	1	0.0	0.01	0.01	0.11	0.07	0.08	0.01	0.01	0.01	0.05	0.04	0.04
	5	0.0	0.0	0.0	0.23	0.16	0.19	0.01	0.02	0.01	0.1	0.08	0.08
	10	0.0	0.0	0.0	0.27	0.2	0.23	0.03	0.04	0.03	0.12	0.1	0.11
Bloomz-1b1	0	0.04	0.19	0.07	0.61	0.21	0.31	0.08	0.17	0.11	0.24	0.16	0.15
	1	0.02	0.03	0.02	0.25	0.14	0.18	0.03	0.04	0.03	0.11	0.08	0.09
	5	0.01	0.01	0.01	0.28	0.22	0.25	0.02	0.03	0.03	0.12	0.11	0.12
	10	0.03	0.04	0.04	0.31	0.27	0.29	0.03	0.05	0.04	0.14	0.15	0.14
Bloomz-3b	0	0.01	0.06	0.01	0.32	0.32	0.32	0.07	0.12	0.09	0.15	0.22	0.16
	1	0.0	0.01	0.0	0.24	0.19	0.21	0.02	0.04	0.02	0.1	0.1	0.09
	5	0.01	0.06	0.02	0.33	0.28	0.3	0.03	0.07	0.05	0.13	0.14	0.13
	10	0.01	0.03	0.01	0.4	0.32	0.36	0.04	0.1	0.06	0.16	0.17	0.16
Bloomz-7b1	0	0.05	0.26	0.08	0.5	0.35	0.41	0.11	0.19	0.14	0.22	0.28	0.21
	1	0.01	0.05	0.02	0.2	0.17	0.18	0.02	0.05	0.03	0.09	0.1	0.09
	5	0.02	0.08	0.03	0.32	0.29	0.31	0.06	0.12	0.08	0.15	0.18	0.16
	10	0.07	0.12	0.08	0.42	0.34	0.37	0.01	0.03	0.01	0.18	0.18	0.17
	0	0.01	0.02	0.02	0.2	0.02	0.04	0.0	0.0	0.0	0.07	0.03	0.02
	1	0.02	0.05	0.03	0.18	0.15	0.16	0.01	0.03	0.01	0.07	0.07	0.07
	5	0.02	0.05	0.03	0.37	0.29	0.33	0.0	0.01	0.0	0.14	0.13	0.13
	10	0.04	0.09	0.06	0.41	0.31	0.36	0.01	0.04	0.01	0.17	0.16	0.15

TABLE 15 – Résultats de Bloom et bloomz pour la reconnaissance d'entités nommées dans le domaine clinique (corpus EMEA) en faisant varier la taille du modèle et le nombre d'exemples pour chaque type d'entité nommée (précision, rappel et F-mesure en utilisant la mesure *fuzzy_list*). Nous rapportons les résultats pour les trois types d'entités nommées dans le corpus, ainsi que les résultats tout type confondu.